Dissertation
submitted to the
Combined Faculties of the Natural Sciences and Mathematics
of the Ruperto-Carola-University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Put forward by
Nina Hernitschek
born in Freyung
Oral examination: 26.01.2017

Astrophysical Modeling of Time-Domain Surveys

Referees: Prof. Dr. Hans-Walter Rix
Prof. Dr. Norbert Christlieb

"... the ways by which men arrive at knowledge of the celestial things are hardly less wonderful than the nature of these things themselves"

- Johannes Kepler

Astrophysikalische Modellierung zeitaufgelöster Himmelsdurchmusterungen

Zusammenfassung:
Das Ziel dieser Arbeit ist die Entwicklung und Anwendung algorithmischer Methoden für die Modellierung zeitaufgelöster Beobachtungsdaten. Solchen Methoden kommt im Kontext aktueller und zukünftiger großangelegter zeitaufgelöster Himmelsdurchmusterungen besondere Bedeutung zu. Der Fokus der Arbeit liegt auf der Quantifizierung und Charakterisierung der Variabilität astrophysikalischer Objekte ausgehend von nichtsimultanen, lückenhaften zeitaufgelösten Multiband-Lichtkurven, die exemplarisch auf Daten des Pan-STARRS1 (PS1) 3π angewandt werden. Variabilitätsamplituden und Zeitskalen werden hierbei mittels Lichtkurven-Strukturfunktionen abgeschätzt. Anhand von PS1 3π-Daten in mit SDSS S82 überlappenden Regionen, für die bereits eine Klassifizierung vorliegt, wird ein Klassifizierungsalgorithmus aus dem Bereich des Maschinellen Lernens trainiert, um QSOs und RR Lyrae anhand ihrer Variabilität und mittleren Farben zu bestimmen.
Dieser Ansatz ermöglicht eine variabilitÃ€tsselektierte, annähernd vollständige und reine Auswahl von QSO und RR Lyrae (außerhalb der galaktischen Scheibe, die in ihrer Kombination aus Flächenabdeckung, Tiefe (Erfassung schwacher Objekte) und Zuverlässigkeit beispiellos ist. Sie enthält $\sim 4.8 \times 10^4$ hochwahrscheinliche RR Lyrae-Kandidaten im galaktischen Halo, sowie $\sim 3.7 \times 10^6$ hochwahrscheinliche QSO-Kandidaten.
Die resultierende Karte der RR Lyrae-Kandidaten deckt 3/4 des Himmels ab und zeigt Strukturen bis in 130 kpc Entfernung mit 3% Entfernungsgenauigkeit. Insbesondere kann der Sagittarius stream, dargestellt durch RR Lyrae, in noch nie dagewesener Qualität kartiert werden.
Darüber hinaus werden die Eigenschaften von PS1 3π und seine Rolle als Pilotprojekt für den zukünftigen LSST dargelegt.

Astrophysical Modeling of Time-Domain Surveys

Abstract:
The goal of this work is to develop and apply algorithmic approaches for astrophysical modeling of time-domain surveys. Such approaches are necessary to exploit ongoing and future all-sky time-domain surveys. I focus on quantifying and characterizing source variability based on sparsely and irregularly sampled, non-simultaneous multi-band light curves, with an application to the Pan-STARRS1 (PS1) 3π survey: variability amplitudes and timescales are estimated via light curve structure functions. Using PS1 3π data on the SDSS "Stripe 82" area whose classification is available, a supervised machine-learning classifier is trained to identify QSOs and RR Lyrae based on their variability and mean colors.
This leads to quite complete and pure variability-selected samples of QSO and RR Lyrae (away from the Galactic disk), that are unmatched in their combination of area, depth and fidelity. The sample entails $\sim 4.8 \times 10^4$ likely RR Lyrae in the Galactic halo, and $\sim 3.7 \times 10^6$ likely QSO.
The resulting map of RR Lyrae candidates across 3/4 of the sky reveals targets to $\sim$130 kpc, with distances precise to $\sim$3%. In particular, the sample leads to an unprecedented map of distance and width of Sagittarius stream, as traced by RR Lyrae.
Furthermore, the role of PS1 3π as pilot survey for the upcoming LSST survey is discussed.

List of Acronyms

AGN	Active Galactic Nuclei
CCD	charge coupled device
DRW	Damped Random Walk
FOV	field of view
Gyr	Gigayears, 10^9 years
HB	horizontal branch
HR diagram	Herzsprung-Russell diagram
intergalactic medium	IGM
IR	infrared
LMC	Large Magellanic Cloud
near-IR	near-infrared
PB	Petabyte (10^{15} byte)
PS1 3π	Pan-STARRS1 3π
RFC	Random Forest Classifier
RGB	red giant branch
SDSS	Sloan Digital Sky Survey
SMC	Small Magellanic Cloud
WISE	Wide-field Infrared Survey
ZAHB	zero-age horizontal branch
ZAMS	zero-age main sequence

Contents

Chapter 1

Introduction

During the last decades, a number of astronomical surveys has been carried out, monitoring and mapping the sky with ranges from our own solar system and cosmic neighborhood, but also to survey deeply into the cosmos. Whether they observe all-sky or not, whether they are ground- or space-based: what many of them have in common is that they are multi-epoch surveys. Advances both in instrumentation as well as in computer hardware and software design made these challenging tasks possible.

With upcoming surveys like the Large Synoptic Survey Telescope (LSST, see the *LSST Science Book, Version 2.0*), monitoring a large fraction of the sky at a high cadence, the data produced at each night will easily exceed the terabyte scale.

Such data rates make data processing and astrophysical analysis even more challenging, but these surveys will enable a more detailed view on variable and transient phenomena.

This thesis deals with the development of methods for quantifying astronomical properties of non-simultaneous, sparse multi-band time-domain surveys in order to identify and classify variable sources such as QSOs, RR Lyrae and Cepheids. The focus of the work is to develop and test a methodology, based on so-called multi-band light-curve structure functions, that is then applied to Pan-STARRS1 3π, but can be used for time-domain surveys in general.

In Chapter 2, an extensive introduction to variable astrophysical sources is given. The time-domain appearance of such variable sources, as well as their astrophysical background is described. The case of RR Lyrae and Cepheids as two examples for variable stars, as well as QSOs as an example for a non-stellar variable source, are laid out in greater detail. Subsequently, an overview of time-domain surveys is given.

Chapter 3 deals with the properties of Pan-STARRS1 3π as time-domain survey and addresses how it can serve as a pilot survey for LSST. Sky coverage, wavebands, observational baseline and the observational strategy, as well as caveats are described.

After describing pre-conditions and the environment of this thesis, the following chapters show new research both in methodology and results.

In Chapter 4, the methodological concepts of finding variable sources, quantifying their astro-statistical properties as well as automated source classification by machine-learning methods are

introduced. Here a focus is given to the question how to deal with the challenges that come up when developing such methodology for non-simultaneous multi-band surveys. Additionally to presenting known variability measures, such as single-band structure functions, in this chapter also the newly developed multi-band structure function fitting is given. Chapter 4 also deals with automated classification of sources by machine-learning classifiers and with the question how to test and quantify the reliability of such methods. Here, concepts are given that are applied in the following chapters to Pan-STARRS1 3π data.

The methodology presented in Chapter 4 – the newly developed multi-band structure fitting as well as a machine-learning classifier – is applied in Chapter 5 to data from Pan-STARRS1 3π in order to estimate variability measures such as amplitudes and time scales for all point sources brighter than r_{P1}=21.5 mag. The identification of QSOs and RR Lyrae candidates lead to a catalog of 25.8×10^6 variable sources. Chapter 5 extends a recent publication (Hernitschek et al. 2016) as well as shows further development and science.

Using the identified RR Lyrae candidates, the extent and geometry of Sagittarius (Sgr) stream is mapped in Chapter 6. The geometry of the Sgr stream, as traced by RR Lyrae candidates, is explored and quantified by fitting its spatial extend and width as a function of the angle $\tilde{\Lambda}_\odot$ in its orbital plane.

Chapter 7 offers a summary and discussion of the presented methodology and obtained results. The broader astrophysical context as well as the application of the new methodology to upcoming surveys is discussed in more detail.

The Appendix gives a summary on time series analysis and Markov Chain Monte Carlo Methods, as used in Chapter 5.

1.1 Research Questions

Many different astrophysical systems vary in brightness with time, Therefore, source classification is one of the key issues of large all-sky time-domain surveys. The amount of data available from nowadays surveys, and even more expected from upcoming, makes it absolutely necessary to apply automated methods that are reliable, general and fast. The process has to be as automated as possible, robust and reliable, it has to operate with sparse and heterogeneous data, and it has to maintain a high purity (low false alarm rate). Furthermore, it has to be extensible for new types of sources to identify.

In this context, this thesis deals with the development, test and application of a new approach for quantifying statistical properties of non-simultaneous, sparse, multi-color light curves, assigning measures, like amplitude and time scale, as well as fitting light curves to get derived mean magnitudes.

As a first step, the light curve "structure function" is generalized to operate on multi-band light curves in a consistent way. This approach is applied to data from the Pan-STARRS 1 3 π sky survey (Chambers 2011). It is used to estimate variability measures for all point sources brighter than $r_{\mathrm{P1}} = 21.5$ mag and having a reasonable number of observational epochs.

The next challenge is then to use the photometric and variability information of the sources to associate classification probabilities that any given source belongs to known types of variable astrophysical sources. While doing so, various questions need to be addressed, related both to data quality required for carrying out this task, caveats related to reddening and crowding, as well as which statements can be made regarding reliability of the resulting classification.

To carry out such source classification from the PS1 3π data, the fact is used that the survey overlaps with SDSS's Stripe 82 (S82), an area on the sky with a rather complete inventory of identified variable sources. These S82 classifications are taken as "ground truth", and then a Random Forest classifier is used to identify RR Lyrae and QSOs based on their variability.

The sources identified as being variable will then be used to build a catalog of all variable point sources brighter than $r_{\mathrm{P1}} = 21.5$ mag within PS1 3π. Based on the RR Lyrae candidates, the distance precision of the RR Lyrae candidates will be estimated. As an astrophysical application, the spatial extent of Sagittarius stream as well as its width can be traced.

Chapter 2

Variable Sources in Time Domain Surveys

In this chapter, an introduction to variable and transient astrophysical sources is given, and how they are reflected in light curves.

The flow of this chapter follows mostly parts of the textbook by Catelan and Smith (2015)[1]. It gives an own compressed summary, adapted to the background of this thesis. If not indicated otherwise, figures are own graphics.

From an observational perspective, *transients* are sources that eventually fall below the detection limit when they are faint (such supernovae), whereas *variables* are sources that are always detectable, but change in brightness on various time scales.

The chapter is organized as follows. After a short overview of the detection of variable sources in history, variable sources are then grouped in types and sub-types by their light curve properties and physical causes of their variability. Pulsation is then identified as cause of variability for many variable stars, based on Catelan and Smith (2015). Subsequently, an overview of surveys being capable of detecting such sources is given.

2.1 Variable Sources in History

An astronomical source is considered as *variable*, if its appearance in brightness or color changes over time. Both the time-scales on which the variability happens and the amount (amplitude) of variability can differ by order of magnitudes, with time scales ranging from milliseconds to years, and brightness changes from fractions of a magnitude to several. Some sources vary regularly, or periodically, some not.

Despite a few recordings of stars appearing and disappearing as well as changing their visual appearance are known for more than 1000 years (Winkler et al. 2003), stars were considered as having fixed properties, at least for most of the time between their birth and death.

The bright supernova in Cassiopeia, seen in 1572 described by Tycho Brahe, and o (Omicron) Ceti by David Fabricius in 1596 are recognized as being the first recordings of variable stars (Hockey 2009).

[1] Catelan, M. and Smith, H. A. (2015). Pulsating stars. Wiley-VCH.

Describing o Ceti as a nova by David Fabricius, it was re-discovered in 1638 by the Frisian astronomer Jan Fokkes van Holwerd, known as Johannes Phocylides Holwarda (Hockey 2009; Catelan and Smith 2015). He soon recognized that this star not only disappears, but subsequently reappears. Holwarda assigned it a regular 11-month cycle.

During the following years, Johannes Hevelius carried out a detailed study of o Ceti and re-named it Mira, the name used since then. Only a few years later, the period of the star was determined to 33 days by Ismaël Boulliau, known as Ismaël Bullialdus (Hockey 2009). The discovery and description of Mira, being the archetype of later so-called *Mira variables*, led to a few more discoveries, like the discovery of the periodicity of χ Cygni in 1686 by Gottfried Kirch (Hockey 2009). However, the cause of their periodicity was not understood yet. Despite Bullialdus suggested in 1667 that Mira's change might be an effect of rotation, it took until the beginning of the 20th century to understand more about variable stars based on spectroscopic measurements. By this time, many types of variable stars today named Mira variables, RR Lyrae, Cepheids have been discovered.

2.2 Time-Domain Appearance of Variable Sources

The time-domain appearance of variable sources is closely related to their discovery.

Before photographic methods or even telescopes came up, astronomers denoted time, position and visual appearance of stars. Nowadays, methods can detect flux changes of fractions of a percent, time scales from milliseconds to years, and also detect variability of faint sources like QSOs. Sophisticated methods of observation and analysis can indicate the reason for the appearance of a source.

A historic example on how a detailed light curve analysis helped with revealing the physics behind the appearance is the star Algol, nowadays a prototype for eclipsing binaries. Early assumptions on variable stars included the case of eclipsing binaries as well brightness changes caused by one side of the star being significantly brighter than the other. Despite this is a reasonable assumption, and indeed is the cause of variability in many cases, this does not explain the variability of all stars. Detailed light curve analysis for certain classes of variable stars, such as Cepheids, revealed that this cannot hold as a general model. Pointing to the difference in light curve shapes between Algol-type and Cepheid-type stars, Plummer stated that the brightness change of Cepheids and stars of similar light curves would be hard to explain with the assumption of eclipsing binaries (Plummer 1913). This led to the hypothesis of radial pulsation.

So, light curves are both useful for detection of variable sources as well as they give hints on the origin of variable sources. Once a classification scheme is built, light curve analysis is a major resource for assigning types to light curves by classification.

2.2.1 Light Curves

Light curves are time series gained from repetitive photometric observations of the same astronomical source, showing the brightness variation of an object in certain bandpasses over a given time.

They are extracted from exposures of distinct sky regions. The time sampling (cadence) can be more or less regular. In general, ground based observations lead to more irregular sampling as the repeated observation of sources is affected by seasonal effects as well as weather, atmospheric and maintenance effects. Additionally, scheduling of multi-band surveys can also cause irregular sampling. Many surveys, for instance Pan-STARRS1 (PS1) 3π, cannot observe simultaneously in multiple bands; the Sloan Digital Sky Survey (SDSS) however is able to observe almost simultaneously. Irregularity in sampling is also caused by the order filters are used, the time needed to change the filters, and the usage of certain filters only under certain atmospheric and astronomical conditions as required in the survey design.

As an example for irregularity, a light curve from PS1 3π is shown in Fig. 2.1.

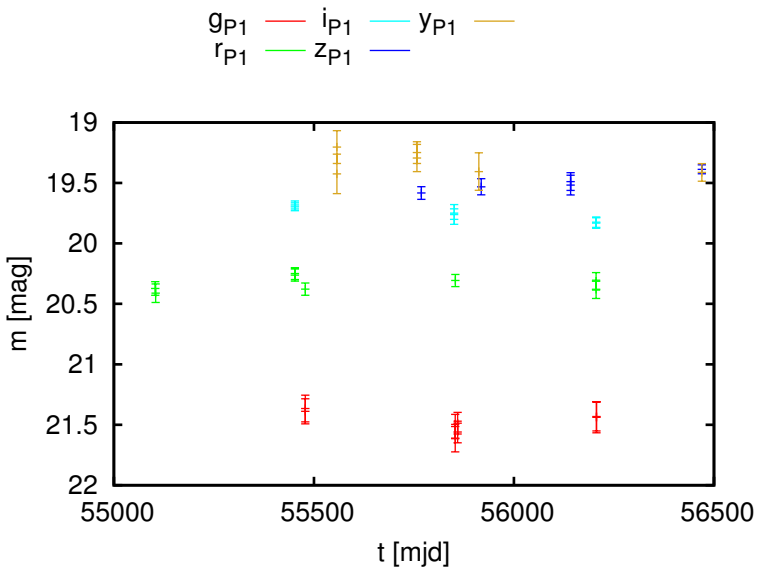


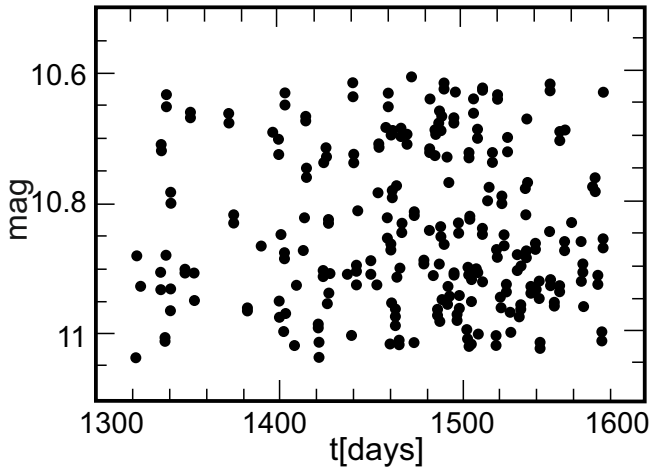
Figure 2.1 Example light curve from PS1 3π, with 33 observational epochs within five bands over 3.75 years. Magnitude uncertainties are indicating. Time is given in units of MJD = JD-2400000.5.

2.2.2 Analysis of Light Curves

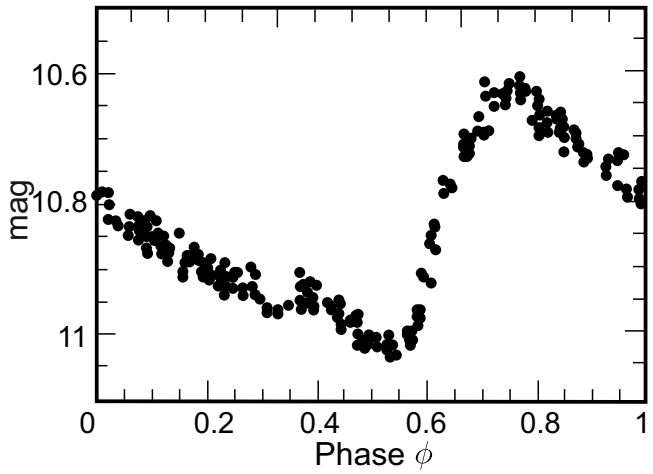
Among light curve analysis, one distinguishes basically two types of methods: such made for periodic variables, and such not requiring any periodicity in the light curve shape.

When dealing with periodic variables, the light curves can be compressed from the original observed light curve to a phase- (or period-)folded one.

Given the period $\mathcal{P}$ for a light curve with observations at time t_i, the light curve can be phase-folded by replacing the time axis by a phase $\phi = \left(\frac{t-t_0}{\mathcal{P}}\right) - E(t)$ (Hoffmeister et al. 1985), where t is the time of an observation, t_0 is some reference time, $\mathcal{P}$ is the period, and $E(t)$ indicates the integer part of $(t - t_0)/\mathcal{P}$, so phases are in the range $[0, 1[$. An example is shown in Fig. 2.2.



(a)



(b)

Figure 2.2 The unfolded (a) and folded (b) light curve of the variable star AT And, based on observations obtained during the Northern Sky Variability Survey (Woźniak et al. 2004). Time is given in units of MJD-50,000. The period-folded light curve in Subfigure (b) shows the periodicity of the light curve clearly, whereas it is concealed in the light curve itself. Taken from Catelan and Smith (2015).

By applying maximum-likelihood methods, this approach can lead to period determination. One example for doing so is the Lomb-Scargle periodogram (named after Lomb (1976) and Scargle (1982)), a method for finding periodicity in irregularly-sampled data. It is in many ways analogous to the more familiar Fourier Power Spectral Density (PSD) often used for detecting periodicity in regularly-sampled data.

However, not all variable sources show a periodic behavior. Whereas e.g. RR Lyrae, Cepheids and eclipsing binaries do (see Section 2.3), e.g. QSOs usually don't show periodic variability. In order to analyze such light curves, methods to determine variability time scales (instead of periods) and amplitudes (instead of determining the amplitude from maximum and minimum flux) are used, such as structure functions (Kozłowski et al. 2010; Kelly et al. 2009).

A structure function describes the mean squared difference (or, sometimes, root mean square difference) between pairs of observations of some object's brightness (or other property) as a function of the time lag between the observations. In more detail, the structure function is a description of a second-order statistic of the source's brightness history.

A detailed introduction to structure functions and other methods for non-periodic as well as irregularly sampled light curves can be found in Chapter 4.

Such approaches are both helpful for light curves of sources like QSOs showing stochastic behavior as well as light curves of sources who vary periodically but whose light curve is irregularly due to e.g. observational cadence and thus conceals periodicity. In the latter case, the more general non-periodic approaches help with applying a common parametric description to all sources of a given sample, irrespective if the individual source shows periodicity, and thus pre-selection of sources that vary and might possibly be periodic. More computationally expensive methods, being able to estimate periods even in such cases, can be applied subsequently. Such approaches are developed and carried out as part of this thesis work, to first assign variability measures not demanding periodicity, pre-selecting candidates for different types of variable sources, such as RR Lyrae, Cepheids and QSOs among them, and apply period-estimation techniques to only the candidates for variable sources.

Especially among the irregularly sampled light curves, approaches differ a lot for different signal-to-noise ratio and cadence at hand. Also, care must be taken in carrying out cleaning of photometric outliers. Approaches on outlier cleaning carried out as part of this thesis can be found in Chapter 5.

2.3 A Tree for Variable Sources

As mentioned before, this thesis focuses on QSOs and a few classes of periodically variable stars. Yet, nowadays more than 110 classes and subclasses of variable sources can be found in the *General Catalogue of Variable Stars* (GCV, Kholopov et al. 1998, Combined General Catalogue of Variable Stars, 4.1 Edition). Within the catalog, properties of the associated light curves are given, such as their light curve amplitudes and (in the case of periodic variables) their periods, as well as static properties such as luminosity classes.

Variable sources can be grouped in variability *types* and *classes* (Catelan and Smith 2015). The shape of their light curves divide them into three variability types: regular, semi-regular and irregular. Regular light curves show patterns repeated over time; a period can be assigned. Whereas

semi-regular light curves show some periodicity superimposed with a non-periodic signal amount, irregular light curves show no periodic behavior at all.
Variability classes can be assigned based on the cause for the variability (Watson et al. 2006). Light curves can vary by physical effects of the source itself, or external by alignment effects due to e.g. rotation or the observer's position. The former are called *intrinsic*, whereas the latter are called *extrinsic* variables. Examples for intrinsic variables are pulsating and eruptive variables, whereas eclipsing variables belong to the class of extrinsic variables. Rotational variables are assigned to the class of intrinsic or extrinsic variables, depending on the author, e.g. Eyer and Mowlavi (2008) vs. Catelan and Smith (2015).

The "variability tree" by Eyer and Mowlavi (2008), shown in Fig. 2.3, gives an overview of the most common variable stars and other variable sources like AGN and asteroids. The diagram is based on photometric variability. Spectral line profile variations are not taken into account for the scheme shown in the diagram.

In the following, a description of light curve appearance as well as underlying mechanisms is given for some of them, based on Catelan and Smith (2015). A more detailed description of the sources the analysis done in this thesis is carried out for – RR Lyrae, Cepheids and QSOs – is given in Section 2.5.

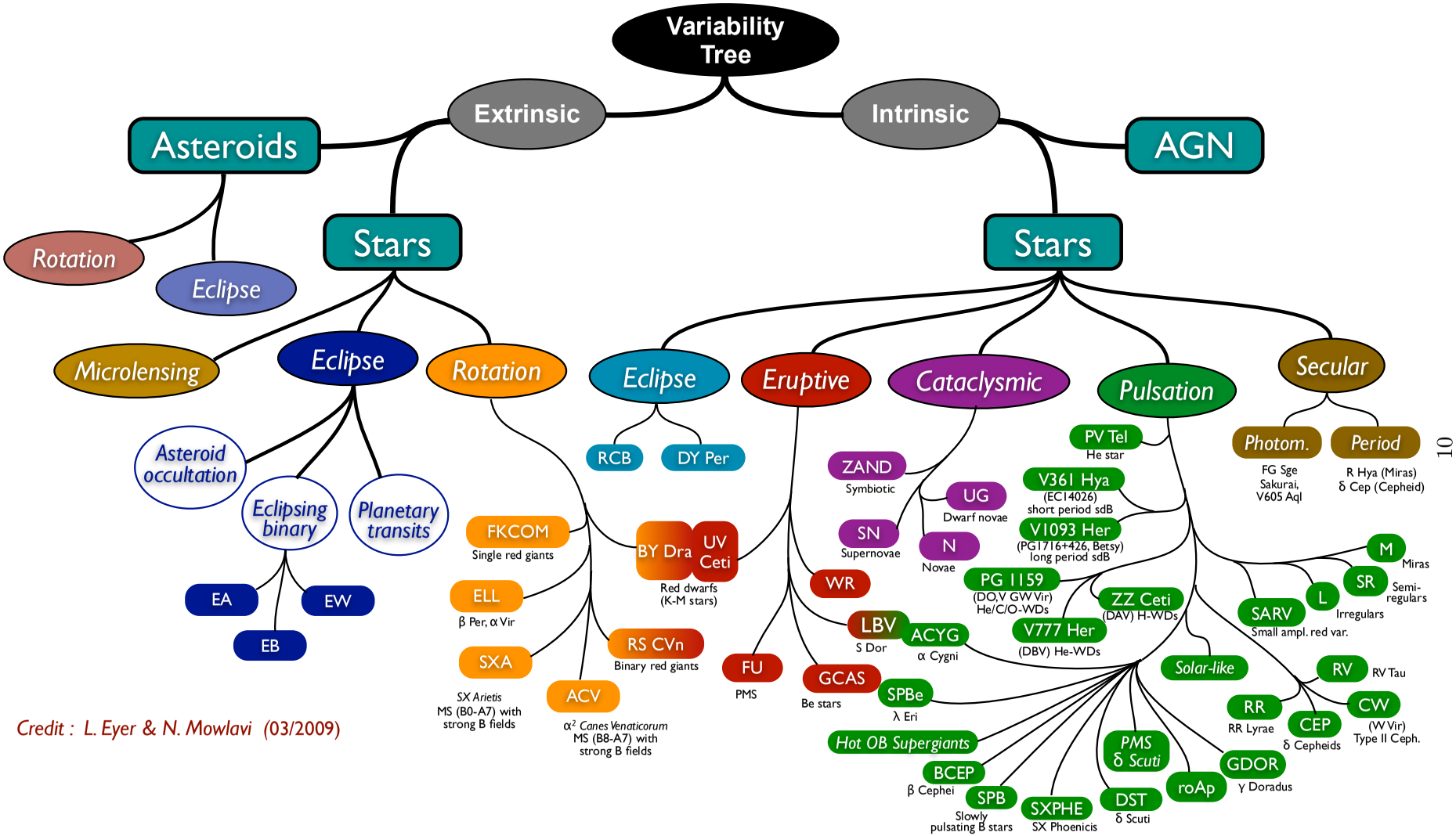


Figure 2.3 A tree for the most common photometric variable sources. Credit: N. Eyer & N. Movlawi.

2.3.1 Extrinsic Variables

In the case of extrinsic variables, the cause for the variability in the observed light curve is due to line-of-sight effects in binary or multiple star systems or the effect of the rotation of the source itself. In binary or multiple systems, variability is caused by the fact that one component is passing in front of the other as seen by the observer, whereas rotation can lead to variability due to star spots, magnetism or changing in shape.

Among them, there are both regular variables like eclipsing binaries and rotating stars as irregular variables like most asteroids.

Eclipsing Binaries

In eclipsing binary systems, variability in their brightness is caused by one component passing in front of the other when orbiting around its companion (see Fig. 2.5). A nearly edge-on alignment is required. Eclipsing binary stars play an important role in astrophysics, as they provide a robust method to derive stellar properties like radii, masses and ages (Popper 1980; Andersen 1991; Paczyński 1996).
Eclipsing binary stars can be divided into subclasses depending on the separation between the components. In Fig. 2.4, the configuration for detached, semi-detached, and contact binaries is shown. The black contour indicates the *Roche volume*, and the primary and secondary star are indicated.

The Roche volume, or Roche lobe is the region around a star in a binary system bounded by a critical gravitational equipotential. Within each Roche lobe, orbiting material is gravitationally bound to that star. When the extent of stellar material of a star within a binary system exceeds its Roche lobe, mass transfer to the companion star can occur. This is referred to as *Roche-lobe overflow*.

Table 2.1 gives a summary of eclipsing binaries and their properties.

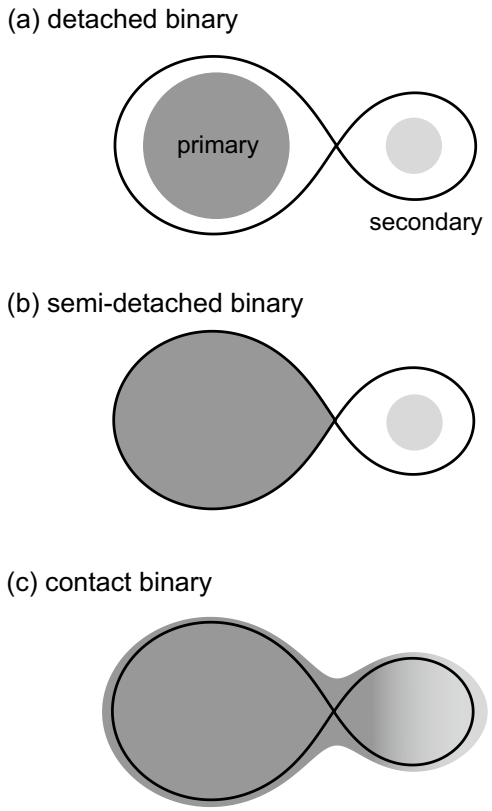


Figure 2.4 The configuration for detached, semi-detached, and contact binaries. The black contour indicates the Roche volume of the two stars.

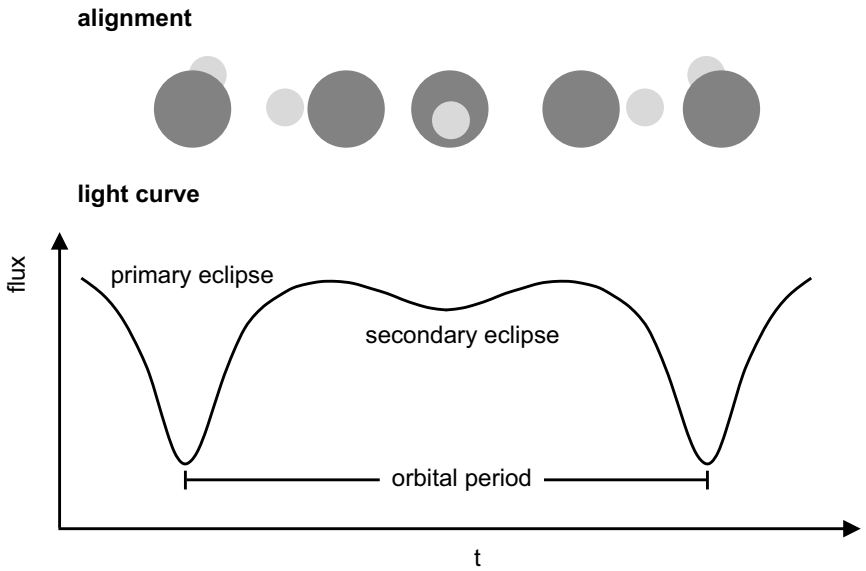


Figure 2.5 Alignment and resulting light curve for eclipsing binaries. The primary eclipse (where the secondary star goes behind the primary) as well the secondary eclipse (where the secondary star goes in front of the primary) are indicated.

Rotational Variables

For rotational variables, variability in their brightness is caused by either a non-uniform surface brightness or a flattened shape. To see variability, the rotation axis must not coincide with the line of sight. Spots on stars, like also seen on our Sun, are caused by strong local magnetic fields, leading to reduced surface temperature as the magnetic flux inhibits convection. Stars with ellipsoidal shapes also show changes in brightness as they present varying areas of their surfaces to the observer. The light curve is then modulated by these effects, depending on how the spots pass the line of sight.

Fig. 2.6 shows the effect of rotation on light curves of stars with significant spots. The light curve is clearly modulated on how the large spot passes the line of sight. Table 2.2 lists the most important types of rotational variable stars.

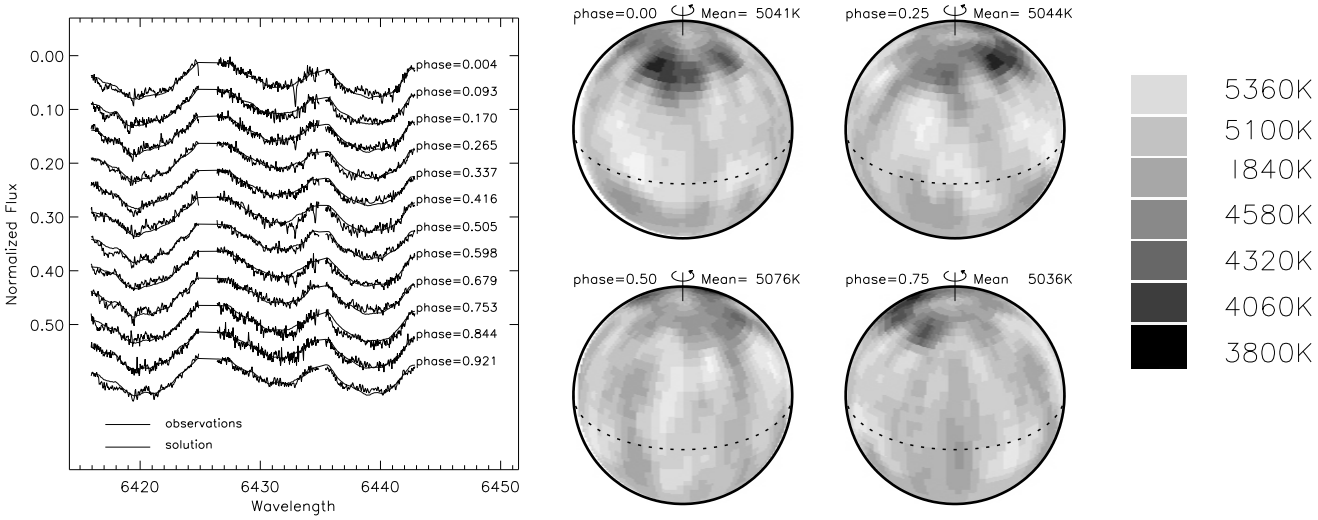


Figure 2.6 The effect of rotation on light curves of stars with significant spots, as shown on the example of FK Comae Berenices. The left panel shows photometric observations (squares) as well as the reconstructed light curve (lines). The map of the star is given on a grid of 40 latitudes and 80 longitudes across the stellar surface. Taken from Korhonen et al. (1999) with modifications.

Table 2.1. Summary of Eclipsing Binaries

Type	Description
Algol-type eclipsing binaries (EA)	• named after β Per (Algol) • primary and secondary minima have striking different depth • beginning and end of eclipse well defined • semi-detached or detached systems with components having very different spectral types
β Lyrae eclipsing binaries (EB)	• primary and secondary eclipses both prominent and different • close proximity of components results in ellipsoidal shapes of the stars
W Ursae Major (UMA) eclipsing binaries (EW)	• periods shorter than 1 day • very tight systems • contact binaries • unlike EB, very small difference between primary and secondary eclipse
R-type binaries	• one component has a significant contribution from the reflected light of its companion

Table 2.2. Summary of Rotational Variables

Type	Description
AP stars, α^2 Canum Venaticorum (CVn) stars (ACV)	• partially slow rotators, periods 1/2 day to decades • chemical pecularities due to strong constant magnetic fields ($\sim$kG) stabilize stellar atmosphere
SX Arietis (SXA)	• B-type main sequence stars • strong magnetic fields, intense He I and Si III lines • high-temperature analogues of ACV
By Draconis (Dra) stars (BY Dra)	• cool main-sequence stars • variability caused by starspots and fast rotation • amplitude < 0.3 mag in V, $\mathcal{P} \lesssim 5$ days • long-term trends superimposed

Planetary Transits

When Johannes Kepler figured out that the motion of the Venus is predictable, he derived that Venus would pass in front of the Sun in 1631. This was the first prediction from what is called a *Venus transit*, or in general, *planetary transit*. Despite it was not observed, as the Sun was below the horizon as seen from Europe, later the calculations were proven to be right. Measurements of the Venus' motion across the Sun were made during the next transit in 1639. These measurements by Jeremiah Horrocks and William Crabtree (Hockey 2009), carried out from two different spots in England, enabled them to calculate the geometry between the Earth, Venus and the Sun.

Venus transits occur 4 times in 243 years: after 8, additionally 12.5, 8, and additionally 105.5 years. The reason for why this happens so rarely is because the planets are not exactly lined up at the same angle towards the Sun.

Nowadays, the transit method is one of the ways that astronomers discover planets orbiting stars other than the Sun, exoplanets. When a planet perfectly passes directly between us and a star, a drop in the star's brightness is detected. If such a brightness drop is detected at regular time intervals and lasts a fixed length of time, then it is very probable that a planet is orbiting the star and passing in front of it once every orbital period.

From the amount of decrease in brightness, the diameter ratio between the star and the planet can be directly calculated. As the size of the star is known with considerable accuracy, the light curve analysis gives a good estimate of the orbiting planet's diameter.

Exoplanet surveys are e.g. carried out by the Kepler mission (Borucki et al. 2010). The Kepler mission, launched in March 2009, has already detected thousands of planetary candidates, including several being Earth-sized and orbiting in their star's habitable zone.

Asteroids

Asteroids are small bodies of the solar system, having sizes in the range of meters to hundreds of kilometers. Most of them orbit in the so-called *main belt* between Mars and Jupiter, whereas some of them have orbits that come close to the Earth's orbit or even cross it, called near-Earth asteroids. Asteroids can be described as irregular solid solar system bodies without any atmosphere.

Similarly to planets, asteroids shine by light reflected from the Sun. The brightness being detected changes for several reasons: As the distance of an asteroid to the Sun and the Earth changes while it orbits the Sun, the brightness shows temporal change. Additionally, asteroids show brightness variations caused by their irregular shape and their rotation, exhibiting different parts of the surface. Also, eclipsing asteroids are possible.

The analysis of asteroid light curves gives information on the shape of the asteroid, its rotation period and the spin axis direction. So far, models for more than 900 asteroids have been derived

this way. They are stored in the Database of *Asteroid Models from Inversion Techniques* (DAMIT, Ďurech et al. 1999).

Fig. 2.7 shows the light curve of the asteroid 2867 Steins, one of the two asteroids that Rosetta was flying by on its way to comet 67 P/Churyumov-Gerasimenko. Asteroid 2867 Steins is one of two asteroids (the other being 21 Lutetia) that Rosetta was flying by on its way to comet 67 P/Churyumov-Gerasimenko. The light curve (Küppers et al. 2007) shows the variation of the asteroid's apparent magnitude as measured by the OSIRIS Narrow Angle Camera on March 11, 2006 from a distance of 1.06 AU. The data show the asteroid has a spin period of 6.052 hours, in good agreement with ground based observations.

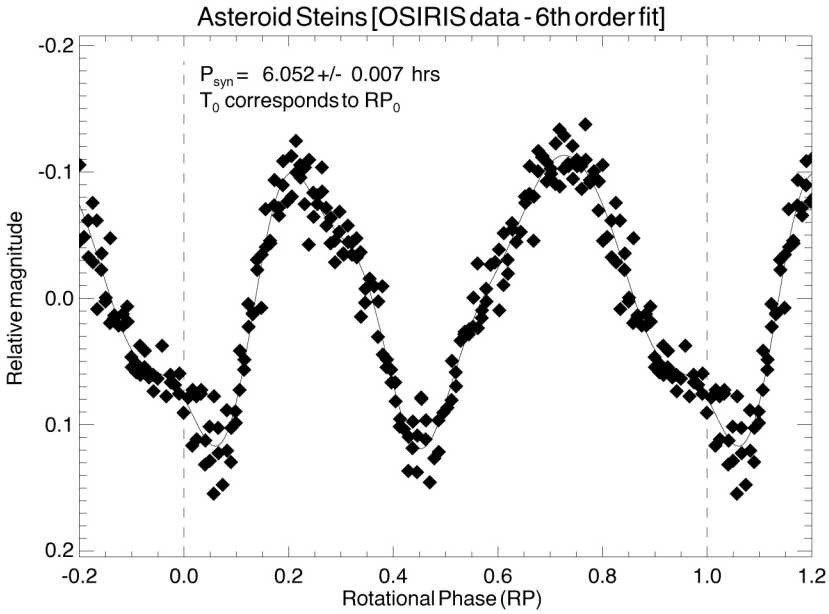


Figure 2.7 The light curve of the asteroid 2867 Steins, showing an amplitude of 0.1 mag as the asteroid is rotating. Taken from (Küppers et al. 2007).

2.3.2 Intrinsic Variables

In the case of intrinsic variables, light curve variability is caused by effects inherent to the source. Among intrinsic variable stars, one can distinguish between the sub-classes of pulsating and eruptive as well as cataclysmic variables. Other reasons for intrinsic variability of sources are e.g. accretion effects found in AGN.

Pulsating Variables

During most stages of their life, most types of stars are in a stable equilibrium. But there are certain stages in the life of stars where a stable equilibrium cannot be maintained. When this occurs, the star is radiating more than its average luminosity, causing the star's outer layers to expand. The luminosity increases, as the density of the layer decreases due to expansion,

Table 2.3. Summary of Pulsating Variables

Type	Description
RR Lyrae stars, RR Lyrae	• periods between 0.2 and 1 day • amplitudes of 0.5 to 1.5 mag • standard candles used to measure distances to systems containing old stellar population (e.g. Milky Way's halo) • for a more detailed description, see Section 2.5.1
δ Cephei stars, Cepheids	• periods between 1 and 5 days • amplitudes of 0.5 to 2 mag in V • standard candles, strong concentration towards the Galactic plane • for a more detailed description, see Section 2.5.2
Mira variables	• cool red giant star of spectral type Ke, Me, Se, or Ce • period of 100–1,000 days, amplitude > 1 mag in IR, 2.5–11 mag in V • strong stellar wind
δ Scuti stars	• amplitudes from 0.003 to 0.9 mag in V • period of a few hours • used as standard cndles • radial and non-radial pulsations

so it eventually cools, its ionization drops and it becomes more transparent to radiation. The expansion thus reduces the internal pressure, leading the star to contract by gravity. When the star contracts, the internal pressure will increase again to the point that it exceeds the gravitational force contracting the star. It then expands, increasing its luminosity, and the cycle repeats.

Stars showing this pulsating behavior are present in various subclasses throughout the Hertzsprung-Russell (HR) diagram. It turns out that there is a certain region in the HR diagram where stars having a combination of temperature and luminosity have the proper conditions for this pulsation mechanism. A schematic overview is shown in Fig. 2.9. Pulsating variables can be classified in terms of their radial or non-radial pulsation, their excitation mechanisms triggering the pulsations, as well as their evolutionary status in the HR diagram.

The subclass of radial pulsators includes RR Lyrae and Cepheids – the two classes among variable stars, the analysis within the thesis deals with – as well as Mira variables.

Table 2.3 lists the most important types of pulsating variable stars.

A detailed description of the underlying pulsation mechanisms is given in Section 2.4, as well as for RR Lyrae and Cepheids in Section 2.5.1 and 2.5.2.

Eruptive Variables

Eruptive variable stars vary in brightness because of processes associated with magnetic fields, such as flares that occur within the stellar atmosphere. The changes in luminosity coincide with mass outflow in the form of stellar wind, or interaction with outside interstellar medium.

Eruptive variable stars exhibit irregular or semi-regular brightness variations caused by material being erupted from the star. Eruptive variables include protostars – stars which haven't reached the *main sequence* yet – showing impressive flares, as well as giants and supergiants, who lose their matter relatively easily and may also experience eruptions.

Eruptions are well-known also in our Sun; Fig. 2.8 shows a prominence eruption on the Sun, where a giant eruption of solar material exploded off the surface of the Sun and is falling right back.

Table 2.4 lists the most important types of eruptive variable stars.

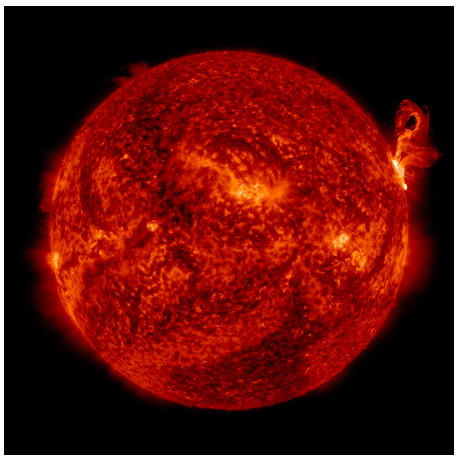


Figure 2.8 On March 2, 2012, a giant eruption of solar material exploded up off the surface of the Sun, as captured in this image from NASA's *Solar Dynamics Observatory*. Known as a prominence eruption, most of the material usually falls right back down on to the Sun. Credit: NASA/SDO.

Table 2.4. Summary of Eruptive Variables

Type	Description
UV Ceti stars (flare stars)	• flares range from radio to X-ray • flares can increase star's brightness by up to 6 mag in V • presumably young stars, preceeding T Tauri phase
T Tauri stars (TTS)	• identified by their optical variability and strong chromospheric lines • pre-main-sequence stars in the process of contracting to the main sequence along the Hayashi track • large areas of starspot coverage, and they have intense, variable X-ray and radio emissions, often powerful stellar winds
FU Orionis (FUOr) stars	• pre-main sequence stars undergoing rapid accretion • brightness increase by $\gtrsim$4 mag, slow decline • G-type supergiants, K–M giants/supergiants
Ex Lupi (EXor) stars	• amplitude of 1 to 4 mag in V • outbursts lasting 10 to 100 days, separated by several months • emission lines similar to T Tauri stars
Wolf-Rayet stars (WR)	• strong broad emission lines of highly ionized He, N, C • very high surface temperatures of ∼30,000–200,000 K • highly luminous, $\sim 1000 \times L_\odot$
Luminous blue variable (LBV) stars	• unstable supergiant stars • periodic outbursts, occasionally larger eruptions • temperature between 10,000 K and 25,000 K • luminosity of $250,000 - 10^6 L_\odot$
Herbig Ae/Be stars	• pre-main-sequence star • spectral type earlier than F0 • show Balmer emission lines • show IR radiation excess due to circumstellar dust
R Coronae Boralis (R CrB) stars	• luminosity varies in two modes: low ampluitude pulsation with a few tenths of a mag, irregular with fading by 1 to 9 mag • supergiants in the spectral classes F and G
γ Cassiopeiae (Be) stars	• spectra vary over time • non-supergiant stars with temperatures between 10,000 and 30,000 K

Cataclysmic and Nova-like Variables (CV)

Cataclysmic variables are contact binaries consisting of a white dwarf and a low-mass main-sequence star (K or M dwarf). The latter one, called *secondary* (in contrast to the *primary* white dwarf) has filled the Roche volume, resulting in mass transfer onto the primary.

There are various subclasses of cataclysmic variables, mostly characterized by their magnetic fields.

When an accretion disk develops due to the mass transfer, the cataclysmic variable is called a *dwarf nova*. Also, among them there are various sub-classes depending on frequency and intensity of outbursts.

The class of nova-like stars also includes supernovae. Initially assuming they are new stars (Hockey 2009), they are the most noticeable class of variable stars. Supernovae occur during the last stellar evolutionary stages of a massive star's life. For a short time of a few days, this causes the appearance of a seemingly 'new' star with an apparent brightness of up to -6 to -7.5 mag, depending on their distance. Then, it slowly fades from sight over weeks to months.

Supernovae can be classified according to their light curves and the absorption lines of different chemical elements that appear in their spectra. The first element for division is the presence or absence of a H line. If a supernova's spectrum contains H lines, it is classified type II; otherwise it is type I. In each of these two types there are subdivisions according to the presence of lines from other elements or the shape of the light curve. The most important one is type Ia.

Type Ia supernovae show a line at 615.0 nm caused by singly ionized silicon (Si II). They happen when a white star as part of a binary system accretes matter from its companion and thus reached its *Chandrasekhar limit* of 1.4 $M_\odot$. When this mass limit is reached, the star becomes unstable and undergoes a thermal runaway nuclear fusion reaction. The fact that all type Ia supernovae explode at about the same mass results into a very narrow range of absolute magnitudes. This makes them very useful as standard candles.

The blue and visual peak magnitudes of type Ia supernovae are given as (Hillebrandt and Niemeyer 2000):

$$M_B \approx M_V \approx -19.3 \pm 0.3 \text{ mag}\,. \tag{2.1}$$

Supernovae of type Ia are crucial in establishing the cosmological *distance ladder* to extragalactic distances. They had been deciding in discovering the accelerated expansion of the Universe. Supernovae are the source of many elements, especially the ones heavier than Fe, which are produced in supernovae and ejected out into space.

Table 2.5 lists properties of supernovae, as well as cataclysmic variables and nova-like variables in general.

Table 2.5. Summary of Cataclysmic Variables

Type	Description
Supernovae (SN)	• explosive event occurring at the last evolutionary stages of a massive star • seemingly 'new' star with an apparent brightness of up to -6 to -7.5 mag • supernovae are classified according their light curves and absorption lines • used as standard candles
Novae	• close binary system in which a white dwarf accretes matter from its companion • nova results of the rapid fusion of the accreted H on the white dwarf's surface • steep rise to peak, steadily decline • brightens by >12 mag, decays over ∼25–80 days by 2 mag
Dwarf Novae (Geminorum-type variable star)	• close binary system in which a white dwarf accretes matter from its companion • luminosity effects attributed to changes in the accretion disk • depending on sub-type, one or multiple outbursts can happen
Recurrent Novae	• same mechanism as for novae • at least 2 outbursts over the past century, intervals 10–100 years • brightens by 8–9 mag during outburst • currently 10 recurrent novae known

2.4 Pulsation as Cause of Variability

Despite many variable stars were known by the beginning of the 20th century, the cause of their variability was not understood until this time. Historically, among the first assumptions for the cause of variability were mostly eclipse or rotational models, pointing towards extrinsic causes for variability.

The observation by Bélopolski (1895) of a radial velocity change during the light cycle of δ Cephei (Hockey 2009; Catelan and Smith 2015) was actually an indicator for pulsation, but was misinterpreted as an indicator for a binary star. However, the light curve shape didn't match to known binaries. In 1900, Schwarzschild (Schwarzschild 1900) found the change of color and brightness in the Cepheid η Aquilae. This behavior doesn't fit to binaries, but was also not understood. The assumption of pulsation was brought up in 1914 by Harlow Shapley (Shapley 1914), making the binary hypothesis more unlikely by several arguments.

In the following, an overview is given on the structure and evolution of stars, and the driving mechanisms behind the pulsation of stars.

2.4.1 Stellar Structure and Evolution

Nowadays, the theory of stellar structure and evolution is based on equations describing the hydrostatics (or hydrodynamics) of the stellar interior. Major contributions were made by Atkinson (1931), Bethe (1939), Bethe and Marshak (1939) and Gamov (1939).

The description of stellar structure and evolution below is based mostly on Catelan and Smith (2015).
In the following, a spherically symmetric, self-gravitating star is assumed. The properties of the stellar matter at any point of the stellar interior are described with density ρ, temperature T, pressure P, entropy per unit mass S, coefficient of thermal conductivity per unit volume λ, and the chemical composition, based on the abundance of elements X_i. The mass of the star is assumed to be M, its radius R.

The basic set of equations of a star in hydrostatic and thermal equilibrium then consists of the following four time-independent differential equations (Kippenhahn and Weigert 1990). Of them, the first two describe the mechanical structure of the star, and the last two its energetic and thermal structure:

$$\frac{\partial r}{\partial M_r} = \frac{1}{4\pi r^2 \rho} \tag{2.2}$$

$$\frac{\partial P}{\partial M_r} = -\frac{GM_r}{4\pi r^4} \tag{2.3}$$

$$\frac{\partial L}{\partial M_r} = \epsilon - \epsilon_\nu - \epsilon_g \tag{2.4}$$

$$\frac{\partial T}{\partial M_r} = -\frac{GM_r T}{4\pi r^4 P}\nabla \tag{2.5}$$

with

M_r: stellar mass enclosed in radius r of a star with total mass M (2.6)

L: luminosity in units of $\mathrm{J\,s^{-1}}$ (energy per unit time) (2.7)

ϵ: energy generation rate in the form of thermonuclear reactions (2.8)

ϵ_ν: energy loss rate in the form of neutrinos, important in late stages of evolution (2.9)

ϵ_g: work performed on the gas during any expansion or contraction of the star, (2.10)

i.e., the total heat flux through a spherical shell with radius r, is given as luminosity $L = 4\pi r^2 F = -4\pi r^2 \lambda \partial T/\partial r$.

The temperature gradient $\nabla \equiv \frac{\partial \ln T}{\partial \ln P}$ in Equ. (2.5) depends on the modus of energy transport, being primarily radiative (∇_{rad}), conductive (∇_{c}) or convective (∇_{conv}).

In the case of radiative transport, ∇ takes the form

$$\nabla = \nabla_{\mathrm{rad}} = \frac{3}{16\pi acG}\frac{\kappa_R L P}{M T^4} \tag{2.11}$$

where $a = \frac{4\sigma}{c} = 7.5657 \times 10^{-16}\ \mathrm{Jm^{-3}K^{-4}}$ is the radiation constant, with the Stefan-Boltzmann constant σ. The Rosseland mean opacity κ_R (per unit mass) is defined as

$$\kappa_R = \frac{\int_0^\infty \frac{1}{\kappa_\nu}\frac{\partial B_\nu(T)}{\partial T}\,\mathrm{d}\nu}{\int_0^\infty \frac{\partial B_\nu(T)}{\partial T}\,\mathrm{d}\nu} \tag{2.12}$$

with the coefficient of monochromatic radiative opacity (per unit mass) $\kappa_\nu = \frac{4acT^3}{3\rho}\frac{1}{c/\nu}$, and the monochromatic Planck function B_ν (Catelan and Smith 2015).

Convective regions in the interior of a star are identified by the Ledoux criteria (Ledoux 1947),

$$v_{\rm rad} > v_{\rm ad} - \frac{\chi_\mu}{\chi_T}\nabla_\mu \tag{2.13}$$

with the molecular weight μ and

$$\chi_\mu \equiv \left(\frac{{\rm d}\ln P}{{\rm d}\ln \mu}\right)_{\rho,T} \tag{2.14}$$

$$\chi_T \equiv \left(\frac{{\rm d}\ln P}{{\rm d}\ln T}\right)_{\rho,\mu} \tag{2.15}$$

$$\nabla_\mu \equiv \left(\frac{{\rm d}\ln \mu}{{\rm d}\ln P}\right). \tag{2.16}$$

In the absence of a chemical composition gradient ∇_μ, this reduces to the Schwarzschild criterion (Schwarzschild 1906) for the onset of convection $\nabla_{\rm rad} > \nabla_{\rm ad}$.

In the above, partial derivatives have been used to emphasize the non-stationary nature of the physical solutions. They evolve over time as a consequence of the nuclear processes in the interior of the star.

The fact that stars are radiating away energy, because they are luminous, implies they are not stationary. Also, due to the nuclear processes providing the energy, their chemical composition has to change over time. Therefore, the equations above have to be supplemented by a set of equations to describe the evolution of the abundance $\partial X_i/\partial t$ and thus introducing a nuclear time scale:

As the energy released by fusing a mass ΔM of H into He is $\sim$0.007 $\Delta M c^2$, the time until the H is exhausted, given the star's current luminosity, will be:

$$t_{\rm nuc} = \frac{0.007\ \Delta M c^2}{L}, \tag{2.17}$$

which is $\approx 10^{10} - 10^{11}$ yr for our Sun.
However, the actual lifetime of a star is only one tenth of $t_{\rm nuc}$ because it changes its luminosity to become brighter during its evolution.

In the following, the Equations (2.2) to (2.5) are modified to describe the time evolution of a spherical symmetric star having a given distribution of chemical abundances $X_i(M_r)$.

In the case of a pulsating star, the system is outside hydrostatic equilibrium, as there is no longer perfect balance between pressure gradient and gravity. In this case, Equ. (2.2) to (2.5) changes as the mass element will undergo acceleration, and becomes:

$$\frac{\partial r}{\partial M_r} = \frac{1}{4\pi r^2 \rho} \tag{2.18}$$

$$\frac{\partial P}{\partial M_r} = -\frac{GM_r}{4\pi r^4} - \frac{1}{4\pi r^2}\frac{\partial^2 r}{\partial t^2} \tag{2.19}$$

$$\frac{\partial L_r}{\partial M_r} = \epsilon - T\frac{\partial S}{\partial t} = \epsilon - \epsilon_\nu - \epsilon_g \tag{2.20}$$

$$\frac{\partial T}{\partial M_r} = -\frac{3\kappa L_r}{64\pi^2 acT^3 r^4}. \tag{2.21}$$

Here, three kinds of derivatives appear:

- $\partial_r^2/\partial t^2$ in Equ. (2.19) describes hydrodynamical changes to the stellar structure. These changes occur on the dynamical time scale $\tau_{\rm dyn} = \left(\frac{R^3}{GM}\right)^{1/2} \simeq (G\bar{\rho})^{-1/2}$.
- $T\partial S/\partial t$ in Equ. (2.20) which is often written as an additional energy generation term ϵ_g:

$$\epsilon_g = -T\frac{\partial s}{\partial t}. \tag{2.22}$$

This term describes changes to the stars thermal structure, resulting from contraction ($\epsilon_g > 0$) or expansion ($\epsilon_g < 0$). Such changes occur on the Kelvin-Helmholtz timescale

$$t_{\rm KH} = \frac{GM^2/R}{L}. \tag{2.23}$$

The Kelvin-Helmholtz time scale indicates the time required to radiate the current gravitational binding energy of the Sun at its current luminosity; this is the timescale on which the Sun would contract if its nuclear energy sources were turned off. For our Sun, $t_{\rm KH} \sim 3 \times 10^7$ yr.

As $\tau_{\rm nuc} \gg \tau_{\rm KH} \gg \tau_{\rm dyn}$, changes in the stellar chemical abundance occur on much larger timescales than the dynamical timescale.

If the time derivative in Equ. (2.19) vanishes, the star is in hydrostatic equilibrium. If the time derivative in Equ. (2.20) vanishes, the star is in thermal equilibrium.

It is always assumed – no matter if the star is in hydrostatic or thermal equilibrium – that the conditions of a local thermodynamic equilibrium are satisfied.

These equations have to be supplemented with boundary conditions (Catelan and Smith 2015). The boundary conditions for the differential equations of stellar evolution constitute an important part of the overall problem. At the stellar center, two adjustable parameters exist: The central density ρ_c, and the central temperature T_c. At the stellar surface, there are also two adjustable parameters: the stellar luminosity L and the radius R.

Central boundary conditions

At the center ($r = 0$) of the star, the enclosed mass M_r, the radius r and the luminosity L_r have to vanish and the energy generation rate must remain finite. Therefore, both M_r and L must vanish at the center:

$M_r = 0$, $L_r = 0$ at $r = 0$.

As nothing is known a priori about the central values of P and T, the remaining two boundary conditions must be specified at the surface rather than at the center.

Surface boundary conditions

At the surface ($M_r = M$), the boundary conditions are generally much more complicated than at the center. The simplest option is to take the zero boundary conditions $T = 0$ and $P = 0$ at the surface. However, in reality, T and P never become zero because the star is surrounded by an interstellar medium with low, but finite density and temperature. Another option is to set the outer boundary conditions to $\rho = 0$, $T = \left(\frac{L}{8\pi R^2 \sigma}\right)^{1/4}$, where σ is the Stefan-Boltzmann constant.

A more realistic option is to identify the surface with the star's photosphere, which is where the bulk of the radiation escapes and which corresponds to the visible surface of the star. The photospheric boundary conditions approximate the photosphere with a single surface at optical depth $\tau = 2/3$ (Catelan and Smith 2015). One can write

$$\tau_{\rm ph} = \int_R^\infty \kappa\rho \, {\rm d}r \approx \kappa_{\rm ph} \int_R^\infty \rho \, {\rm d}r, \tag{2.24}$$

where $\kappa_{\rm ph}$ is an average value of the opacity over all layers above the photosphere. Assuming the atmosphere is geometrically thin, thus its extent is very small compared to R, it follows $\frac{{\rm d}P}{{\rm d}r} = -\frac{GM}{r^2}\rho \Rightarrow P(R) \approx \frac{GM}{R^2}\int_R^\infty \rho \, {\rm d}r$. Since $\tau_{\rm ph} = 2/3$ and $T(R) \approx T_{\rm eff}$, the boundary conditions can be written as

$$M_r = M: \; P = \frac{2}{3}\frac{GM}{\kappa_{\rm ph} R^L}, \; L = 4\pi R^2 \sigma T^4. \tag{2.25}$$

The Hertsprung-Russell Diagram

In 1912, two astronomers – Ejnar Hertzsprung and Henry Norris Russell – found independently that when stars are plotted accordingly their temperature and luminosity, the majority of them fall on a smooth curve (Hertzsprung 1911). The properties plotted originally were properties which can be determined observationally, e.g. the absolute visual magnitude M_V vs. the $B - V$ color index. The resulting diagram turned out to be a significant tool to understand stellar evolution, and was thus named after Hertzsprung and Russell.

Fig. 2.9 shows a Hertzsprung-Russell diagram (HR diagram) in $\log L/\mathrm{L}_\odot$ vs. $\log T_{\mathrm{eff}}$ with common types of stars. Also, different types of pulsating stars within the region of instability, the *instability strip*, are shown.

The basic HR diagram is a luminosity vs. temperature graph. The temperature may be replaced or supplemented with spectral class or color index. The main spectral classes in order from hottest to coolest are O, B, A, F, G, K and M. Within the HR diagram, the stars which lie along the nearly straight diagonal line are known as *main sequence stars*. Those stars are still burning hydrogen in their cores. The main sequence accounts for about 90 percent of the stellar population above 0.5 $\mathrm{M}_\odot$ (Arnett 1996).
There is a correlation between a main sequence star's mass and luminosity. Stars that are high up on the main sequence are more massive. The relation, the *mass-luminosity relation* (Catelan and Smith 2015), says that $L \propto M^{3.5}$.
Stars younger and older than main sequence stars are called *pre- and post-main sequence stars*. Stars that have evolved well beyond the main sequence are often on the *red giant branch* of the HR diagram, or might be *asymptotic giant branch* stars. RR Lyrae are found on the *horizontal branch*, or β Cephei stars on the *upper main sequence*.
The cooler, dimmer stars are found towards the lower right of the HR diagram, and the hotter, more luminous stars in the upper left. Old red giants are found at the *red giant branch* (RGB).
Our own star, the Sun, is located nearly in the middle of both the temperature and luminosity scales relative to other stars.
Because of the relation $L = 4\pi R^2 \sigma T^4$, stars above the main sequence (having higher luminosity, with the same temperature as cooler main sequence stars) have larger radii. Also, stars having the same luminosity as dimmer main sequence stars, but are to the left of them (being hotter), have smaller radii.

Stellar Evolution and the Instability Strip

One of the key concepts of modern astronomy is that stars change over time – they are born from clouds of interstellar gas and dust, they shine over billions of years by light created through nuclear fusion of hydrogen (H) in their cores, and eventually run out of their nuclear fuel and die. During their last stages, they return some of their mass back to interstellar space, that will be taken up into new generations of stars. The process of change a star undergoes during its lifetime is called *stellar evolution*.

As such processes take millions to billions of years for a star, we can't observe them directly. However, there are many pieces of evidence that formed the current understanding of stellar evolution. One was the understanding of the nuclear physics responsible for why stars shine over such a long time, and the subsequent realization that their fuel is nuclear, and thus large but finite.
Another piece of evidence was the observational study of star clusters – groups of stars all born at the same time and place, and thus from the same composition under the same conditions –

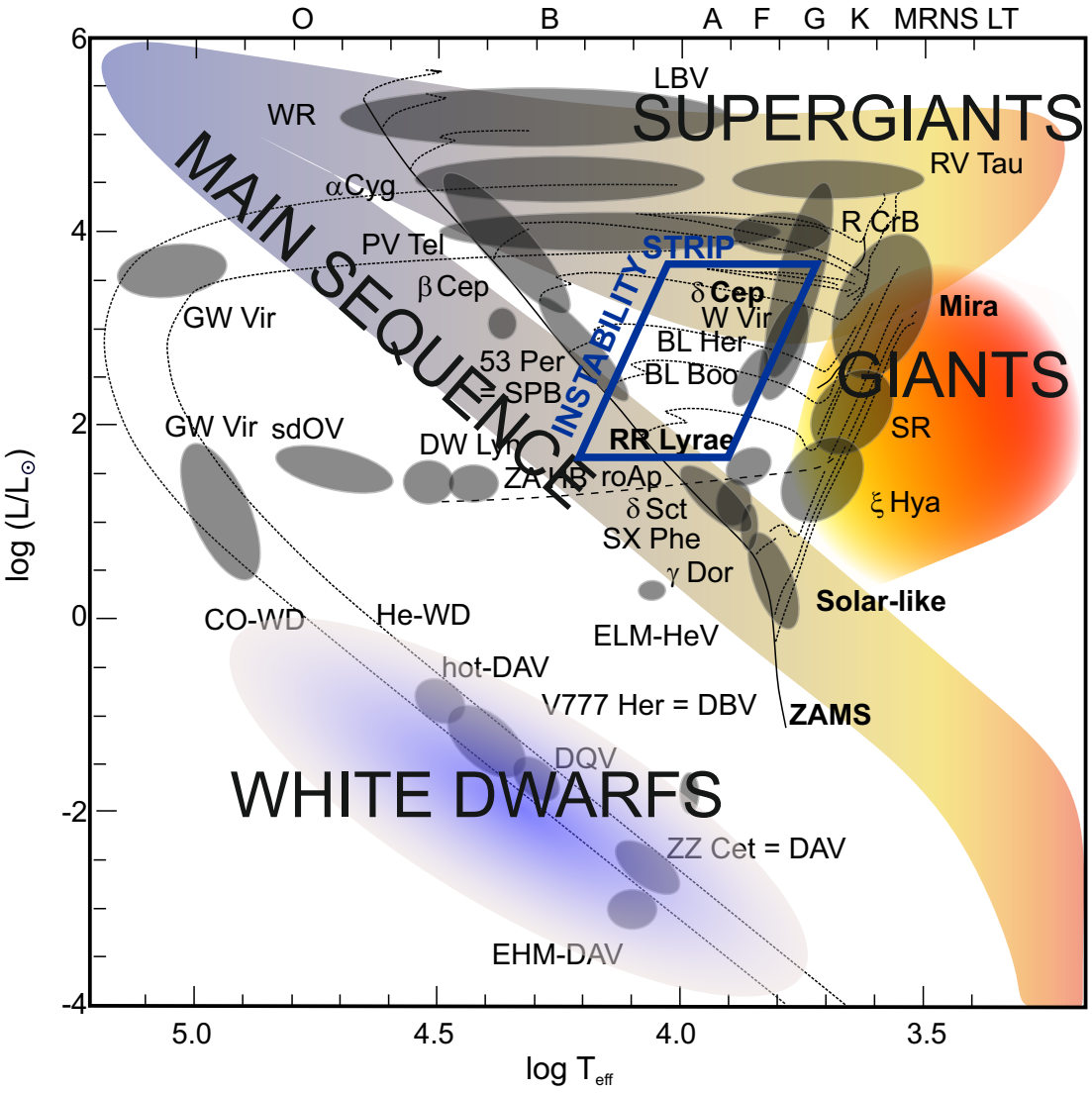


Figure 2.9 Schematic distribution of different types of pulsating stars across the Hertzsprung-Russell diagram. The instability strip, including RR Lyrae and Cepheids, is indicated. Own figure; main features are taken from Catelan and Smith (2015).

and the eventual realization that the properties of star clusters differ depending upon how old they are.

During their evolution, almost all stars have phases of instability, leading to variability in their light curves. Above, theoretical models of stellar structure and evolution and statements for instability were given. Here, the problem is described and discussed from a more observational view.

Stars being in a state of instability can be found in an area of the HR diagram called the *instability strip* (Gautschy and Saio 1996). This is an area around 1000 K wide just above the main sequence, as shown in Fig. 2.9. It includes Cepheid variables where it intersects the supergiants, RR Lyrae variables where it crosses the horizontal branch, as well as δ Scuti stars, rapidly oscillating Ap (roAp) stars and others near the main sequence.

Stellar Mass and Stellar Evolution

The HR diagram indicates that a star of a given brightness could only lie within a certain range of colors, and a star with a given color could only be found within a certain range of brightnesses. More observational and theoretical research showed that the HR diagram represents a snapshot of the evolutionary states of the stars within the diagram. As a star evolves, it changes in brightness and color in a very predictable way, and stars of different masses change in very different ways. The progress of a star's life is predetermined by its mass, because mass is what determines the amount of energy being produced and how fast its evolution will be.

In the HR diagram, four large regions can be identified (Catelan and Smith 2015):

- red dwarfs: $M < 0.7\ \mathrm{M}_\odot$, their main sequence lifetime exceeds age of the Universe
- low-mass stars: $0.7\ \mathrm{M}_\odot < M < 2\ \mathrm{M}_\odot$, they end lives as white dwarfs and possibly planetary nebulae
- intermediate-mass stars: $2\mathrm{M}_\odot < M < 8 - 10\ \mathrm{M}_\odot$, similar to low-mass stars, but higher L, they end as higher mass white dwarfs and planetary nebulae
- high-mass stars: $M > 8 - 10\ \mathrm{M}_\odot$, distinctly different evolution paths, end as supernovae, leaving neutron stars or black holes.

The age of a star tells which evolutionary stages it has passed. Both of this quantities are hard to measure directly, but are related to temperature and luminosity. Within the HR diagram, evolutionary tracks can be identified. Fig. 2.10 to 2.12 show them depending on the mass ranges specified above. The following description of the evolutionary tracks of stars as depending on their mass is oriented on Catelan and Smith (2015), with the evolutionary tracks therein based on simulations by A. V. Sweigart using a 1D hydrostatic code.

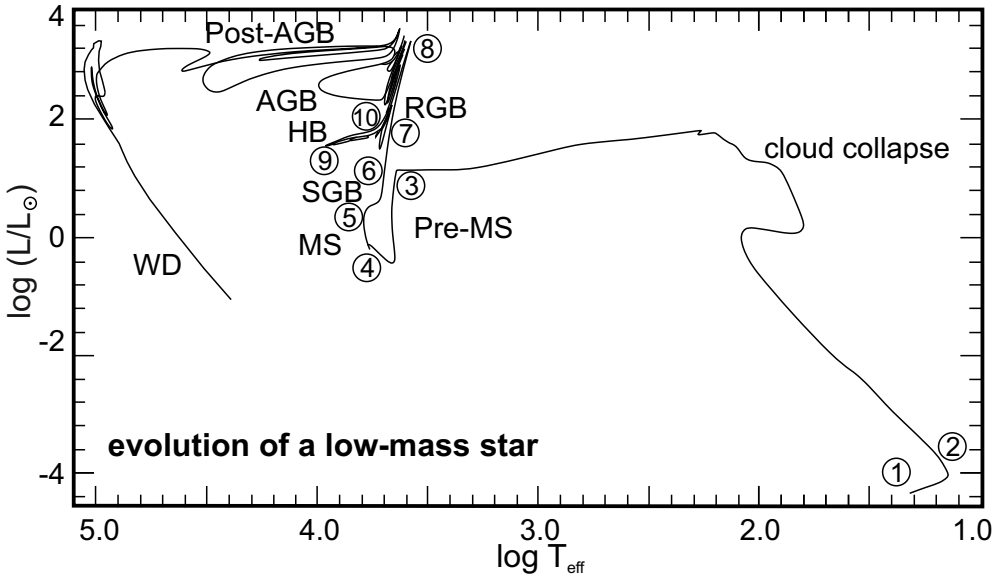
Low-mass stars (Fig. 2.10) start from a molecular cloud that becomes unstable according to the Jeans criterion for instability (Jeans 1902) and thus collapses and fragments (Bodenheimer 2003). The temperature at this early phase (isothermal phase) is of order 10 K and the cloud is optically thin.
Eventually the opacity increases and the temperature rises, forming a proto-star(**1**). A hydrostatic core forms after 1.5×10^5 years (Wuchterl and Klessen 2001)(**2**), luminosity and temperature increase steadily. When accretion stops, the photosphere of the proto-star becomes visible(**3**). The star becomes fully convective, and luminosity decreases at about constant temperature, leading to the vertical segment of the evolutionary track, the *Hayashi track*. After this point, the proto-star becomes hotter and increases in brightness, while shrinking in size. Increasing core temperature leads to incomplete CNO processing, with ${}_{12}\mathrm{C}$ being consumed into the core. Until ${}^{12}\mathrm{C}$ is exhausted, additional 5×10^7 years will have passed. The star has now reached the zero-age main sequence (ZAMS)(**4**). The star evolves along the much longer nuclear time scale, which is of order 10^{10} years for low-mass stars. Temperature and luminosity slightly increase with time as H is steadily transformed into He. When the star reached the turn-off point(**5**), H burning stops to be

a central process and instead becomes a shell-burning process. Not all the energy released by the H-burning shell reaches the surface: part of it expands the star's envelope. The star begins to cool down and enters the subgiant phase. When the convective envelope reaches its maximum inwards extension after additionally 10^9 years(6), the dredge-up of material occurs. This material has been partially processed nuclearly and includes a small amount of He. The H-burning shell continues to advance outward in mass(7), thus leading to a continued increase of mass in the He core. The H-burning shell actually encounters the chemical composition discontinuity that was left behind from the dredge-up phase. The remaining time on the RGB(8) is characterized by the increasing size of the He core, which keeps on contracting and heating up. At this stage, large amounts of energy are lost in the form of neutrinos. When temperature finally becomes high enough for the He-burning reactions, this happens in a shell inside the He-rich core, leading to the *helium flash*. The zero-age horizontal branch (ZAHB)(9) marks the onset of the He-burning phase. In the case of low-mass stars, it is referenced to as the horizontal branch (HB) phase, because HB stars have very nearly the same luminosity irrespective of mass. When He has been exhausted at the center after additionally 10^8 years(10), the star continues as a low-mass asymptotic giant branch (AGB) star. An AGB star has an inert core comprised mostly of C and O, that burns He in a shell and H in another shell further out. The onset of He-shell burning causes a temporary reversal in the star's evolutionary direction. This leads to the AGB clump in the observed color-magnitude (CM) diagrams of well-populated globular clusters and old Local Group galaxies. The details of the AGB phase, which lasts of order 10^7 years, depend strongly on the poorly known mass loss rates. Once the AGB star's envelope mass has become very low, a final mass ejection phase may take place, the so-called *superwind phase*, leading to the formation of a post-AGB star. After a quick evolution to the blue, the star finally settles on the white dwarf (WD) cooling sequence.

The evolutionary track of **intermediate-mass stars** is shown in Fig. 2.11 based on a 5 $M_\odot$ rotating star model from Ekström et al. (2012).
The star is evolving along the ZAMS(1), until the amount of H in the core becomes insufficient to support the structure of the star through nuclear reactions. Thus the star contracts(2). He is exhausted in the core(3), so a H-burning shell-narrowing phase follows. The convective envelope extends inward(4). When the star reaches the base of the RGB, dredge-up of nuclearly-processed material toward the stellar surface happens(4). After He ignition(5), it goes to the "blue loop" (6) and finally forms a He-burning shell(7).
The basic principles shown for low-mass stars apply also here. The main difference between low- and higher-mass stars is whether or not ignition of core He-burning occurs under degenerate conditions. In low-mass stars, the He core becomes highly degenerated by the time the triple-α process is ignited at the tip of the RGB, leading to the helium flash. In turn, in intermediate- and high-mass stars, He-burning commences long before degeneracy happens in the stellar core. This results into an actual RGB sequence being much shorter than for low-mass stars.
Another difference between low- and high-mass stars is the presence of convective cores in the latter due to the fact that temperatures are higher in the core of high-mass stars. A direct consequence of the presence of fully mixed convective cores is that, by the time the core H is

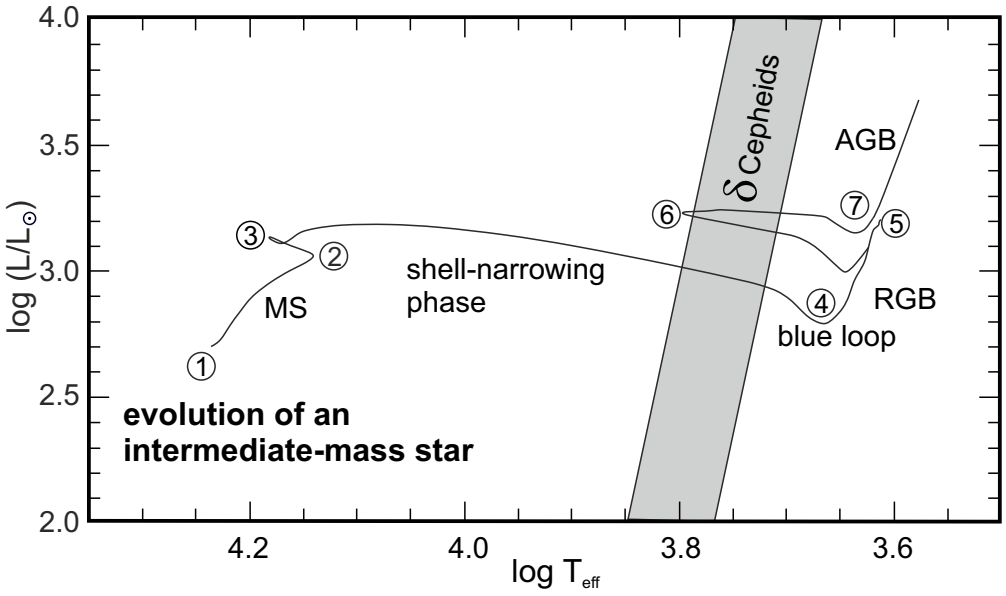
exhausted, the region that becomes devoid of its nuclear fuel is significantly larger than for low-mass stars. In intermediate-mass stars, the core He-burning phase is characterized by prominent "blue loops". They can cross the Cepheid instability strip and thus give raise to the classical Cepheids.
As in the case for low-mass stars, dredge-up occurs, but here, a second and third dredge-up phase is possible. The second dredge-up, which takes place in stars more massive than 3 $M_\odot$, is analogous to the first. The third, however, is unique to AGB stars, and is intimately related to thermal pulses.

The evolutionary track of a **high-mass star** is shown in Fig. 2.12 based on a 40 $M_\odot$ rotating star model from Ekström et al. (2012).
The star is evolving along the ZAMS(**1**), until the amount of H in the core is not able any longer to support the structure of the star through nuclear reactions. This leads the star to contract (**2**). He is exhausted in the core(**3**), so a H-burning shell-narrowing phase follows. The convective envelope extends inward(**4**). When the star reaches the base of the RGB, dredge-up of nuclearly-processed material toward the stellar surface sets on(**4**). The C-burning phase(**5**) starts.
Even more so than in the case of intermediate-mass stars, the details of the evolution, and especially the final stages, are affected by the assumptions regarding mass loss and overshooting from the convective core.
What distinguishes high-mass stars clearly from intermediate-mass stars is the final product of their evolution. For the same assumptions regarding mass loss, chemical composition and rotation, stars above a certain initial mass of $8 - 10$ $M_\odot$ will not produce white dwarfs, but neutron stars or black holes in their final stage.
Depending on the mass of the star, nuclear fuel burning may proceed all the way to the Si-burning phase. The duration of each such phase is dramatically shorter than the previous one. By comparing the evolutionary tracks in Fig. 2.10 to 2.12, one can notice that some high-mass stars are expected to spend a significant fraction of their lives as either PV Tel, α Cyg or LBV stars. High-mass stars, such as the one depicted in Fig. 2.12, spend quite little time, if any, as red supergiant. As a consequence, one expects an upper limit for the masses of red variables: for example, the so-called *ultra-long period Cepheids* with periods longer than about 80 days may have masses reaching up to 15 or 20 $M_\odot$. In contrast, lower-mass Cepheids likely have masses in the range between 2.5 and 4.5 $M_\odot$.
After the shell-narrowing phase(**3**)(**4**), high-mass stars evolve back to blue. While doing so, the lower-mass ones are missing entirely the red supergiant phase. In contrast, stars with masses higher than 20 to 32 $M_\odot$ may become Wolf-Rayet stars.



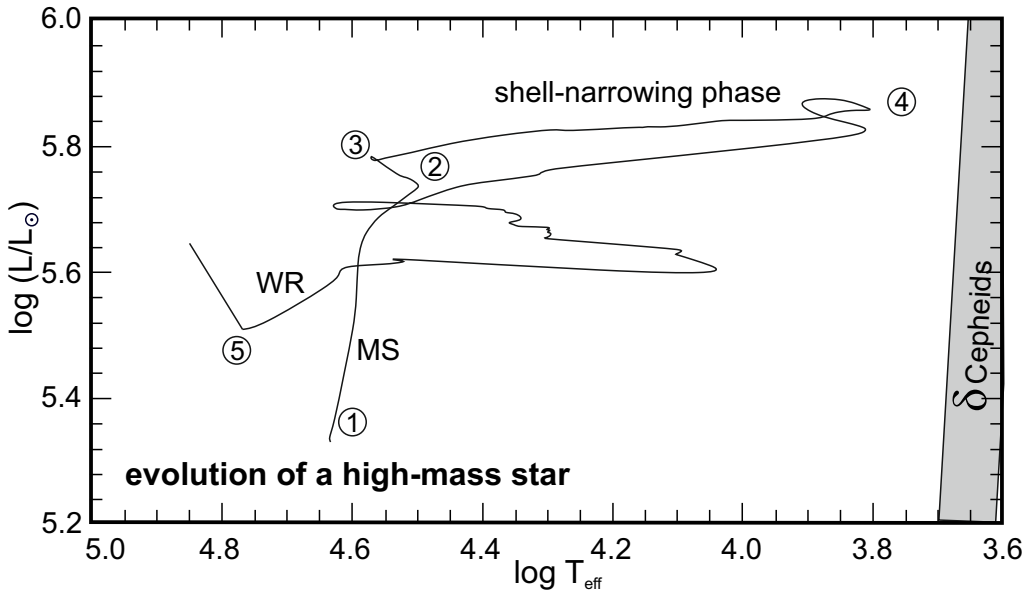
(a)

Figure 2.10 The evolution of a low-mass star in the Hertzsprung-Russell diagram. The numbers correspond to specific episodes in the life of the star, as described in the text. Adapted from Catelan and Smith (2015).



(a)

Figure 2.11 The evolution of an intermediate star in the Hertzsprung-Russell diagram. The numbers correspond to specific episodes in the life of the star, as described in the text. Adapted from Catelan and Smith (2015).



(a)

Figure 2.12 The evolution of a high-mass star in the Hertzsprung-Russell diagram. The numbers correspond to specific episodes in the life of the star, as described in the text. Adapted from Catelan and Smith (2015).

2.4.2 Stellar Pulsation Theory

Stars on the instability strip pulsate due to He III (double-ionized helium). Whereas He is neutral in the photoshere of A-F-G stars, deeper below the photosphere, at about 25,000 - 30,000 K, the He II layer (single-ionized helium) begins (Catelan and Smith 2015).
When the star contracts, the density and temperature of the He II layer increases. He II starts to transform into He III at about 35,000 - 50,000 K. The opacity of this layer increases due to the ionization, and thus the energy flux from the interior of the star is effectively absorbed. The star expands, and the temperature rises. After expansion, He III begins to recombine into He II and the opacity of the star drops. This lowers the surface temperature of the star. The outer layers contract and the pulsation cycle starts from the beginning.

Between a star's radial pulsation and brightness variations, a phase shift can be observed. For most Cepheids, this creates a distinctly asymmetric light curve, rising rapidly to maximum and falling slowly back down to minimum, see Fig. 2.14. This phase shift is caused by the distance of the He II zone from the stellar surface.

The variations of these stars could be understood in terms of pulsations in the first radial mode, where the star expands and contracts while preserving its spherical symmetry. It was realized by Shapley (1914), that the period is approximately given by the dynamical time scale of the star:

$$\tau_{\mathrm{dyn}} \simeq \left(\frac{R^3}{GM}\right)^{1/2} \simeq \sqrt{G\bar{\rho}}, \tag{2.26}$$

where R is the radius of the star, M its mass, $\bar{\rho}$ is the mean density and G is the gravitational constant.
Later on, major contributions to the understanding of stellar pulsations were made by Eddington (Eddington 1926). However, the identification of the actual cause of the pulsations, and of the reason for the distinct instability strip, was first arrived at independently by Zhevakin (1953) and by Cox and Whitney (1958).
The pulsation of a star, being a hydrodynamic phenomenon, should take place on the dynamical time scale Equ. (2.26) roughly equal to the sound-crossing time. The speed of sound is given by

$$v_s = \sqrt{\frac{\Gamma_1 P}{\rho}}, \tag{2.27}$$

where Γ_1 is the *first adiabatic coefficient*, $\Gamma_1 \equiv 1 + \left(\frac{\partial \ln p}{\partial \ln \rho}\right)_s$.

Assuming an ideal gas equation of state, this becomes

$$v_s = \sqrt{\frac{\Gamma_1 k_B T}{\mu m_{\mathrm{H}}}}, \tag{2.28}$$

where for ionized matter, the mean molecular weight μ is represented by $\mu^{-1} = 2\mathrm{X} + \frac{3}{4}\mathrm{Y} + \frac{1}{2}\mathrm{Z}$ with the mass fractions X (of H), Y (of He) and Z (of all other elements, so-called "metals").
For Γ_1, the ratio of specific heats for an ideal monoatomic gas, $\gamma = 5/3$, is adopted. A reasonable approximation for the temperature in Equ. (2.28) is provided by the temperature of the second ionization of He of around 35,000 - 55,000 K (Zhevakin 1963; Christy 1966). This results into $v_s \approx 32.2$ km s^{-1}.
Following Catelan and Smith (2015), the timescale for the propagation of a sound wave through the interior of a Cepheid can be assumed to be given by $\mathcal{P} \sim 2R/v_s$. This equation can be rewritten in terms of the star's gravitational potential energy Ω, $\Omega \sim -\frac{GM^2}{R}$ by using the virial theorem. In its simplest form, the virial theorem can be written as

$$\Omega = -3 \int_M \frac{P}{\rho} \, \mathrm{d}M_r. \tag{2.29}$$

Inserting (2.27) into (2.29) gives

$$\Omega = -3 \int_M \frac{v_s^2}{\Gamma_1} \, \mathrm{d}M_r = -3 \frac{\int_M \frac{v_s^2}{\Gamma_1} \, \mathrm{d}M_r}{\int_M \mathrm{d}M_r} M = -3 \left\langle \frac{v_s^2}{\Gamma_1} \right\rangle M, \tag{2.30}$$

where $\langle \cdot \rangle$ averages over the whole star.

Using the approximation $\langle v_s^2/\Gamma_1 \rangle \approx \langle v_s^1 \rangle / \langle \Gamma_1 \rangle$, one gets the following expression for the speed of sound:

$$v_s \approx \left(\frac{-\Omega \langle \Gamma_1 \rangle}{3M} \right)^{1/2}. \tag{2.31}$$

Equation (2.4.2) now gives

$$\mathcal{P} \sim \frac{2R}{v_s} \sim 2 \left(\frac{3}{\langle \Gamma_1 \rangle} \right)^{1/2} \frac{\left(MR^2\right)^{1/2}}{(-\Omega)^{1/2}} \sim \left(\frac{I_{\mathrm{osc}}}{-\Omega} \right)^{1/2}, \tag{2.32}$$

where the oscillatory moment of inertia I_{osc} is defined by

$$I_{\mathrm{osc}} \equiv \int_r r^2 \, \mathrm{d}M_r. \tag{2.33}$$

Catelan and Smith (2015) point out that Equ. (2.32) is analogous to many other expressions describing oscillating mechanical systems. For instance,

$$\mathcal{P} = 2\pi \sqrt{\frac{m}{k}} \tag{2.34}$$

describes the oscillation period of an object of mass m attached to a spring of constant k (Hooke's Law),

$$\mathcal{P} = 2\pi\sqrt{\frac{I}{K}} \tag{2.35}$$

describes the oscillation period of an object with moment of inertia I suspended by a string of *torsion coefficient* K.

The calculations presented so far could be improved by taking into account the fact that the sound speed changes along the path of the sound wave across the stellar interior. In this case, Equ. (2.4.2) changes, as an estimate of the sound wave travel time back and forth across the diameter of the star changes to

$$\mathcal{P} = 2\int_0^R \mathrm{d}t(r) = 2\int_0^R \frac{\mathrm{d}r}{v_s(r)} = 2\int_0^R \frac{\mathrm{d}r}{\sqrt{\frac{\Gamma_1(r)P(r)}{\rho(r)}}}. \tag{2.36}$$

For a proper integration of Equ. (2.36), the full form of the functions $P(r)$, $\rho(r)$ and $\Gamma_1(r)$ is needed. They can be calculated using the equations of stellar structure.
In the homogeneous case with the assumption of constant ρ and Γ_1, a *period – mean density relation* can be derived, as first obtained by Ritter (1879). His relation

$$\mathcal{P}\sqrt{\langle\rho\rangle} = \sqrt{\frac{3\pi}{2\Gamma_1 G}} \Rightarrow \mathcal{P} \propto \langle\rho\rangle^{-1/2} \tag{2.37}$$

indicates that denser pulsating stars should have shorter pulsation periods than less dense stars. Indeed, white dwarfs and δ Scuti stars show shorter periods than Cepheids and Mira.

Pulsation and Energy Conservation

In order to properly explain pulsating stars as thermodynamic engines, one must realize that the ϵ_g term in Equ. (2.20) is necessary. As found by Eddington (1926), without the ϵ_g term, the energy released by nuclear reactions or lost by neutrino emission would always be identical to the change in the outward luminosity. This means that no individual star layer would be able to cyclically absorb and release energy during a pulsation cycle.

Since ϵ_g represents the rate $\mathrm{d}Q/\mathrm{d}t$ at which energy is absorbed or released per unit mass in a given layer, following Catelan and Smith (2015) one can write Equ. (2.20) as

$$\epsilon_g = \frac{\mathrm{d}Q}{\mathrm{d}t} = \epsilon - \epsilon_\nu - \frac{\partial L}{\partial M_r}, \tag{2.38}$$

where the partial derivative symbol was used for Q, as Q is not a state variable.

According to the *First Law of Thermodynamics*, the rate of heat input or loss Q per unit mass into the layer can be expressed in terms of the rate change of the internal energy E per unit mass of a given layer, minus the work performed by the layer upon its surroundings (Catelan and Smith 2015):

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\partial E}{\partial t} + P\frac{\partial}{\partial t}\left(\frac{1}{\rho}\right). \tag{2.39}$$

This can be rewritten as follows, by expressing E as $E(\rho, P)$:

$$\frac{\partial E}{\partial t} = \left(\frac{\partial E}{\partial P}\right)_\rho \frac{\partial P}{\partial t} + \left(\frac{\partial E}{\partial \rho}\right)_P \frac{\partial \rho}{\partial t} \tag{2.40}$$

$$\frac{\partial Q}{\partial t} = \left(\frac{\partial E}{\partial P}\right)_\rho \frac{\partial P}{\partial t} + \left(\frac{\partial E}{\partial \rho}\right)_P \frac{\partial \rho}{\partial t}\frac{\partial \rho}{\partial t}. \tag{2.41}$$

It is convenient to write this in terms of luminosity L and effective energy generation rate $\epsilon_{\mathrm{eff}} \equiv \epsilon - \epsilon_\nu$. The detailed steps for doing so are e.g. described in Catelan and Smith (2015). One finally obtains

$$\frac{\partial \ln T}{\partial t} = (\Gamma_3 - 1)\frac{\partial \ln \rho}{\partial t} + (c_v T)^{-1}\left(\epsilon_{\mathrm{eff}} - \frac{\partial L}{\partial m}\right) \tag{2.42}$$

with the *third adiabatic coefficient* Γ_3, $\Gamma_3 \equiv 1 + \left(\frac{\partial \ln T}{\partial \ln \rho}\right)_s$.

Stability Conditions

The equations shown so far depict the adiabatic theory, which is successful at describing the pulsation period of stars, but fails in at least two aspects: first, it cannot explain the phase lags that are often observed between different physical quantities of a pulsating star; second, it is unable to predict at all which stars will pulsate.

Starting with the conservation of momentum equation, it is described now, based on Catelan and Smith (2015), how a non-adiabatic theory can describe which stars pulsate.

From Equ. (2.19), the conversion of momentum equation becomes

$$\frac{\partial^2 r}{\partial t^2} = -4\pi r^2 \frac{\partial P}{\partial m} - \frac{GM_r}{r^2}. \tag{2.43}$$

Multiplying both sides of Equ. (2.43) by $\frac{\partial r}{\partial t}$ results in

$$\frac{\partial r}{\partial t}\frac{\partial^2 r}{\partial t^2} = -4\pi r^2 \frac{\partial r}{\partial t}\frac{\partial P}{\partial M_r} - \frac{GM_r}{r^2}\frac{\partial r}{\partial t}. \tag{2.44}$$

The left-hand side of this equation can be also written as

$$\frac{\partial r}{\partial t}\frac{\partial^2 r}{\partial t^2} = v\frac{\partial v}{\partial t} = \frac{1}{2}\frac{\partial}{\partial t}v^2, \tag{2.45}$$

so that

$$\frac{1}{2}\frac{\partial}{\partial t}v^2 = \left(-4\pi r^2\frac{\partial P}{\partial M_r} - \frac{GM_r}{r^2}\right)\frac{\partial r}{\partial t}. \tag{2.46}$$

Integration over the volume of the whole star gives

$$\int\limits_M \frac{1}{2}\frac{\partial}{\partial t}v^2\,\mathrm{d}M_r = \int\limits_M \left(-4\pi r^2\frac{\partial P}{\partial M_r} - \frac{GM_r}{r^2}\right)\frac{\partial r}{\partial t}\,\mathrm{d}M_r \tag{2.47}$$

$$= -\frac{\mathrm{d}}{\mathrm{d}t}\left(\int\limits_M -\frac{GM_r}{r}\,\mathrm{d}M_r\right) - \int\limits_M \left(4\pi r^2\frac{\partial P}{\partial M_r}\right)\frac{\partial r}{\partial t}\,\mathrm{d}M_r. \tag{2.48}$$

The integral in the first term on the right-hand side is the total gravitational potential energy of the star. When performing the integration in the last term by parts, this gives

$$\int\limits_M \frac{1}{2}\frac{\partial}{\partial t}v^2\,\mathrm{d}M_r = -\frac{\mathrm{d}\Omega}{\mathrm{d}t} - \left[4\pi r^2 P\frac{\partial r}{\partial t}\right]_0^M + \int\limits_M P\frac{\partial}{\partial m}\left(4\pi r^2\frac{\partial r}{\partial t}\right)\mathrm{d}M_r. \tag{2.49}$$

As the pressure at the surface of the star is many orders of magnitude lower than in the interior,

$$\left[4\pi r^2 P\frac{\partial r}{\partial t}\right]_0^M \simeq 0. \tag{2.50}$$

Therefore, with good approximation,

$$\int\limits_M \frac{1}{2}\frac{\partial}{\partial t}v^2\,\mathrm{d}M_r \simeq -\frac{\mathrm{d}\Omega}{\mathrm{d}t} + \int\limits_M P\frac{\partial}{\partial M_r}\left(4\pi r^2\frac{\partial r}{\partial t}\right)\mathrm{d}M_r. \tag{2.51}$$

From the conservation of mass, it follows that

$$\frac{\partial}{\partial t}\left(\frac{1}{\rho}\right) = \frac{\partial}{\partial M_r}\left(4\pi r^2\frac{\partial r}{\partial t}\right), \tag{2.52}$$

and therefore

$$\int\limits_M \frac{1}{2}\frac{\partial}{\partial t}v^2\,\mathrm{d}M_r \simeq -\frac{\mathrm{d}\Omega}{\mathrm{d}t} + \int\limits_M P\frac{\partial}{\partial t}\left(\frac{1}{\rho}\right)\mathrm{d}M_r. \tag{2.53}$$

If one integrates this equation over a complete pulsation cycle, the total mechanical work that is transferred into kinetic energy of motion of the stellar layers is obtained:

$$W = -\int\limits_{\mathcal{P}} \frac{\mathrm{d}\Omega}{\mathrm{d}t} + \int\limits_{\mathcal{P}} \mathrm{d}t \int\limits_M P\frac{\partial}{\partial t}\frac{1}{\rho}\,\mathrm{d}M_r. \tag{2.54}$$

Accordingly,

$$W = -\left[\Omega\right]_0^{\mathcal{P}} + \int_{\mathcal{P}} \mathrm{d}t \int_M P \frac{\partial}{\partial t}\left(\frac{1}{\rho}\right) \mathrm{d}M_r = \int_{\mathcal{P}} \mathrm{d}t \int_M P \frac{\partial}{\partial t}\left(\frac{1}{\rho}\right) \mathrm{d}M_r, \tag{2.55}$$

as the gravitational potential is purely conservative. Averaging over a full pulsation cycle gives

$$\left\langle \frac{\mathrm{d}W}{\mathrm{d}t} \right\rangle \equiv \frac{W}{\mathcal{P}} = \frac{1}{\mathcal{P}} \int_{\mathcal{P}} \mathrm{d}t \int_{\mathcal{M}} P \frac{\partial}{\partial t}\left(\frac{1}{\rho}\right) \mathrm{d}M_r. \tag{2.56}$$

Following (Catelan and Smith 2015), this finally implies a *condition for pulsation*:

If $\langle \mathrm{d}W/\mathrm{d}t \rangle > 0$, the star maintains pulsation. When defining Ψ as the total pulsation energy of the star (see e.g. Moya and Rogríduez-López 2010), a natural timescale for the growth (or damping) of pulsations can be defined as follows (Rosseland 1949):

$$\tau \equiv -\frac{1}{2} \frac{\mathrm{d}W/\mathrm{d}t}{\Psi}. \tag{2.57}$$

Associated with this time scale, one also defines the so-called *stability coefficient* κ, $\kappa \equiv \tau^{-1}$.
$\kappa > 0$ implies overall damping and thus stability, whereas for $\kappa < 0$, the instabilities grow over time, leading the star to pulsate.
Any regions in the star that contribute positively to Equ. (2.56) are called *driving layers*, whereas those that contribute negatively are called *damping layers*.

Modifying Equ. (2.56) gives answers to the question which layers behave driving or damping. From the First Law of Thermodynamics, one obtains at each of the star:

$$\int_{\mathcal{P}} P \frac{\partial}{\partial t}\left(\frac{1}{\rho}\right) \mathrm{d}t = \int_{\mathcal{P}} \frac{\mathrm{d}Q}{\mathrm{d}t} \mathrm{d}t + \int_{\mathcal{P}} \frac{\partial E}{\partial t} \mathrm{d}t = \int_{\mathcal{P}} \frac{\mathrm{d}Q}{\mathrm{d}t} \mathrm{d}t + \left[E\right]_0^{\mathcal{P}} = \int_{\mathcal{P}} \frac{\mathrm{d}Q}{\mathrm{d}t} \mathrm{d}t, \tag{2.58}$$

where the integral of the internal energy over a closed cycle cancels out. Therefore,

$$\left\langle \frac{\mathrm{d}W}{\mathrm{d}t} \right\rangle = \frac{1}{\mathcal{P}} \int_{\mathcal{P}} \mathrm{d}t \int_M \frac{\mathrm{d}Q}{\mathrm{d}t} \mathrm{d}t. \tag{2.59}$$

Using Equ. (2.38) and (2.42), one finds that

$$\frac{\partial \ln P}{\partial t} = \Gamma_1 \frac{\partial \ln \rho}{\partial t} + \frac{\rho}{P}\left(\Gamma_3 - 1\right) \frac{\mathrm{d}Q}{\mathrm{d}t}. \tag{2.60}$$

At maximum compression, when $\partial \ln \rho / \partial t = 0$, one gets

$$\left(\frac{\partial \ln P}{\partial t}\right)_{\text{max.compr.}} = \frac{\rho}{P}\left(\Gamma_3 - 1\right) \frac{\mathrm{d}Q}{\mathrm{d}t} \tag{2.61}$$

with $(\Gamma_3 - 1)\rho/P) > 0$.

Therefore, if a layer of the star gains heat ($\mathrm{d}Q/\mathrm{d}t > 0$) during maximum compression, it will also be building up pressure ($\partial \ln P/\partial t > 0$) at this point of time. This implies that the pressure will be higher during expansion than during contraction. Exactly this is required to efficiently maintain pulsation. In such a *driving region*, therefore the following relation is satisfied:

$$\int_{V_{\min}}^{V_{\max}} PdV > -\int_{V_{\max}}^{V_{\min}} PdV. \tag{2.62}$$

What exactly are now the driving mechanisms for pulsation?

2.4.3 Driving Mechanisms

As pointed out by Eddington (1926), to maintain pulsations, stars must operate pretty much as thermodynamic engines with heat being added to matter at a high temperature only to be withdrawn at a low temperature. The question is now, how does this exactly occur in stars?

Accounting to (Catelan and Smith 2015), the equation for the stability coefficient κ in the m-th pulsation mode can be written as

$$\kappa_m = -\frac{\int_M (\delta T/t)_{m,\mathrm{ad}}\, \delta\epsilon_{\mathrm{eff}}\, \mathrm{d}m}{2\omega_m^2 J_m} + \frac{\int_M (\delta T/t)_{m,\mathrm{rad}} \left(\frac{\partial \delta L}{\partial M}\right)_m \mathrm{d}m}{2\omega_m^2 J_m} \tag{2.63}$$

with J_m is the corresponding oscillatory moment of inertia.
The first term in Equ. (2.63) is associated with energy generation, and the second with energy transfer. In the first case, when pulsations are excited, one refers to the ϵ *mechanism*, whereas in the latter case, the κ *and* γ *mechanisms* are at play.

The ϵ Mechanism

Within the region of the star where the thermonuclear reactions take place, the temperature increases during compression. This leads to an increase in the rate of energy generation, and vice versa during the expansion. Thus, energy is gained by these layers during the compression, and released during the expansion. As stated by Catelan and Smith (2015), this mechanism works exactly as required to establish pulsational instability according to Equ. (2.56) and (2.63).

Thermonuclear reactions show a strong dependence on temperature. If the amplitude of the temperature fluctuations in the energy-generation regions is sufficiently high in the course of pulsations, such a supply of energy will indeed fluctuate over time, thus being naturally able to maintain the pulsations. This is the so-called ϵ *mechanism* of stellar pulsation, where the ϵ is the nuclear energy generation rate. In classical pulsators, such as RR Lyrae and Cepheids, the ϵ

mechanism does not play an important role, whereas in other types of pulsating stars, it has been claimed to be of considerable interest (Catelan and Smith 2015).

The κ and γ Mechanism

For most stars, the energy transfer, rather than the energy generation, is the main cause for pulsation. Pulsation will be excited when the stability coefficient κ in Equ. (2.63) becomes negative. The second term in Equ. (2.63) then requires that during maximum compression (i.e., $\delta T/T > 0$), $\partial \delta L/\delta m$ being negative, implying an increase in δL with increasing M_r (i.e., towards the surface of the star). Thus, at least some layers of the star must gain energy during compression, and release energy during expansion to maintain pulsation. Such layers are called *driving layers*. They are typically associated with H and He partial ionization zones.

Assume now, that the Rosseland mean opacity Equ. (2.12) in a given layer of the star can be approximated for simplicity by

$$\kappa_R \propto \rho^n T^{-s}. \tag{2.64}$$

In the case of free-free absorption in a non-degenerate, fully ionized gas, the so-called Kramers opacity law can be applied, by setting $n = 1$, $s = 7/2$:

$$\kappa_R \propto \rho T^{-7/2} \tag{2.65}$$

that was derived by Eddington (1926) based on Kramer's opacity law. According to this expression, there is a tendency that during compression, opacity decreases in the layers of a star, caused by the rise in temperature.

However, there are a few "bumps" in the opacity, caused by the ionization of H and partial ionizations of He. In these regions, there is a tendency for the opacity to actually increase with increasing temperature, so the s value in Equ. (2.64) becomes negative. The consequence of an increasing opacity during compression is that the corresponding region of the star will "concentrate" energy during compression, and more easily release it during the expansion, leading to pulsation.
This increase in the opacity is known as κ *mechanism* (Baker and Kippenhahn 1962). The effect was first studied by S. A. Zhevakin and J. P. Cox (Cox and Whitney 1958; Cox 1960; Zhevakin 1963).
The increased ability of the same layers participating in the κ mechanism to gain heat during compression is called γ *mechanism* (Cox 1963).
The classical κ and γ mechanisms explain the excitation of pulsation instabilities in stars within the instability strip, such as RR Lyrae, Cepheids and δ Scuti stars.

Non-Radial Pulsations

The mechanisms presented so far describe radial pulsations. However, not all pulsating stars pulsate radially.

The discovery of the Sun's 5-minute-oscillations (Leighton 1960; Leighton et al. 1962) hinted that stars might pulsate in non-radial modes. Furthermore, it has led to the assumption that similar pulsations might be detected in other stars when observational techniques improve.
Nowadays, the study of oscillation in stars – asteroseismology – requires the detection of a huge range of pulsation modes, many of which are non-radial. Nowadays, asteroseismological studies of stars have grown enormously in the course of the past several years. The results of the CoRoT (Auvergne et al. 2008) and Kepler (Borucki et al. 2010) missions, among others, have enabled us to gain insight into the physical processes of star interiors.

The κ and γ mechanisms, as described before, are successful in describing the pulsation of stars located within the instability strip, but fail for many different types of other pulsating stars, most of them non-radial pulsators. In hot pulsating stars, the metal opacity bump (due to an increase in the heavy element contribution), rather than the opacity bump discussed before being associated with the H and He partial ionization zones, is responsible for driving the oscillations (*metal bump mechanism*, Simon (1982); Cox et al. (1992).

2.5 The Physics of Variable Sources

After giving an overview of the theory that describes the pulsation of stars, as well as the mechanisms causing them to pulsate, this section focuses on the sources the analysis done in this thesis is based on: RR Lyrae and Cepheids, as well as – despite being neither stars nor showing periodic behavior – QSOs. All three types of variable sources can be detected and classified by using the same methods shown in Chapter 5 and all three are of great interest for various purposes.
If not stated otherwise, in the following, periods are always given in units of days.

2.5.1 The Physics of RR Lyrae

With periods between 0.2 and 1.0 days, RR Lyrae stars are one of the most useful types of variable stars used for exploring the distances and properties of old stellar populations. In the HR diagram, they are found in the instability strip with absolute visual magnitudes near 0.6 and mean effective temperatures ranging between about 6000 and 7250 K (Catelan 2004). RR Lyrae stars are only found in systems that contain a stellar component older than about 10 Gyr, and they are thus an important standard candle for determining distances to very old systems.

The prototype of this class of variable stars, RR Lyrae itself, was discovered by Williamina Fleming on Harvard College Observatory photographs (Pickering et al. 1901). The class of variable stars was then defined through observations of RR Lyrae stars in globular clusters. Between 1895 and 1898, Bailey and Pickering found more than 500 variable stars in a search of 23 globular clusters (Bailey and Pickering 1913). Bailey noticed that many of these variables showed similar

properties: their periods were mostly shorter than a day, and their amplitudes on the blue-sensitive photometric plates were typically about 1 magnitude. These stars were first called *cluster variables*, of which most are stars we call nowadays *RR Lyrae stars*, or short, *RR Lyrae*.

RR Lyrae Types and Light Curve Properties

Bailey (1902) divided the RR Lyrae stars in ω Cen (ω Centauri, NGC 5139) into three sub-classes, now called *Bailey types*. They can be distinguished by light curve shape, period and amplitude.

Bailey type *a* stars show the largest amplitudes and the steepest rise to the maximum amplitude. RR Lyrae of type *b* are similar to those of type *a*, but with smaller amplitudes and longer periods. Type *c* RR Lyrae have shorter periods and lower amplitudes. Their light curves are more symmetric than those of types *a* and *b*, and show an almost sinusoidal shape.

As the type *a* and *b* RR Lyrae stars form a continuous sequence in an amplitude-period diagram, it is now usual to combine them into a single type RRac, leaving only the original Bailey type *c* distinct (RRc). Additionally there is a type called RRd stars, which are double-mode pulsators, unlike RRac or RRc (Nemec 1985). Among all types of RR Lyrae, RRab variables are the most common, making up ∼91% of all observed RR Lyrae (Smith 2004). RRc variables account for ∼9% of the observed RR Lyrae, RRd are the rarest RR Lyrae and make up only ∼1% (Smith 2004).

Fig. 2.13 shows light curves of typical RRab and RRc stars. The light curves are given in the *ugriz* filters used by the SDSS survey (York et al. 2000).
RR Lyrae show an increase in their amplitude as one goes from the near-infrared *z* filter to the *g* filter, but with only a small change as one continues to the *u* filter. Towards the ultraviolet, they reach amplitudes up to 4 magnitudes (Downes et al. 2004). When proceeding to the infrared instead, the decline of the amplitude with increasing wavelength continues.

Whereas some RR Lyrae stars have light curves that repeat nearly perfect from one pulsation cycle to the next, some RR Lyrae stars have light curves that change in a secondary period that can be tens or hundreds of days long. These changes are called the *Blazhko effect* (Blažko 1907).

RR Lyrae Stars as Standard Candles

RR Lyrae stars are especially important as they can be used to measure distances to systems containing old stellar populations, such as the Milky Way's halo. They were first used by Harlow Shapley to determine distances to globular clusters, leading to the awareness that the Sun is located far from the center of our Galaxy (Shapley 1914).
Within a single globular cluster, all RR Lyrae have about the same visual apparent magnitude, as they are in the HB stage of the evolution of low-mass stars. This makes RR Lyrae very good *standard candles*.

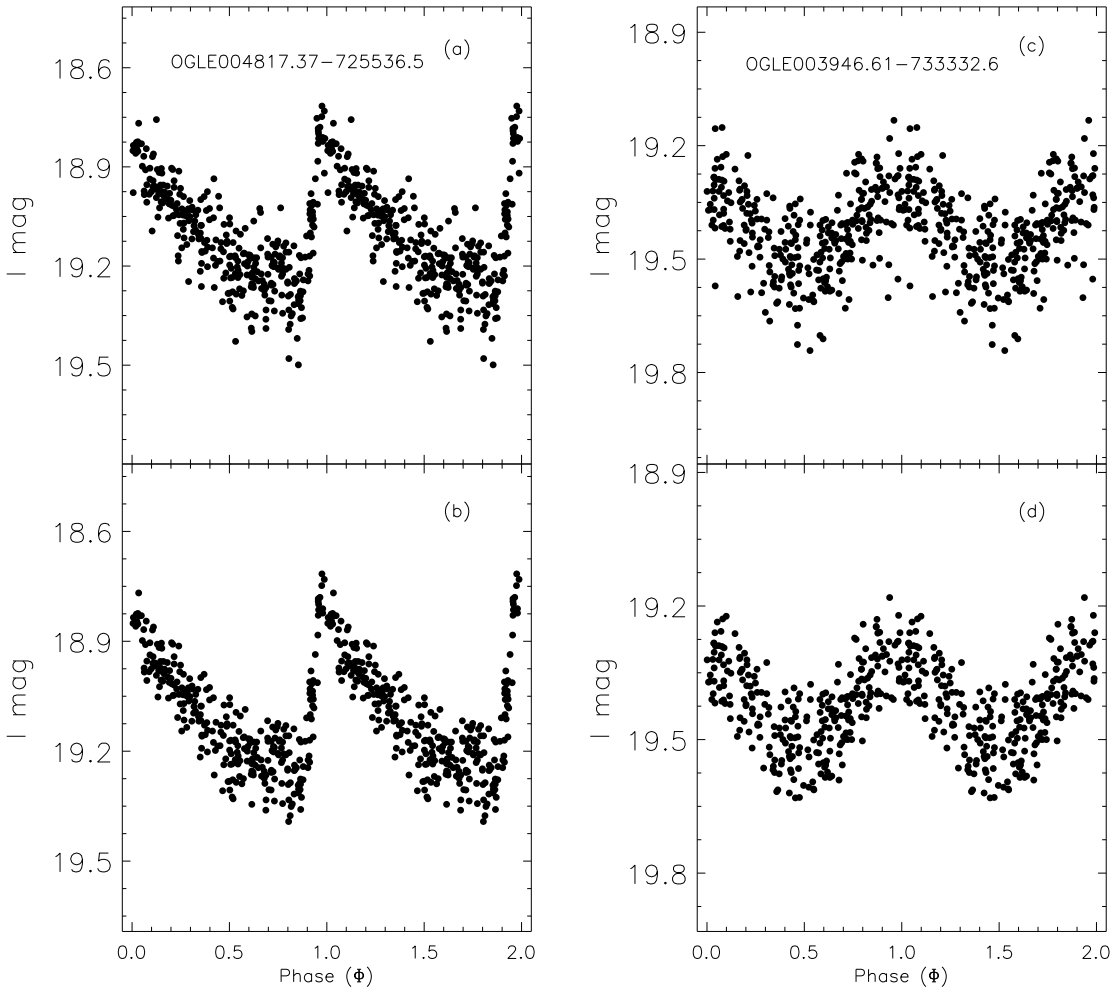


Figure 2.13 Examples of RRab and RRc light curves from OGLE-II. (a) shows a raw RRab light curve, (c) a raw RRc light curve. In panels (b) and (d), the corresponding outlier removed smoothed light curves are shown. The steep rise in the RRab light curve, in contrast to the almost symmetric nature of the RRc light curve, is clearly visible. Taken from Deb and Singh (2010).

For distance determination, the apparent magnitude must be transformed to absolute magnitudes. Until the 1960s, it was assumed that all RR Lyrae have the same absolute V band magnitude. Later on, a dependency on metallicity was found (Sandage 1990a), that is generally expressed as $\langle M_V \rangle = a + b[\mathrm{Fe/H}]$. A calibration by Benedict et al. (2011) results into the relation

$$\langle M_V \rangle = (0.214 \pm 0.047)([\mathrm{Fe/H}] + 1.5) + (4.5 \pm 0.05). \tag{2.66}$$

Longmore et al. (1986) found a linear relationship between the mean infrared K-band magnitude ($\lambda \simeq 2.20$ μm) and the logarithm of the RR Lyrae star's fundamental-mode period. This infrared *period-luminosity relation* has the advantage of being relatively insensitive to interstellar

extinction. Also, it is relatively insensitive to the star's [Fe/H] value. Updated relationships can be found in Cáceres and Catalan (2008), indicating

$$M_{\rm z} = 0.839 - 1.295 \log \mathcal{P} + 0.211 \log Z \tag{2.67}$$

$$M_{\rm i} = 0.908 - 1.035 \log \mathcal{P} + 0.220 \log Z \tag{2.68}$$

for SDSS i and z bandpasses and metallicity Z, $\mathcal{P}$ in days.

The Evolution of RR Lyrae Stars

RR Lyrae are stars who have already left the main sequence (see Fig. 2.9), ascended the RGB, undergone the He flash, and settled down to core He burning that characterizes stars on the HB. It takes more than 10 Gyr until they reach the HB. The lifetime of RR Lyre stars is expected to be in the order of 10^8 years (Koopmann et al. 1994). The variability of RR Lyrae stars is caused by pulsation, being mainly driven by κ and γ mechanisms. The zone within the star where He goes from being singly to doubly ionized is most important for driving the pulsations. RRab stars are pulsating in the fundamental radial mode, whereas RRc stars are pulsating in the first-overtone mode.

Period Changes

RR Lyrae can undergo period changes. This came apparent as the time spanned by observations of the same RR Lyrae reached 100 years and more. Whereas some have stable periods, others undergo significant changes. Such period changes have been observed for RR Lyrae in a number of the Milky Way's globular cluster (Catelan and Smith 2015). It was suggested by Sweigart and Renzini (1979) that discrete mixing events in the semi-convective zone of RR Lyrae could lead to period noise and thus period change. Cox (1998) proposed that small changes in the gradient of the He composition in the regions of RR Lyre stars below the H and He convective zones might produce the period changes.

Period Distributions: The Oosterhoff Groups

Oosterhoff, working on RR Lyrae within 5 globular clusters (Oosterhoff 1939), noted that they could be divided into two groups according their period, now known as the *Oosterhoff groups*. The globular clusters with a mean period of their RRab $\langle \mathcal{P}_{\rm ab} \rangle$ near 0.55 days became known as *Oosterhoff type I* clusters, whereas those with $\langle \mathcal{P}_{\rm ab} \rangle$ near 0.65 days became known as *Oosterhoff type II* clusters. Analysis of [Fe/H] showed that globular clusters of Oosterhoff type I are more metal rich than those of Oosterhoff type II. There are various approaches to explain this difference. RR Lyrae stars in Oosterhoff type II clusters are more luminous than those in Oosterhoff type I clusters. The longer periods of Oosterhoff type II RR Lyrae would then be a result of their lower

densities, according to the pulsation equation $\mathcal{P}\sqrt{\langle\rho\rangle} = Q$. As a possible explanation, Sandage (1981) suggested that a higher He abundance in the Oosterhoff type II clusters might account for their different period. However, this does not explain why Oosterhoff type II clusters have higher fractions of RRc stars than Oosterhoff type I clusters. Also it does not explain the higher period change rates found in Oosterhoff type II clusters.

A different explanation that accounts for both the higher RRc fraction as well as mean period, was proposed by van Albada and Baker (1973). They suggested the existence of a so-called hysteresis zone near the center of the instability strip. Within this zone, both the fundamental and the first-overtone modes can in principle be excited. RR Lyrae stars entering this zone of the HR diagram would keep the pulsation mode that they had.

2.5.2 The Physics of Cepheids

Classical Cepheid variable stars are supergiants with periods in the range of 1–5 days. The light curve amplitude is typically between 0.5 and 2 magnitudes in the V band, and the velocity amplitude due to the pulsation is in the range of 30–60 km s^{-1}.

Different to RR Lyrae, which can be found at all Galactic latitudes, Cepheids are strongly associated with the Galactic plane. More than 800 Cepheids are known in the Milky Way, and a few 1000 have been found in the two nearest galaxies, the Magellanic Clouds.

Cepheids show a close relationship between period and luminosity, which was found by Henrietta S. Leavitt in 1912 (Leavitt and Pickering 1912). This relation has given Cepheids a unique role in establishing the distances of near galaxies and hence the distance scale of the Universe, the "distance ladder".

Cepheid Types and Light Curve Properties

Among Cepheids, two types can be distinguished: Classical Cepheids (or type I Cepheids) are comparatively young stars of ages $\sim 10^8$ yr with masses of $2 - 3$ $M_\odot$. They show a strong concentration towards the Galactic plane and have low space velocities. Their ages can be estimated from star clusters. Within period-luminosity diagrams, they occupy a narrow strip.
Type II Cepheids are fainter than type I Cepheids of comparable period. From globular clusters, as well as from being present in the Galactic halo, their age can be estimated as being up to 15×10^9 yr. This implies that they must be less massive than type I Cepheids. Type II Cepheids can also be distinguished from type I Cepheids by the shape of their light curves. Most type I Cepheids, of which δ Cephei is a prototype, have asymmetric light curves, showing a steep rise to their maximum and a slower decline. Type II Cepheids, in contrast, show almost sinusoidal light curves.

Fig. 2.14 shows the light curves of the type I Cepheid SU Cygni in different photometric bands. As also for the RR Lyrae stars, the amplitude decreases as one goes from the ultraviolet to the

infrared. Some Cepheids of short periods have nearly sinusiodal light curves with amplitudes of only 0.5 mag.

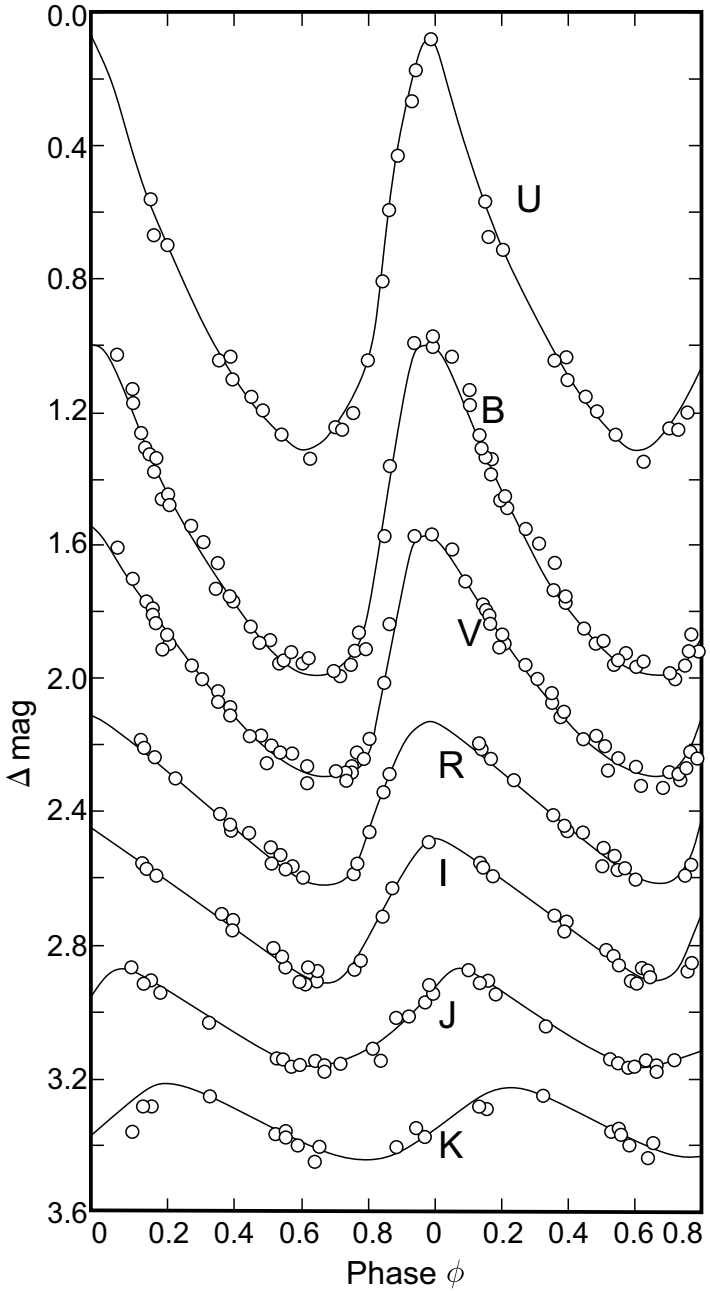


Figure 2.14 Amplitude and phase variation of a typical Galactic Cepheid as a function of increasing wavelength. Note for increasing wavelength the monotonic drop in amplitude, the progression toward more symmetric light variation, and the phase shift of maximum toward later phases. The wavelength increases from top (ultraviolet, blue, and visual) to bottom (red and near-infrared out to K=2.2 μm). Taken from Madore and Freedman (1991).

Cepheids as Standard Candles

The relation between period and luminosity of a Cepheid comes directly from the Stefan-Boltzmann law (Catelan and Smith 2015). When expressed in bolometric magnitude units,

$$M_{bol} = -5\log(R) - 10\log(T_{\rm eff}) + \text{const.} \tag{2.69}$$

Combining this with the pulsation equation Equ. (2.37), one gets

$$\log \mathcal{P} + 0.5\log(M) + 0.3M_{\rm bol} + 3\log T_{\rm eff} + \text{const} = \log Q, \tag{2.70}$$

where M is the stellar mass. From this, at a constant effective temperature, the period should increase with increasing luminosity.

The period-luminosity relation was first found empirically by Leavitt and Pickering (1912), and then calibrated by Shapley (1918).

Fig. 2.15 shows near-infrared period-luminosity relations for type I and type II Cepheids in the Large Magellanic Cloud. For type I and type II, both the offset and the slope differ.

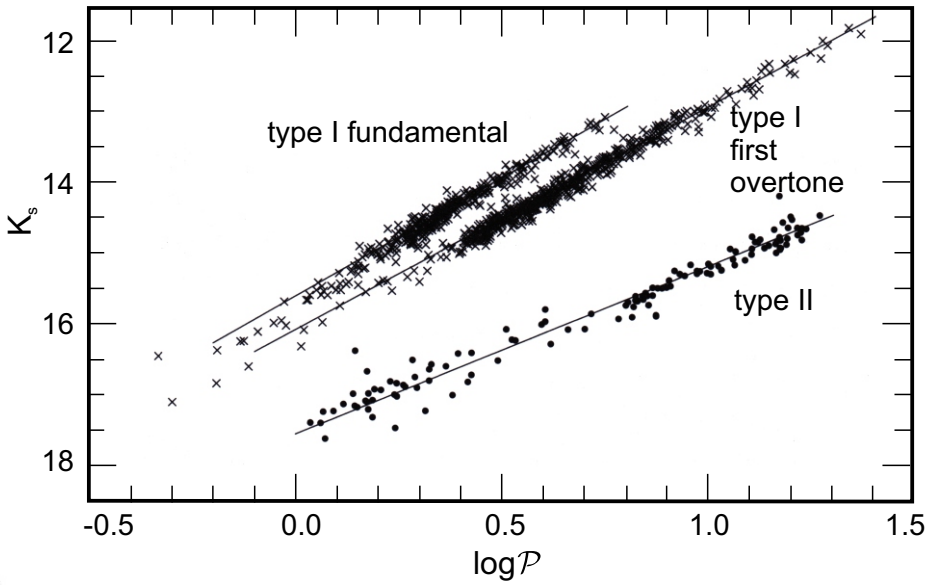


Figure 2.15 Near-infrared period-luminosity relations for Cepheids of type I and II in the Large Magellanic Cloud. Among type I, both fundamental and first overtone Cepheids are indicted. Adapted from Matsunga et al. (2009).

The Evolution of Cepheids

Cepheids of type I are stars who are more massive than the Sun, having evolved from main sequence stars of 2–20 $M_\odot$. Such a star starts pulsating as a Cepheid when it crosses the instability strip on its way to the RGB and, later, during a blueward loop as it fuses He in its core.

Like for other types of variables within the instability strip, κ and γ mechanisms within the H and He partial ionized zones are the most important drivers for pulsation. The variability of Cepheids of type I is caused by pulsations, being mainly driven by κ and γ mechanisms. The zone where He goes from singly to doubly ionized is mostly important to the driving of the pulsation.
As the lifetimes of such massive stars like Cepheids of type I are short, they are relatively young stars with ages in the range of 10^7 years for the brightest and most massive ones, to a few 10^8 for the faintest and less massive ones. For this reason, Cepheids of type I are found in systems that have experienced recent star formation. Thus, within the Milky Way, Cepheids belong to the young disk population.

In contrast, Cepheids of type II are old, evolved stars with low masses of about 0.5–0.6 $M_\odot$. They have evolved away from the main sequence, up to the giant branch, down the horizontal branch, back up the AGB, but are experiencing He flashes as He burning briefly switches on. This shifts the star to higher temperature and over the instability strip (Catelan and Smith 2015).

2.5.3 The Physics of QSOs

This type of variable sources differs in many ways from the ones discussed previously. First, QSOs aren't stars but are associated with the centers of active galaxies. Second, they show no periodic behavior but stochastic. Third, their astronomical application is not distance estimation, but establishing an astrometric reference frame.

QSOs are, like their higher-level type AGN, composed of supermassive black holes (SMBH) in the order of 10^5 to 10^9 $M_\odot$ and surrounding accretions disks. The gas in the disk heats up during accretion, resulting into the production of emission in the optical and ultraviolet range. Some QSOs also show radio or X-ray emission. They luminosities can be as large as 10^{47} ergs s^{-1} in tiny volumes ($\approx 2 \times 10^{14}$ cm, Edelson et al. 1996).

QSOs were discovered during the first radio surveys in the late 1950s. Due to their star-like appearance as point sources, but showing properties inconsistent with stars, they were named "quasi-stellar objects". The exact cause for their enormous total luminosities – of up to 10^4 times the luminosity of a typical galaxy – within a small volume (as implied by Spitzer and Saslaw 1966) were unclear until the physics of accretion disks were understood and imaging and spectroscopic observations were available. Such observations can give evidence for the existence of massive compact objects at the centers of galaxies (e.g. Kormendy et al. 1996a,b; Magorrian et al. 1998; van der Marel et al. 1997). Observed line broadening (e.g. Peterson 1997) indicates the presence of gas moving in a relativistic potential wall.

Light Curve Properties

Unlike variable stars, whose variability is often periodic or a least dominated by periodic components, AGN (and thus QSOs) show mostly no periodic variability. In consequence, QSO light

curves are described as a stochastic process, e.g. a Gaussian process (Rybicki and Press 1992), whose parameters can be determined by using a structure function (Rybicki and Press 1992).

QSOs vary in every waveband. Continuum variability in the optical was established even before the emission-line redshifts were understood.

Variability of QSOs occur on many different time scales. They range from weeks for changes on the thermal time scales in the accretion disk, over months for superpositions of stochastic processes up to several years for changes in the large-scale structure of accretion disks or lens crossing times.

Most, but not all, AGN continuum spectra have a spectacularly different appearance from normal galaxy spectra. Whereas in the UV, large fluctuations are common and occur on time scales from weeks to years, in the optical, the fluctuations are rather small.
A particularly well observed example, NGC 4151, is shown in the top panel of Fig. 2.16 in its UV, as well as one of NGC 5548 in the optical.

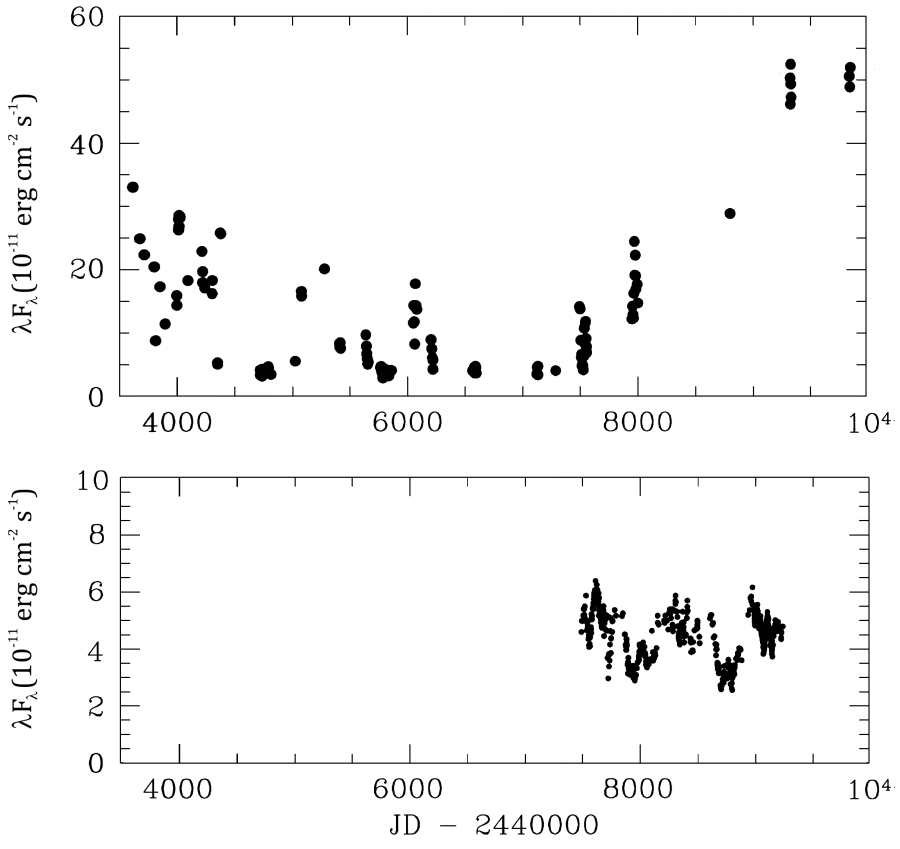


Figure 2.16 (Top) A long-term UV (1455 Å) light curve for NGC 4151; (bottom) a shorter optical (5100 Å) light curve for NGC 5548. In the UV, fluctuations of several are common and can occur on timescales ranging from weeks to years. In the optical band, the fluctuations tend to be rather smaller. Taken from Krolik (1999).

The Importance of QSOs

Originally, active galactic nuclei were considered to be a rare phenomenon. However, studies on the Palomar Survey (L. C. Ho and A. V. Filippenko and W. L. Sargent 1995) point a different picture. Out of the Palomar Survey sample, 86% of the galaxies show emission lines, among them, ∼40% of the galaxies show H II emission (an indicator for star formation), and ∼50% belong to the active galaxies, in detail, ∼30% are low-ionization nuclear emission-line regions (LINER), ∼13% are Seyfert I and II, and ∼10% of all galaxies have a broad Hα component.

It is assumed that almost all galaxies undergo active phases during their evolution, so AGN evolution is assumed to be closely related to galaxy formation and evolution in the Universe. Their large luminosities make them to be traced even at high redshifts and thus large distances and early stages of the Universe. As AGN are very good tracers of distributions of both visible and dark matter (Ferrarese et al. 2001), they enable a view on the large-scale structures in the early Universe. The evolution of supermassive black holes (SMBH) residing in the centers of AGN can be probe the intergalactic medium (IGM). "Feedback" from AGN affects the host galaxies and IGM (Silk and Rees 1998). Feedback in a galaxy is any process that heats or disrupts gas, and hence decelerates star formation, as hot turbulent gas will collapse into stars much more slowly than cold and stationary gas.

Despite from probing the early Universe, QSOs have another important application. The most stable celestial reference frames used so far build on the positions of extragalactic sources such as QSOs. The current IAU standard frame defining the coordinates on the sky is the ICRF-2, built using radio interferometry (Fey et al. 2015). Its accuracy is 100 μas.

Such reference frames are used for astrometry, but also for geodesy and navigation. With the data from the Gaia mission, for the first time an additional μas reference frame, but based on observations in the optical wavelengths, will be available.

The Physics and Evolution of QSOs

Gravitational accretion onto compact objects provides very efficient conversion of potential and kinetic energy into radiation. Such processes give a reasonable explanation for the observed high luminosities and rapid variability of such sources.

When the accretion disk heats up, the ultraviolet and X-ray continuum emission is able to photo-ionize and excite the diffuse cold atomic gas clouds close to the black hole. This leads to the production of emission lines, which are then broadened due to the high velocities of the clouds,reaching up to 10,000 kms^{-1} (Peterson 1997).

Despite such accretion processes, there is no evidence up to now what actually gives rise to QSO variability. Large-scale changes in the amount of in-falling material (as discussed e.g. in Hopkins and Beacom 2006) as well as disk instabilities (e.g. Schmidt et al. 2012) are considered as the most probable causes for most of the observed variability.

Other effects being discussed to contribute to the variability are microlensing by the host galaxy (Hawkins 1996; Zackrisson and Bergvall 2003) and starbursts in the host galaxy (Aretxaga et al. 1997). Central SMBHs are widely believed to be found in the centers of most or all galaxies. Furthermore, it is believed that almost all, or even all galaxies undergo active phases for about 10^7 to several billion years (Hopkins et al. 2005).

Whereas there are rare cases known in which AGN vary by several percent over a few nights (Pollock et al. 1979), most AGN (and thus QSOs) show variation of just a few percent over weeks to years. The fact that some show variations on very sort timescales is an indicator for the existence of a very small region causing the variability, ranging from a few light-months to a few light-days in diameter.

2.6 Surveys for Variable Objects

Due to the importance of various classes of variable sources – from QSOs to variable stars to exoplanets – many of surveys focused and focuses on their observation.

In this section, a brief review on surveys enabling science in the time domain is given. Some of them observe only small fields, whereas others are observing large portions of the sky ("all-sky surveys"). Surveys represent a fundamental data basis for astronomy. They are used to map the Universe in a systematic way, and thus discover new types of objects or phenomena. Some science can be done with the survey data alone, some requires the combination of data from different surveys, some requires a targeted follow-up of potentially interesting sources. Surveys can be used to generate large, statistical samples of objects that can be studied as populations, or as tracers of larger structures to which they belong. They can be also used to discover samples of rare objects, and may lead to discoveries of some previously unknown types.

The exponential growth of data volume and complexity makes a broader application of data mining and knowledge discovery technologies critical in order to take full advantage of this wealth of information.

Surveys can be classified in regard to their scientific motivation and strategy, their wavelength regime, ground-based vs. space-based, the temporal character (single- vs. multi-epoch, time domain), the type of observation (e.g. imaging, spectroscopy), their area coverage on the sky and their depth.

Progress in astronomy has always been driven by technology. From the viewpoint of surveys, the milestone technologies include the development of astrophotography (late 1800's), Schmidt telescope (1930's), radio electronics (1940's), space telescopes (1960's) and digital detectors, notably the CCDs (1980's).

In this section, a review on past, current and planned time-domain surveys is presented, with emphasis on photometric imaging surveys in the optical and infrared. The general and variety

properties of photographic, digital, ground- and space-based surveys are presented, illustrated by more details for a few outstanding examples. Finally, some prospects for the future are discussed.

2.6.1 Photographic Surveys

Repeatedly observing the sky was the starting point of astronomy in ancient times. Positions and brightnesses of the same objects – mostly stars and planets – were reported for mythological as well as calendrical reasons. The same striking sources were monitored repeatedly, leading to the first sky charts of something nowadays called targeted astronomical observations.

In contrast, modern sky surveys are not targeted, but aim to map and characterize the astronomical content of large fractions of the sky in a systematic manner.

Historically, surveying of the sky started with the naked eye, continued with telescopes, and one could consider Charles Messier's catalog[2] from the middle of the 18th century – first listed 1771 – as a pioneer in searching for and describing astronomical sources.

The way of surveying the sky was revolutionized by the invention of photography at the end of the 19th century. The first surveys in the modern sense, not only containing positions and descriptions but images, took place by systematically documenting large areas of the sky on photographic plates. One of the most notable of those is the *Harvard Plate Collection*, which spans over a century of sky coverage. It is currently digitized by the *Digital Access to Sky Century on Harvard* (DASCH, Laycock et al. 2010) project.
One important discovery made from analyzing repeated observations on photographic plates was the period-luminosity relation for Cepheids, found by Henrietta Leavitt (Leavitt and Pickering 1912) from stars in the Magellanic Clouds. This discovery laid the groundmark for the cosmological distance scale and the breakthrough discovery of the expanding Universe by Edwin Hubble and others in the 1920's.

Despite being not a time-domain survey, it's worth to mention here the *Henry Draper Catalogue* (HD) from the early 20th century. The first version was published between 1918 and 1924 for 225,300 stars, and successively extended until 1949. In total containing ∼360,000 stars, it gives spectral types based on objective prism plates. It is still in use. Currently, the Catalogue and Extension are available from the VizieR service of the Centre de Données astronomiques at Strasbourg as catalogue number III/135A.

In the 1930's, Fritz Zwicky pioneered the field of sky surveys in a way that affected much of the subsequent work. He built the first telescope on Mt. Palomar, an 18-inch Schmidt telescope, and used it to search for supernovae.

A major milestone was the first *Palomar Observatory Sky Survey* (POSS-I), conducted from 1949 to 1958, which mapped 3/4 of the entire sky with two observations per source, one using blue-

[2]The Messier Catalog. SEDS Messier Database. http://messier.seds.org/

and one using red-sensitive plates, down to 21 mag. It was continued from 1980 to 1999 as POSS-II. It is an important resource for star movements due to the large temporal baseline of about 4 decades. POSS-I is currently available as digitized scans of the photographic plates from *The Minnesota Automated Plate Scanner Catalog of the Palomar Observatory Sky Survey*, (MAPS Catalog).

2.6.2 Digital/ CCD Surveys

Photographic surveys, as well as their derived catalogs, were published as books or small sets of volumes that can be looked up. But as the data volumes rapidly increased in the 1990's, such catalogs soon contained millions of objects, so there was a transition to purely electronic publications, nowadays being available as web-accessible archives.

Aside from the digitized versions of the photographic sky surveys, a major milestone was the advent of the CCD surveys, like the Sloan Digital Sky Survey (SDSS). Due to such surveys, astronomy transitioned from a relatively data-poor science, dealing with a few sources and a few epochs of observations, to an immensely data-rich one. Thanks to the advent of several large-scale surveys observing large fractions of the sky multiple times, our understanding of the time-variable Universe has increased rapidly.

In the following, the most important digital sky surveys are ordered chronologically.

As one of the first digital sky surveys, the *All Sky Automated Survey* (ASA, Pojmanski 1997) covers the entire sky using a set of small telescopes at Las Campanos Observatory, Chile, and Haleakala, Maui. It was observing in V and I bands, with limiting magnitudes $V\sim$14 mag and $I\sim$13 mag. The ASAS-3 Photometric V-Band Catalogue contains over 15 million light curves.

Another key contribution at that time was the *Two Micron All-Sky Survey* (2MASS, Skrutskie et al. 2006), carrying out all-sky observation in the infrared. 2MASS was observing between 1997 and 2001, in two different locations at the U.S. Fred Lawrence Whipple Observatory on Mount Hopkins, Arizona and at the Cerro Tololo Inter-American Observatory in Chile, each using a 1.3-meter telescope for the northern and southern hemisphere, respectively. The survey covered four infrared bands, J (1.235 μm), H (1.662 μm), K_s (2.159 μm).
The goals of this survey included: Detection of galaxies in the "Zone of Avoidance", a strip of sky obscured in visible light by the Milky Way; detection of brown dwarfs; an extensive survey of low mass stars; cataloging of all detected stars and galaxies.
Although 2MASS was primarily a single-epoch survey, approximately 30% of the sky was observed multiple times. It produced an astronomical catalog with over 300 million observed objects, including minor planets of the Solar System, brown dwarfs, low-mass stars, nebulae, star clusters and galaxies. In addition, 1 million objects were cataloged in the 2MASS Extended Source Catalog. The final data release for 2MASS occurred in 2003, and is served by the Infrared Science Archive.
The science products of 2MASS are: Point Source Catalog (PSC) consisting of over 500 million

stars and galaxies; Extended Source Catalog (XSC) consisting of 1.6 million resolved galaxies; Large Galaxy Atlas (LGA) consisting of ∼600 nearby galaxies and globular clusters; All-Sky Quicklook and Atlas images providing full coverage of the infrared sky.

The *Sloan Digital Sky Survey* (SDSS, York et al. 2000) is the first CCD photometric survey at high Galactic latitudes, mostly in the northern sky. The imaging survey uses 5 passbands (*ugrizy*) with limiting magnitudes of 22.0, 22.2, 22.2, 21.3 and 20.5 mag, respectively. Additionally, SDSS spectra were collected by a series of spectroscopic programs.
The initial survey SDSS-I (2000-2005) covered ∼8,000 deg^2. SDSS-II (2005-2008) is made up of multiple surveys, among them the *Sloan Legacy Survey* that extended the area coverage to ∼8,400 deg^2, and catalogued 230 million objects, the *Sloan Extension for Galactic Understanding and Exploration* (SEGUE) that obtained almost 2.5×10^5 spectra over ∼3,500 deg^2, and the *Sloan Supernova Survey* which spectroscopically confirmed 500 type Ia supernovae in the redshift range $z = 0.05 - 0.4$.
SDSS-III was running from 2008 to 2014, using the Sloan Foundation 2.5-meter Telescope at Apache Point Observatory in New Mexico. SDSS-III consists of four surveys, BOSS, APOGEE, SEGUE-2, MARVELS.
SDSS-IV, the current survey (2014-2020), is consisting of sub-surveys for extending precision cosmological measurements to a critical early phase of cosmic history (eBOSS), expanding its revolutionary infrared spectroscopic survey of the Galaxy in the northern and southern hemispheres (APOGEE-2), and for the first time using the Sloan spectrographs to make spatially resolved maps of individual galaxies (MaNGA). Two smaller surveys will be executed as subprograms of eBOSS: The Time Domain Spectroscopic Survey (TDSS) will be the first large-scale, systematic spectroscopic survey of variable sources; while the SPectroscopic IDentification of EROSITA Sources (SPIDERS) will provide a unique census of supermassive black-hole and large scale structure growth, targeting X-ray sources from ROSAT, XMM and eROSITA (Clerc et al. 2016).
More than perhaps any other survey, SDSS has transformed the culture of astronomy in regards to sky surveys: A major innovation of SDSS was the effective use of databases for data archiving, as well as web-based interfaces for data access, being not only available within the community, but public. Multiple well-documented public data releases were made using this approach with the recent one, DR13, containing observations through July 2015.

The *Nearby Supernova Factory* (SNfactory, Aldering et al. 2002) operated by the Lawrence Berkeley National Laboratory (LBNL) searches for type Ia supernovae in the redshift range $0.03 < z < 0.08$ in order to establish the low-redshift anchor of the SN Ia Hubble diagram. This survey was carried out from 2003 to 2008.

The *Catalina Real-Time Transient Survey* (CRTS, Drake et al. 2009) uses existing synoptic telescopes and imaging data resources from the *Catalina Sky survey* (CSS, Drake et al. 2009) for near-earth objects and potentially planetary hazard asteroids (NEO/PHA). The solar system objects remain in the domain of the CSS, while CRTS aims to detect astrophysical transient and

variable objects from the same data stream. It started operation in 2007.
CRTS utilizes three wide-field telescopes: the 0.68-m Schmidt telescope at Catalina Station, AZ (CSS), the 0.5-m Uppsala Schmidt (Siding Spring Survey, SSS) at Siding Spring Observatory, NSW, Australia, and the Mt. Lemmon Survey (MLS), a 15-m reflector located at Mt. Lemmon, AZ. They are operated for 23 nights per lunation, centered on new moon. Most of the observable sky is covered up to 4 times per lunation. The total area coverage is $\sim$30,000 deg^2, as it excludes the Galactic plane within $|b| < 10 - 15°$. In a given night, 4 images of the same field are taken, separated by $\sim$10 min. The combined data stream covers up to 2,500 deg^2 per night to a limiting magnitude of $V \sim 19 - 20$ mag, and add up to 275 deg^2 per night to $V \sim 21.5$ mag. Date cover time baselines from 10 min to years.

The *Optical Gravitational Lensing Experiment* (OGLE, Udalski 2003) is a long-term large-scale sky survey searching for various variable and transient sources, among others microlensing. Its main targets are the Magellanic Clouds and the Galactic Bulge. Those regions are the most natural locations to conduct such search, as they have a large number of background stars that are potential targets for microlensing during a stellar transit. As the optical depth for microlensing is very small (about 10^{-6}), a long-term large-scale survey is needed to detect them and to draw conclusions from a statistically significant sample of microlensing events.
OGLE began regular observation in 1992. Since then, it has undergone various phases. The first phase, OGLE-I (1992-1995), was the project pilot phase. For OGLE-II (1996-2000), a telescope dedicated for this project, using an 8-chip mosaic CCD camera, was constructed. OGLE-III (2001-2009) was primarily devoted to detecting gravitational microlensing events and transiting planets the Galactic Bulge, the constellation Carina, and both Magellanic Clouds. OGLE-IV was starting in 2010, using a 32-chip mosaic CCD camera with the main goal being to increase the number of planetary detections using microlensing.
Within the first three phases, OGLE detected as many as 20 new extrasolar planets, more than 4000 microlensing events, and several hundred thousand new variable stars. An *OGLE-III Online Catalog of Variable Stars* is available; its goal is to record all variable sources located in the OGLE-III fields in the Magellanic Clouds and Galactic bulge. The data currently available include: classical Cepheids in the Galactic Bulge, LMC and SMC; type II Cepheids in the Galactic Bulge, LMC and SMC; anomalous Cepheids in LMC and SMC; RR Lyrae stars in the Galactic Bulge, LMC and SMC; Long Period Variables in the Galactic Bulge, LMC and SMC; Double Period Variables in LMC; R CrB stars in LMC; δ Scuti stars in LMC.

SkyMapper (S. C. Keller et al. 2007) was developed at the Australian National University.
The fully automated 1.35 m wide-angle telescope is located at Siding Spring Observatory, Australia. Its camera covers a $\sim$5.7 deg^2 field of view. Started operation in 2007, SkyMapper scanned the entire southern sky 36 times in 6 filters (SDSS *grizy*), a Strömgen system-like u, and a narrow V band near 4000 Å. It will generate $\sim$100 MB of data per second during every observed night, totaling about 500 TB of data at the end of the survey.

Whereas SkyMapper is observing the southern sky, its northern counterpart is Pan-STARRS. The *Panoramic Survey Telescope & Rapid Response System* (Pan-STARRS, Chambers 2011) is a wide-field panoramic sky survey developed at the University of Hawaii's Institute for Astronomy. The survey is operated by an international consortium of institutions. It is envisioned as a set of four telescopes with a 3° field of view each, observing the same region of the sky simultaneously. The telescopes are located at Haleakala Observatory on the island of Maui, Hawaii.
Up to now, only one telescope (PS1) is operating, it started operation in 2010. PS1 can cover up to 6,000 deg^2 per night and generates up to several TB of data per night; however, not all images are saved, and the expected final output is estimated to be ∼1 PB per year.
The primary goal of PS1 is to survey ∼3/4 of the entire sky (the 3π *survey*) with 12 epochs in each of the 5 bands ($g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$, $z_{\rm P1}$, $y_{\rm P1}$). The coadded images should reach considerably deeper than SDSS. A dozen key projects, some requiring additional observations, are also underway. The data is restricted to the consortium members until completion of the PS1 3π survey.
The method on finding and classifying variable sources as part of this thesis work was applied to PS1 3π data. A more detailed introduction to Pan-STARRS, and especially PS1 3π, is given in Chapter 3.

The *Palomar Transient Factory* (PTF, Law et al. 2009) uses a 48-inch Schmidt telescope at the Palomar Observatory. The data are taken in a point-and-stare mode, with 2 exposures per field per night, mostly in the broad R and G bands. The survey is operational since 2009. PTF reaches a depth of $m_{\rm R} = 20.45$ and $m_{\rm G} = 21.0$ and covers a few hundred deg^2 per night. The overall coverage is ∼1/2 of the entire sky. The survey's cadence is mostly optimized for supernova discovery. PTF was continued as i-PTF and led to the development of ZTF (Bellm 2014).

There are many other time-domain surveys using ground-based CCD observations, e.g. the *Lincoln Near-Earth Asteroid Research* survey (LINEAR, Stokes et al. 1998), the *Supernova Legacy Survey* (SNLS, Pritchet 2005), MACHO (Alcock et al. 2001), *UKIRT Infrared Deep Sky Survey* (UKIDSS, Lawrence et al. 2007).

2.6.3 Space Telescope Surveys

Ground-based astronomy has several limitations: The main problems concern atmospheric seeing, scattering of light in the atmosphere, and the absorption of radiation with wavelengths shorter than 290 nm by oxygen and nitrogen.

Electromagnetic radiation of all wavelengths reaches Earth's upper atmosphere from the Universe. As different wavelengths can give information about different astrophysical processes, astronomers are interested in examining the complete spectrum. However, radiation of different wavelengths is absorbed by different amounts in the atmosphere.

O and N completely absorb all radiation with wavelengths shorter than 290 nm. Ozone (O_3) absorbs most of the ultraviolet. Electromagnetic radiation over a large range of infrared (IR) wavelengths is absorbed by water vapor and CO_2. This prohibits ground-based IR observation

with the exception of the near-infrared wavelengths from 1 to 10 mm and the far infrared up to 10 μm.

A wholesale exploration of the mid/far IR regime requires a space-based platform, and the *Infrared Astronomy Satellite* (IRAS, Beichman et al. 1986), launched in 1983, opened a huge new area of research.
Some of the subsequent missions included the *Wide-Field Infrared Survey Explorer* (WISE Wright et al. 2010), launched in 2009. His all-sky survey mapped the sky at four infrared wavelength bands, 3.4, 4.6, 12 and 22 μm. WISE detected more than 250 million objects, including near-earth asteroids (NEOs), brown dwarfs, QSOs, ultra-luminous starbursts and other sources of interest. When the telescope run out of hydrogen for cooling, the mission was continued as NEOWISE, and the survey continued for an additional four months using the two shortest wavelength detectors. NEOWISE carried out measurements of asteroids and comets from images collected by the Wide-field Infrared Survey Explorer (WISE) spacecraft. NEOWISE provides a rich archive for searching for solar system objects.
Data from the original WISE as well as NEOWISE missions have already enabled research in a variety of fields. With its increased sensitivity and time-domain information, combining them to ALLWISE extends this, as well as opens avenues of study that were not possible with the individual data.

The *Hubble Space Telescope* (HST, with its catalog, the *Hubble Source Catalog* as described in Budavári and Lubow 2012), named in honor of astronomer Edwin Hubble, started its operation in 1990. It is observing in the near ultraviolet, visible, and near infrared, using a primary mirror with a diameter of 2.4 m. From its low Earth orbit position, being outside the influence of Earth's atmosphere, it is able to take extremely high-resolution images with an angular resolution of 0.05 arcsec and a pointing accuracy of 0.007 arcsec. The HST has made more than 1.2 million observations since its mission began in 1990, resulting into $\sim$10 TB of new archive data per year and up to now more than 14,000 scientific papers. The HST is still operating, and could continue for decades.

The *Spitzer Space Telescope* (Spitzer, SST, Werner et al. 2004), named in honor of astronomer Lyman Spitzer, who had promoted the concept of space telescopes in the 1940s, is an infrared space observatory launched in 2003.
Spitzer is equipped with a 0.85 m primary mirror, that was cooled to 5.5 K, and follows a heliocentric instead of geocentric orbit. Its three instruments enable astronomical imaging and photometry from 3.6 to 160 μm, spectroscopy from 5.2 to 38 μm, and spectrophotometry from 5 to 100 μm.
The planned lifetime of the mission was 2.5 years with a possible extent of another 2.5 years until the He supply for cooling was exhausted. When this happened in May 2009, the two shortest-wavelength modules of the IRAC camera were still operable with the same sensitivity as before. Spitzer was then continued as the so-called *Spitzer Warm Mission*. All Spitzer data, from both the originally phase with the full waverange as well the limited warm phase, are archived at the

Infrared Science Archive (IRSA).
Among many other spectacular results like finding the youngest stars ever detected, it has directly captured light from exoplanets in 2005, namely from the "hot Jupiters" HD 209458b and TrES-1b (Deming and Seager 2005; Charbonneau et al. 2005).

HST's scientific successor, the *James Webb Space Telescope* (JWST Boccaletti et al. 2015), is scheduled for launch in 2018. Its nominal mission time is five years, with a goal of ten years. Different than the HST, it will observe from long-wavelength (orange-red) visible light, through near-infrared to the mid-infrared (0.6 to 27 μm). It is currently under construction and scheduled to launch in October 2018. The JWST has a larger primary mirror than the HST (6.5 meter, segmented, resulting in a collecting area about five times as large as HST's). The telescope will be located near the Earth–Sun L2 point, allowing it to use a single sunshield to keep the instruments below 50 K.
The design of JWST emphasizes the near to mid-infrared for three main reasons: high-redshift objects have their visible emissions shifted into the infrared, as more distant an object is, the younger it appears; cold objects such as debris disks and planets emit their radiation primarily in the infrared; infrared radiation is better able to pass freely through regions of cosmic dust that scatter radiation in the visible spectrum.
JWST's primary mission encompasses four scientific goals: to search for light from the first stars and galaxies that formed in the Universe after the Big Bang, to study the formation and evolution of the first galaxies, to understand the formation of stars and planetary systems and to study planetary systems including direct imaging of exoplanets. Many of them are beyond the reach of current ground and space-based instruments.

Kepler (Borucki et al. 2010), named after the astronomer Johannes Kepler, is a space telescope launched in 2009 in order to discover Earth-size planets orbiting other stars. Kepler is designed to survey a portion of our region of the Milky Way to discover Earth-size exoplanets in or near habitable zones and estimate how many of the billions of stars in the Milky Way have such planets. Its photometer continually monitors the brightness of over 145,000 main sequence stars in a fixed field of view. This data is transmitted to Earth, then analyzed to detect periodic dimming caused by exoplanets that cross in front of their host star.
The initial planned operational time was 3.5 years, but greater-than-expected noise in the data, from both the stars and the spacecraft, enforced additional time was needed to fulfill all mission goals. Initially, in 2012, the mission was expected to be extended until 2016, but on July 14, 2012, one of the spacecraft's four reaction wheels used for pointing the spacecraft stopped turning, and completing the mission would only be possible if all other reaction wheels remained reliable. Then, on May 11, 2013, a second reaction wheel failed. This meant the current mission needed to be modified to continue its search for exoplanets. Kepler was used further on in the so-called *K2* mission in order to detect habitable planets around smaller, dimmer red dwarfs (Howell et al. 2014).
As of September 2016, Kepler had found 2,330 confirmed exoplanets, along with a further 4,696

unconfirmed planet candidates. Further 129 planets have been confirmed through Kepler's K2 mission, and there are 458 K2 candidate exoplanets.

The *Gaia* mission (Prusti 2014), launched in December 2013 will provide fundamental data for a better understanding of the structure of our Galaxy. Gaia started its scientific mission in July 2014 and has been mapping the Milky Way ever since.
Gaia is an ambitious mission to chart a three-dimensional map of the Milky Way in order to reveal its composition, formation and evolution. Gaia will provide unprecedented positional and radial velocity measurements with accuracies required to produce a positional kinematic census of about one billion stars in the Milky Way and throughout the Local Group. This amounts to about 1 percent of the Galactic stellar population.
As Gaia scans the sky, it creates a precise three-dimensional map of astronomical objects – stars, asteroids, comets and other – throughout the Milky Way and map their motions. Gaia will monitor each object about 70 times over a period of five years. From the observations, astrometric parameters are determined: two corresponding to the angular position of a given object on the sky, two for the derivatives of the object's position over time, and the object's parallax from which distance can be calculated. Gaia will determine the position, parallax, and annual proper motion of 1 billion stars with an accuracy of about 20 μas at 15 mag, and 200 μas at 20 mag. This is an accuracy 100 times better than of Hipparcos. Additionally, positions will be determined a magnitude of $V = 10$ down to a precision of 7 μas between 12 ,and 25 μas down to $V = 15$ mag, and between 100 and 300 μas to $V = 20$ mag. The precision depends the color of the star.
Spectrophotometric measurements are carried out in order to provide the detailed physical properties such as luminosity, effective temperature, chemical composition and gravity for all observed stars.
Similar to its precursor Hipparcos, Gaia is equipped with two telescopes. They provide two observing fields with a fixed angle of 106°.5 between them. Gaia rotates continuously around an axis perpendicular to the two telescopes' lines of sight, and maintains a constant angle to the Sun. While doing so, the spin axis has a slight precession across the sky. Thus, a reference system is obtained by precisely measuring the relative positions of objects from both observing directions. The radial velocity of the brighter stars is measured by an integrated spectrometer, making use of the Doppler effect.
On September 12, 2014, Gaia discovered its first supernova in a galaxy about 500 million light-years away. On July 3, 2015, a map of the Milky Way's star density was released. On 13 September 2016, ESA has released a 3D map based on data from the the first 14 month of the mission, containing over a billion stars, out of them 400 million newly found sources.

2.6.4 The Future of Surveys for Variable Sources

Synoptic surveys are now the largest data producers in astronomy, entering the Petascale regime, opening the time domain for systematic exploration. Planned facilities for the next decade and

beyond, such as LSST, and the *Square Kilometer Array* (SKA, Maartens et al. 2015) will revolutionize our understanding of the Universe with nightly searches for changing sources over large fractions of the sky. A great variety of interesting phenomena, spanning essentially all sub-fields of astronomy, can only be studied in the time domain, and these new surveys are producing large statistical samples of the known types of sources and events for further studies, and have already uncovered previously unknown subtypes of these.

Such surveys are generating a new way of doing science, and prepare for even larger surveys to come, e.g. LSST. Astrophysical and instrumentation knowledge, methodology in both astronomy and data science, and experience that are being accumulated now are crucial to fully exploit such forthcoming surveys.

Time-domain astronomy is clearly a part of astronomy depending strongly on computational systems, and will increasingly depend on novel machine-learning and artificial intelligence (e.g. structure finding) tools. The growth of data quantity, coupled with an improved data homogeneity, its challenging but enables a new generation of statistical or population studies: with samples of millions of sources, one could look for subtle effects being simply not accessible with more limited data sets.

A number of important astrophysical phenomena can be discovered and studied only in time domain, ranging from exploration of the Solar System to cosmological phenomena. In addition to the studies of known time domain phenomena, e.g. variable stars, supernovae and quasars, there is an obvious possibility of discovering new types of objects and phenomena.

Numerous surveys, studies and experiments have been conducted in this area, are in progress, or are in the planning stages, indicating the growth interest in time-domain astronomy, leading to the *Large Synoptic Survey Telescope* (LSST).

The field has been fueled by the advent of the new generations of digital synoptic time domain surveys, which cover the sky many times, as well as the ability to respond rapidly to transient events using robotic telescopes. This new growth area of astrophysics has been enabled by information technology, continuing evolution from large panoramic digital sky surveys, to panoramic digital sky "cinematography", a term used in the context of LSST.

The data streams generated by panoramic digital synoptic sky surveys require rapid, real-time processing of massive data streams, with event detection for follow-up, filtering, characterization, and rapid publication of the data to the astronomical community.

The LSST, as described in the LSST Science Book, is a wide-field telescope that is currently under construction at Cerro Paranal in Chile. The primary mirror will be 8.4 m in diameter, the secondary 3.4 m. The large hole in the primary mirror reduces the collecting area to that of a 6.68 m telescope.

The LSST is planned to start observing in 2019, and will produce a 6-bandpass (0.3-11 μm, *ugrizy*) wide-field, deep astronomical survey over 20,000 deg^2 of the southern sky with up to

1000 visits per field. The camera will be able to cover ∼9.6 $\deg^2$ in individual exposures. LSST will take more than 800 panoramic images each night, with 2 exposures per field. This will lead to 30 TB of data per night. By doing so, the accessible sky will be covered twice a week. The data will be continuously generated and updated every observing night. Calibration and co-added images, and the resulting catalog will be generated on a slower cadence. The final source catalog after 10 years of observation is expected to have a data volume of 60 PB. Processing and analysis of this huge amount of data introduces a number of challenges in the fields of real-time data processing, distribution and archiving. A more detailed description of LSST is given in Chapter 3.

The *Zwicky Transient Facility* (ZTF, Bellm 2014) is an optical synoptic survey that builds on the experience and infrastructure of the PTF. Using a new 47 $\deg^2$ survey camera, ZTF will survey more than an order of magnitude faster than PTF to discover rare transients and variables. ZTF is planned to start observing in 2017. Its main goals are searches for fast transients, young supernovae, rare variables as well as counterparts to gravitational-wave detections.

Chapter 3

PS1 3π as Time-Domain Survey & LSST Pilot Survey

Many astronomers and astrophysicists are not only interested in a couple of sources strongly restricted by area and depth, but in explanations for phenomena that would involve data covering huge ranges in area, magnitude range and also temporal resolution. Such an "ideal" data set would help in mapping the Milky Way in its "complete" content, as well as to constrain simulations such as on galaxy evolution.

Early Milky Way surveys (i.e., before SDSS) have suffered from shortage of data; to cope with this, astronomers working with them had to heavily use analytic density laws (fitting functions for density profiles as well as luminosity functions, being often inspired by extragalactic observations, such as the luminosity function of galaxies, the color-luminosity relation, size-luminosity relation, quantitative morphology of galaxies) to characterize the results. Nowadays large, deep and uniform data sets, like SDSS and Pan-STARRS, have shifted the emphasis from model fitting towards multidimensional mapping. Upcoming surveys covering the Milky Way and beyond will do this even more.

Answering questions related to the important scientific problems of the next decade (such as studying the evolution of our Milky Way and of galaxies in general, discovering the nature of Dark Energy and Dark Matter, and opening up the time domain to discover rare transient and variable objects down to faint magnitudes with a fast cadence) rely strongly on deep, wide-field time-domain imaging of the sky in optical bands.
For the understanding of galaxy formation and evolution, it will be essential to examine the full multidimensional distributions of their properties. As data sets and modeling techniques as well as the performance of computing centers evolve, models will be tested not only by their capacities in reproducing the mean trends in galaxy properties but by their ability to reproduce and explore the full distribution.
The upcoming generation of synoptic sky surveys[3], like LSST, will be operating in the Peta-scale regime to fulfill these requirements. Surveys like Pan-STARRS are already preparing for these challenges. This new generation of surveys does not only depend critically on the state of information and computer technology, but will push it to a new performance regime.

[3]Here, this word is adopted from the LSST Science Book (LSST Science Collaborations and LSST Project 2009), who use it to refer to "looking at all aspects of something", derived from the Greek word "Synopsis".

This chapter deals with the time-domain properties of PS1 3π and why and how PS1 3π can serve as a pilot survey for the upcoming LSST. Here, the question comes up on what is meant by a "pilot survey" and why this is auxiliary for an upcoming survey.

In the case described here, a *pilot survey* is an existing survey that can act as a preliminary survey with respective to an upcoming one. The pilot survey is fully operationally and scientifically used, but can also be used to gather information for upcoming surveys, such as determining the efficiency as well as caveats of future surveys.

For the reasons of similar sky coverage (PS1 3π has 30,000 deg^2, whereas LSST has 20,000 deg^2), as well as for similar bandpasses (PS1 3π has g_{P1}, r_{P1}, i_{P1}, z_{P1}, y_{P1}, whereas LSST $ugrizy$), and also for the greater depth expected from LSST (PS1 3π has a single-exposure depth of $g_{\mathrm{P1}} < 22.0$ mag whereas LSST has $g < 25.0$ mag) as well as LSST's higher cadence (PS1 3π has $\sim$60 epochs, whereas LSST will have $\sim$2000 epochs), PS1 3π can be seen as a pilot survey for LSST. As PS1 3π is multi-band with non-simultaneous observations, has a sparse time sampling and is covering almost the same fraction of the sky as LSST, it can serve as a testing ground for various modeling approaches, for variable sources and beyond. Studies on PS1 3π can help with developing multi-band analysis methods, and also helps with doing science with preliminary LSST data.

In the following, first, a description of both surveys is given. Then, it is discussed why and how PS1 3π can serve as a pilot survey for LSST, and for other multi-band synoptic sky surveys in general.

3.1 The Pan-STARRS1 3π Survey

The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) consists of currently two telescopes, of which the first one, PS1, was used for the Pan-STARRS1 survey. Using this 1.8 m, f/4.4 Ritchey-Chrétien telescope with a 3.2° field of view (FOV) ($\sim$7 deg^2), the PS1 survey began full science observation on 13 May 2010, and the observations for the all-sky survey PS1 3π were completed in April 2014.

The Pan-STARRS project is a collaboration between the University of Hawaii Institute for Astronomy, MIT Lincoln Laboratory, Maui High Performance Computing Center and Science Applications International Cooperation.

Most of the PS1 observing time is dedicated to two surveys: The 3π survey, a survey of the entire sky north of declination -30°, and the medium-deep (MD) survey, a deeper, many-epoch survey of 10 fields, each 7 deg^2 in size (Chambers 2011).

In the following, science goals, as well as technical aspects of the PS1 survey, especially the PS1 3π survey, are described.

3.1.1 The Telescope

The PS1 telescope is designed as a wide-field optical imager devoted to survey operations (Chambers 2011). The telescope has a 1.8 m aperture primary mirror, f/4.4, having a FOV of $\sim$7 deg^2. It is using a 1.4 Gpixel camera (GP1) in its focal plane, with a resolution of 0.26 arcsec/pixel. Its location on the peak of Haleakala on Maui offers a point-spread function (PSF) with a full-width at half-maximum (FWHM) of about 1 arcsec (Hodapp et al. 2004; Kaiser et al. 2010; Tonry et al. 2012).

The camera is equipped with a 8×8 array of orthogonal transfer array (OTA) CCDs. Each OTA is further subdivided into an 8×8 array of "cells", each an independent 590×548 μm pixel CCD (Tonry et al. 2012).

PS1 is observing in the optical and near-infrared (near-IR), spanning 400-1000 nm. The filters are designated $g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$, $z_{\rm P1}$, $y_{\rm P1}$ in order to distinguish PS1 from other photographic systems (see also Sec. 3.1.4). These filters are similar to those used in the SDSS (see Stubbs et al. 2010), with the following differences: (i) PS1 has no u band whereas SDSS has, (ii) the $g_{\rm P1}$ filter extends 20 nm redwards of $g_{\rm SDSS}$, paying the price of 5577 Å sky emission for greater sensitivity and lower systematics for photometric redshifts, (iii) the $z_{\rm P1}$ filter is cut-off at 920 nm, giving it a different response than the detector-response defined $z_{\rm P1}$, (iv) PS1 has an additional near-IR y band whereas SDSS has not; the $y_{\rm P1}$ filter covers the region from 920 nm to 1030 nm with the red limit largely determined by the transparency of the silicon in the detector.
These filters and their absolute calibration are described in Stubbs et al. (2010) and Tonry et al. (2012).

3.1.2 Science Goals

The Pan-STARRS system was originally designed for the purpose of detecting potentially hazardous objects in the Solar System. But the wide-field, all-sky, multi-band and time-series nature of the observations makes it also excellent suited for many other astronomical purposes, ranging from Solar System astronomy to cosmology.

The PS1 Science Consortium defined 12 key projects, covering Solar system science right out to cosmology. In detail, these key projects are: populations of objects in the inner Solar System; populations of objects in the outer Solar System; low-mass stars, Brown Dwarfs, and young stellar objects; search for exoplanets by dedicated stellar transit surveys; structure of the Milky Way and the Local Group; a dedicated deep survey of M31; massive stars and supernova progenitors; cosmology investigations with variables and explosive transients; galaxy properties; AGN and high-redshift quasars; cosmological lensing; large-scale structure.

3.1.3 Observing Strategy

PS1 3π has observed all the sky accessible from Hawaii, resulting in 30,000 deg^2.
The PS1 3π observing strategy (Magnier et al. 2012) has a complex schedule in order to balance the needs for the different survey science projects. At the end of the survey, a total of 12 observations in each of the 5 filters was planned to be available for each part of the observable sky. This rather theoretical value would apply in the case of a perfect survey, having a perfect focal plane without any gaps or overlaps between neighboring observations. Indeed, gaps in the focal plane (between cells, between chips, and from masked pixels) lead to an average fill-factor of ∼80% for single exposures.
For the reason of neighboring exposures having both overlapping areas as well as gaps due to the layout of chip and camera, and for the reason of observational gaps due to weather issues, the total survey mission was extended to 5.7 years.

The temporal distribution of the 12 observations in each filter follows an elaborated schedule. For scheduling observations, the following guidelines were set out as shown by Magnier et al. (2012):

- TTI pairs:
  Any specific field is always observed twice per night in a single filter, where the observations take place within 20-30 minutes. Two observations being related in this way are called "Transient Time Interval" (TTI) pairs. They allow for the discovery of moving objects (asteroids and NEOS), but are also helpful for the detection of fast varying sources. TTI pairs are mutually subtracted as part of the nightly processing; sources detected in the difference image are then reported to the Moving Objects Pipeline Software (MOPS).

- scheduling of blue bands:
  The blue bands (g_{P1},r_{P1},i_{P1}) are observed close to opposition to enable asteroid discovery. These observations normally occur within ∼1.5 months of opposition for any given field. Thus any given field should be observed a total of 12 times in these 3 filters within a 2-3 month window each year.

- scheduling of red bands:
  For the reddest two bands (z_{P1},y_{P1}), the observations are scheduled as far from opposition as feasible in order to enhance the parallax factors and allow for discovery of faint, low-mass objects in the solar neighborhood. This constraint results in 2 observations in each of z_{P1}, and y_{P1} occurring roughly 4-6 month before and 4-6 month after opposition for any given field.

Each year, each field was planned to be observed twice in the same filter with an additional TTI pair of images, making for four images of each part of the sky per year in each of the five PS1 filters. The pointing of individual observations is then designed to both carefully cover the entire 3π region and trade-off between maximal overlaps and optimized image differencing. Whenever

possible, the TTI pairs were obtained in the same pointing to minimize the area loss in the difference image due to mismatched gaps.

The end result of the observing strategy is that the full 3π region is covered in a wide range of time periods in each filter and shows a large range of spatial overlaps. The overlaps are important for carrying out photometric and astrometric solutions.

3.1.4 Photometry

PS1 observations are done with exposure times between 30 and 60 s, leading to the limiting apparent magnitudes for single exposures of $g_{\rm P1} < 22.0$ mag, $r_{\rm P1} < 22.0$ mag, $i_{\rm P1} < 21.9$ mag, $z_{\rm P1} < 21.0$ mag, $y_{\rm P1} < 19.8$ mag (point sources, 5σ) and for stacked images of $g_{\rm P1} < 23.4$ mag, $r_{\rm P1} < 23.4$ mag, $i_{\rm P1} < 23.2$ mag, $z_{\rm P1} < 22.4$ mag, $y_{\rm P1} < 21.3$ mag (point sources, 5σ), as stated by Metcalfe et al. (2013) and Schlafly et al. (2014). The PS1 transmission curves are given in Fig. 3.5 at the end of the Section.

Each image requires about 2 GB of storage. The images are processed through the Image Processing Pipeline (IPP Magnier 2006, 2007; Magnier et al. 2008), performing automatic bias subtraction, flat fielding, astrometry, photometry and image stacking and differencing for every image taken. The nightly processing is carried out in a massively parallel fashion at the Maui High Performance Computer Center.

Magnitudes

Traditionally, astronomical magnitudes are defined as 2.5 times the logarithm of the ratio of fluxes given between the object of interest (and observed with the given telescope) to that of Vega (observed with the same instrumentation). There are a couple of drawbacks of such a system, including the strong dependence on the precision of Vega's magnitudes in the used bandpasses, and the problem of observing a bright star like Vega with modern instruments designed to observe very faint sources.

For this reason, PS1 uses the alternative *AB magnitude system*, in which the magnitude of a source is defined by the integral of the flux density spectrum multiplied by the overall throughput as a function of wavelength for the given telescope (Oke and Gunn 1983).
For a source with a flux density spectrum of f_ν erg/sec/cm^2/Hz and a telescope with a system response of $A(\nu)$, the AB magnitude for a bandpass is defined to be

$$m_{\rm AB} = -2.5 \log \frac{\int f_\nu (h\nu)^{-1} A(\nu) {\rm d}\nu}{\int 3631 \text{ Jy } (h\nu)^{-1} A(\nu) {\rm d}\nu} \tag{3.1}$$

Using AB magnitudes, the accuracy of the calibration is limited by knowledge of the system response including the atmosphere, and our knowledge of the spectral energy distribution of a specific star of interest.

Tonry et al. (2012) have determined the overall system zero points needed to place the PS1 magnitudes onto the AB system. For doing so, they have included information on the relative spectral response of the system, the filter transmission curves and also the transmission of the atmosphere at the site, using observations of selected spectro-photometric standards taken in a photometric night. Additionally, a large number of stars having measured spectra were used to provide additional constraints by making stellar locus diagrams. Comparison of the fluxes predicted by this method, and the magnitudes observed in each of the bandpasses led to the inclusion of tweaks to 12 system parameters in order to obtain the most precise magnitude calibration. These tweaks are all at the ∼1%. Details on the PS1 photometric system can be found in Tonry et al. (2012).

Photometric Calibration

Schlafly et al. (2012) have reported on the photometric calibration of the first 1.5 years of the PS1 survey. In their analysis, performing a highly-constrained relative photometric calculation called "ubercal", they select only the photometric nights and assign each a single fitted zero point and a single fitted value for the airmass extinction coefficient per filter. This requires an external zero point definition; Schlafly et al. (2012) used the zero points from Tonry et al. (2012) for the images. In the subsequent analysis, for each night, the zero point is determined by minimizing the dispersion of the measurements of the stars gained from multiple nights. Additionally, they determine flat-field corrections as part of the minimization process.
Schlafly et al. (2012) determined four distinct time periods ("seasons") having quite consistent flat-field corrections that are clearly different from the other seasons. The cause of this was identified with specific changes in the optical system, namely small scale changes in the vignetting and the PSF structure.

The resulting photometric *ubercal* system is shown by Schlafly et al. (2012) to have reliability across the survey region as high as (8.0, 7.0, 9.0, 10.7, 12.4) millimags in ($g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$, $z_{\rm P1}$, $y_{\rm P1}$), respectively.

3.1.5 Data Releases

The internal 3π stacked catalogs were released in three subsequent processing versions (PV), with each version corresponding to a higher number of individual exposures and improved photometry. The current and final internal data release is PV3 (which will become the public release), reaching a single-exposure depth of $g_{\rm P1} \sim 22$ mag (point sources, 5σ level) and covering a total baseline of 5.7 years, whereas most of the observations are within 5 years.

PS1 3π PV3 contains at total of 3.0×10^{10} detections for 6.0×10^{9} sources.

The work at hand is carried out with PV3 for final results, and using PV2 for pre-analysis and during the methodology was designed. Fig. 3.2 shows the total number of exposures in PS1 3π

for PV2 and PV3. Whereas PV2 has 2.85×10^5 exposures, PV3 has 3.75×10^5 exposures. PV3 observations cover a longer temporal baseline; whereas PV2 covers a baseline of 4.2 years (with most observations within 3.6 years), the baseline of PV3 is ∼550 days longer.
Fig. 3.2 gives the total number of epochs per source in PS1 3π for PV2 and PV3. Whereas PV2 has an average of 55 epochs per source, PV3 has 72 because of its longer baseline. Both figures refer to the total number of epochs taken, so no outlier cleaning was applied.

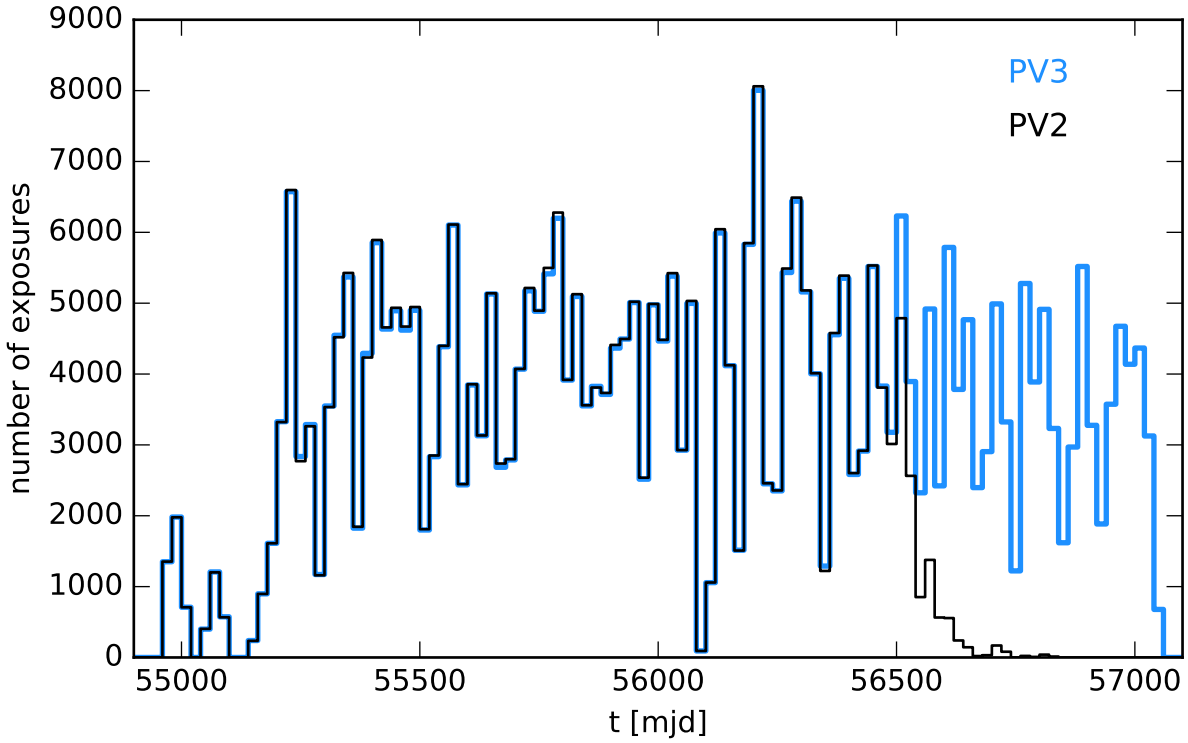
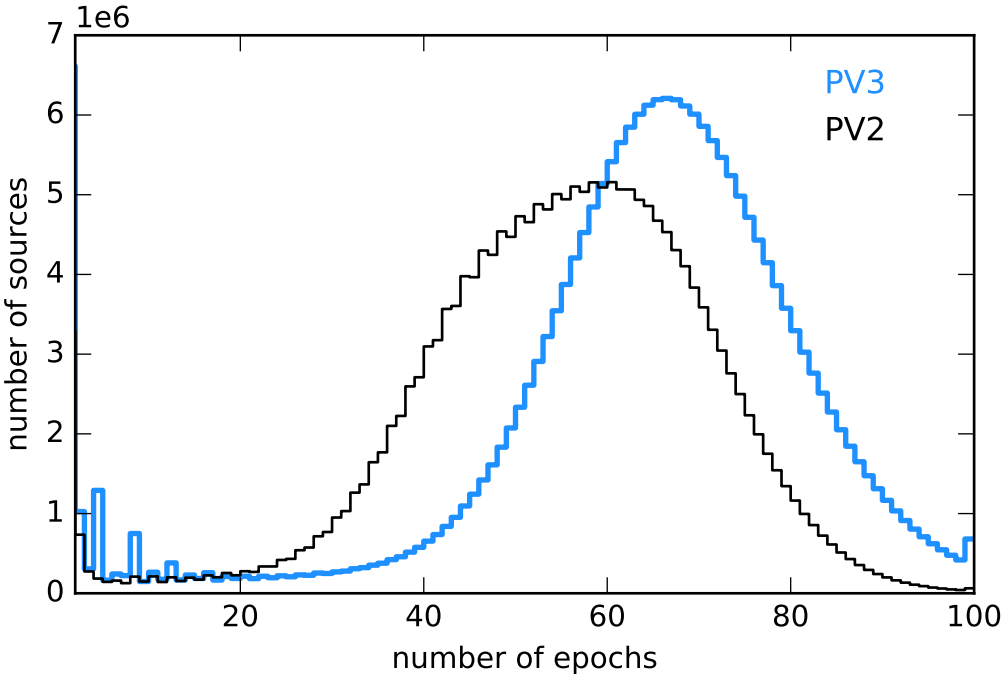
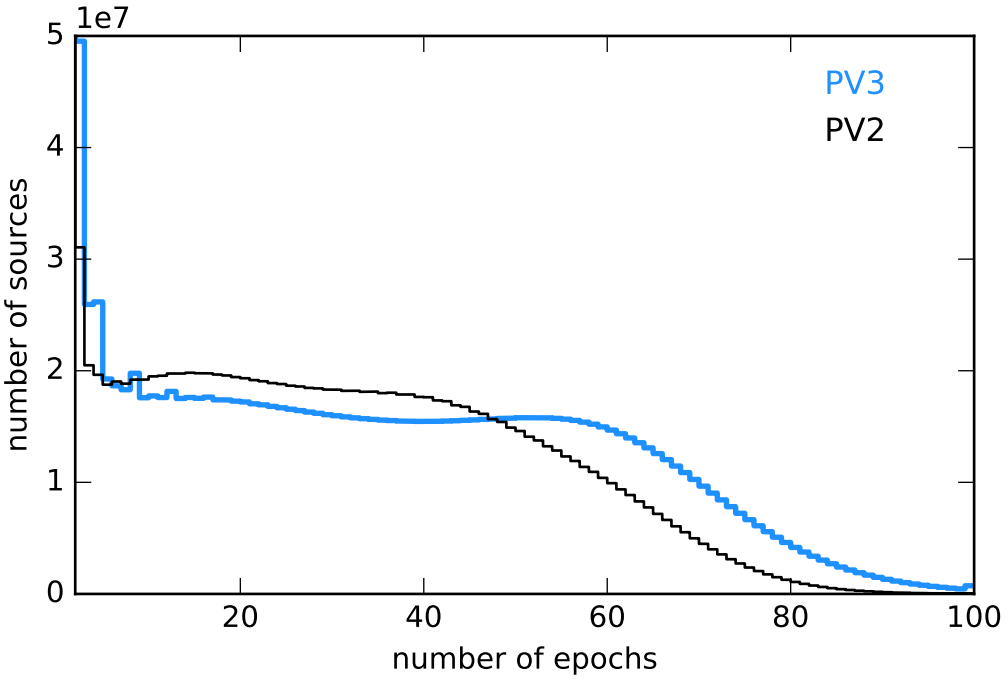


Figure 3.1 Total number of exposures in PS1 3π PV2 and PV3. Whereas PV2 has 2.85×10^5 exposures, PV3 has 3.75×10^5 exposures. PV3 extends PV2 especially at mjd>56500 resulting in a baseline being ∼550 days longer.



(a) bright limit: $15 < i_{\rm P1} < 18$



(b) faint limit: $18 < i_{\rm P1} < 21.5$

Figure 3.2 Total number of epochs per source in PS1 3π PV2 and PV3, over all five bands. Whereas PV2 has an average of 55 epochs per source in the bright limit ($15 < i_{P1} < 18$), PV3 has 72 because of the longer baseline shown in 3.2. In the faint limit ($18 < i_{P1} < 21.5$), as shown in the lower panel, the distribution of the number of epochs is flatter with a lower average number of epochs per source. The peaks at low number of epochs results from very faint sources and detection errors, resulting in sources having only a very small number of epochs.

3.2 LSST

The Large Synoptic Sky Telescope (LSST) is one out of the upcoming generation of synoptic sky surveys that will fulfill the requirements for scientific fields such as exploring the evolution of our Milky Way and of galaxies in general, discovering the nature of Dark Energy and Dark Matter, and makes the discovery of rare transient and variable sources possible down to faint magnitudes (LSST Science Collaborations and LSST Project 2009). All these topics rely strongly on deep, wide-field time-domain imaging of the sky in optical bands.

The LSST will be sited on Cerro Pachón in the Northern Chilean Andes. This location enables the observation of sky regions with up to $\delta < 35.5°$ at an airmass of 2.2 or less, resulting in a 0.6 mag loss of sensitivity at 500 nm compared to an observation in the zenith (LSST Science Collaborations and LSST Project 2009). The site of LSST corresponds to an observable area of 31,000 deg^2; however, the main survey has only a coverage of 20,000 deg^2, as it avoids the confusion-affected parts of the Galactic plane.

In the following, science goals, as well as technical aspects of the LSST are described, based on the LSST Science Book (LSST Science Collaborations and LSST Project 2009).

3.2.1 The Telescope

This science requirement, as described in LSST Science Collaborations and LSST Project (2009), leads to a single wide-field telescope and camera which repeatedly surveys the sky with deep short exposures, enabling a fast cadence. The three-mirror telescope is equipped with with a 8.4 m primary mirror, 3.4 m secondary mirror and 5.02 m tertiary mirror, resulting in an effective aperture of 6.7 meters and FOV of 9.6 deg^2. It is using a 3.2 Gpixel camera, with a resolution of 0.2 arcsec/pixel. LSST will carry out a main survey of 20,000 deg^2 of the sky in six broad photometric bands ranging from u to y, imaging each region of the sky roughly 2000 times over a ten-year survey life time.

LSST is currently under construction on El Peñón Peak of Cerro Pachòn in the Northern Chilean Andes. Cherro Pachòn is also the site where the 8.2 m diameter Gemini-South and 4.3 diameter Southern Astrophysical Research (SOAR) telescope are located. Previous observations with these telescopes have confirmed the excellent imaging quality that can be obtained from this site. LSST is expected to enter operations in 2022.

LSST's wavelengh coverage is 320–1080 nm, ranging from the optical to the near-IR. The filter set consists of *ugrizy*; out of them, five are concurrent in the camera at a time, providing an almost simultaneous survey.

Further information on the main system and survey characteristics can be found in the LSST Science Book (LSST Science Collaborations and LSST Project 2009).

3.2.2 Science Goals

As a sensitive, multicolor time-domain survey over most of the sky, LSST will dramatically impact nearly all fields of astronomy and many new areas of fundamental physics. The aim of LSST is to observe deep, wide and fast; a strategy that will enable a broad range of scientific investigations.

LSST is designed to achieve multiple goals in four main science themes (LSST Science Collaborations and LSST Project 2009): taking an inventory of the Solar system, mapping the Milky Way, exploring the transient optical sky, and probing Dark Energy and Dark Matter.
LSST will take an inventory of the Solar System and extend the boundaries of nowadays surveys both for asteroids and trans-Neptunian objects. It can be anticipated that LSST will detect and characterize over 80% of 140 m or larger killer asteroids, several million main-belt asteroids, and over 100,000 trans-Neptunian objects.
LSST will map the Milky Way out to 400 kpc. The survey will enable the detection of RR Lyrae in the halo out to 400 kpc, main-sequence stars to a distance of 100 kpc, and additionally will provide geometric parallaxes for all stars within 300 pc. Thus, it will take an inventory of the Milky Way, unveiling its formation and accretion history. Additionally, LSST will carry out a census on the stellar content of the Milky Way regarding kinematics and stellar composition (abundances). Previous surveys, such as 2MASS, SDSS and PS1, have shown in great detail that the Galactic halo is composed of stars accreted from companion galaxies. LSST will give a way more detailed look at indicators of how our Galaxy formed and evolved.
As being a time-domain survey with a fast cadence, LSST will explore the transient and variable optical sky with a variety of time scales in the range from 10 sec to the whole sky every few nights, totaling 1000 visits over 10 years of survey mission. By carrying out these observations, LSST will enable the scientific community to characterize a vast amount of objects being members of already known classes – such as RR Lyrae or QSO – as well as finding members of rare classes and discovering new classes. LSST's high cadence in combination with its wide-area coverage and great depth will enable the discovery and detailed analysis of objects being as rare as neutron star and black hole binaries, as well as the optical counterparts to gamma-ray bursts.
Among known classes, especially pulsating variables such as RR Lyrae and Cepheids are of interest. Progress in doing research on variable sources is nowadays limited as this requires observations that not only cover the time domain in detail, but also the parameter space of possible pulsation properties. LSST will be able to contribute by providing a substantial number of "complete" (i.e., having very high sampling) light curves for RR Lyrae in both Galactic and LMC globular clusters. RR Lyrae pairs, being members of eclipsing binary systems are of special interest, as this will enable an important test on stellar models: Models on stellar interiors can explain the pulsation of stars, however, many models differ only in a tiny fraction of mass (e.g. Szabó et al. 2004). To check them against observations at the required level of precision, the mass of the stars is needed at the same precision. The only reliable method for doing so is calculating the mass from a binary system. Additionally, the duration of the eclipses gives the radius of each member of the binary system.

The possibilities provided by LSST will even reach to cosmological scales, enabling to constrain Dark Matter as well as Dark Energy. To achieve this task, LSST will provide a variety of techniques for fundamentally test assumptions on cosmology and gravity theories. LSST will provide a sample of 3 billion galaxies having excellent photometry and shape measurements, over 100,000 clusters of galaxies, and a sample of several million type Ia SNe. By detecting a vast number of AGN, LSST will increase our understanding of such systems dramatically. LSST will enable the scientific community to gain a considerably more detailed insight, and also the possibility to compare observations to theoretical modeling of AGN feedback, which already indicated that AGN play a key role in galaxy evolution. Furthermore, measuring distances, growth of structure and curvature simultaneously for $0.5 < z < 3$, LSST data will tell about the nature of recent acceleration if it is due to Dark Energy or modified gravity. Supernovae will provide high angular resolutions in order to probe the homogeneity and isotropy of the Universe.

These areas highlighted here are just a few of the many on which LSST will have enormous impact.

To give numbers, it is expected that the data gained from LSST will enable (LSST Science Collaborations and LSST Project 2009):

1. The mapping of stellar number density with observations of ∼10 billion main sequence stars to (unextincted) distances of 100 kpc over 20,000 deg^2 of sky.
2. The mapping of stellar metallicity over the same volume, using observations of photometric metallicity indicators in ∼200 million near turn-off main sequence (F/G) stars.
3. The construction of maps of other more luminous tracers, such as RR Lyrae variables, to as far as 400 kpc; this is the approximate virial radius of the Milky Way.
4. The construction of high-fidelity maps of the tangential velocity field out to 10 kpc or more at 10 $\mathrm{km\,s^{-1}}$ precision, and as far as 25 kpc at 60 $\mathrm{km\,s^{-1}}$ precision.

In the following, a more detailed description on two fields is given, namely science enabled by mapping the Milky Way with LSST, and science that can be done with AGN.

As shown in the LSST Science Book (LSST Science Collaborations and LSST Project 2009), maps of stellar distribution gained by LSST will allow measurements of structural parameters such as densities and kinematics of all Galactic components (bulge, disk, halo) including such being only poorly observed yet (e.g., the disk scale length). Putting them together with kinematic information, they will allow for the construction of global dynamical models of the Milky Way, inferring the distribution of mass and the potential of the Milky Way. Furthermore, LSST will put observational constraints on the distribution of matter in the Galactic disk and halo, as well as of Dark Matter in the inner Galaxy.

LSST is able to achieve such a complete map of the Milky Way because its combination of the following capabilities: LSST has a u band, which allows for the measurement of stellar metallicities of near turn-off stars and for mapping them throughout the observed disk and halo volume.

LSST is observing in the near-IR using its y band, which allows for the mapping of stellar number densities and proper motions even in regions of high dust extinction.
Due to LSST's well-sampled time domain nature, it will be capable of identifying variable stars that play a key role as density and kinematic tracers out to large distances.
LSST will enable proper motion measurements for stars being 4 magnitudes fainter than will be obtained by Gaia (LSST Science Collaborations and LSST Project 2009).
In total, LSST data will increase the amount of data available for Milky Way science by two orders of magnitude (Ivezić et al. 2008).

Another large field that will be covered by LSST is science that can be done with AGN. There are three ways on how to identify AGN in LSST data: using their colors in LSST's six-band system, using their variability, and matching the sources to data at other wavelengths gained by different surveys.

For selection by color, LSST can benefit from the u band. At low redshifts ($z \lesssim 2.5$), quasars are blue in $u - g$ and $g - r$, and are well-separated from stars in color-space. At this redshift range, the u-band data are crucial for selection of AGN, to distinguish AGN from white dwarfs and A and B stars. High-redshift AGN will be easily distinguished; the y filter should allow quasars with redshift of 7.5 to be selected.

In addition to color, variability is another strategy for AGN selection. The amplitude of AGN variability depends upon rest-frame variability time scale, wavelength, luminosity, and possibly also redshift (Berk et al. 2004). The cadence of LSST will be especially useful for selecting low-luminosity AGN, which cannot be selected by color as they would be swamped by their hosts. Variability time scales in combination with color also allow for clean separation of AGN from variable stars. An approach for using variability and color information to select quasars in Pan-STARRS 1 3π data is also shown in this thesis (see Chapter 5) and in Hernitschek et al. (2016) (applied to a preliminary version of PS1 3π, PV2). It is expected that the efficiency of AGN selection by variability alone may be comparable to the color selection efficiency (Sesar et al. 2007; Hernitschek et al. 2016).

Fig. 3.3 gives the number of high-redshift ($z > 6$) quasars that are expected to be found by LSST as a function of redshift and limiting magnitude.

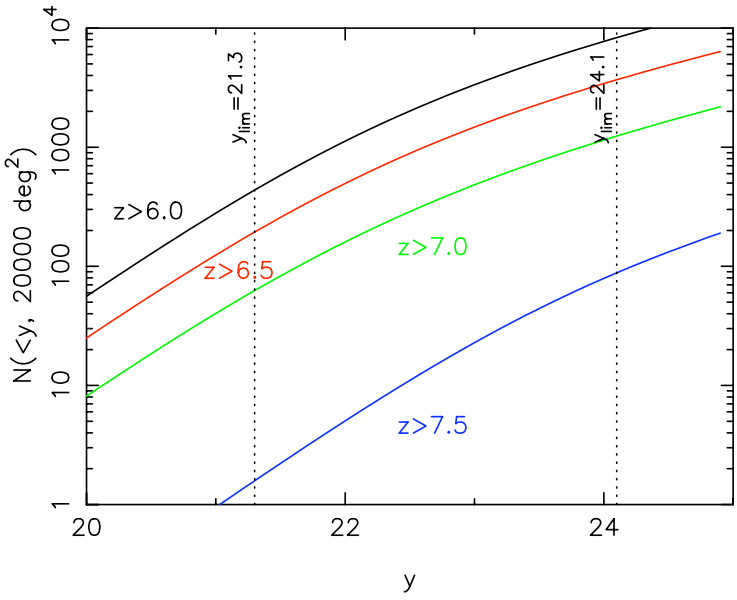


Figure 3.3 Number of high-redshift ($z > 6$) quasars expected to be discovered in a 20,000 deg^2 area as a function of redshift and limiting magnitude. The vertical dashed lines indicate the 10σ detection limit for LSST for a single visit and for the final coadded. Taken from LSST Science Collaborations and LSST Project (2009).

3.2.3 Observing Strategy

The fundamental basis of the LSST concept is to scan the sky deep, wide and fast with a single observing strategy. LSST will use the six bandpasses *ugrizy*, of which five will be concurrent in the camera at a time.

The chosen LSST science themes provide direct and indirect motivation for a sky coverage of 20,000 deg^2 and a coadded depth of $r{\sim}27.5$ mag, and for a number of other system parameters. They also motivate for a uniform cadence, so ∼90% of the time will be spent on a uniform survey. The remaining observation time will then be used to carry out very deep (single-visit $r{\sim}26$) observations, observations with very short revisit times (<1 minute), as well as observations dedicated to regions such as the ecliptic, Galactic plane, and the Magellanic Clouds.

From LSST's site on Cerro Pachòn in the Northern Chilean Andes, observations can be carried out for sky regions with $\delta < 35°.5$, corresponding to an observable area of 31,000 deg^2. However, the main survey is only covering 20,000 deg^2, as it avoids parts of the Galactic plane around the Galactic center. In these regions, the high stellar density would lead to a confusion limit at much brighter magnitudes than in the rest of the survey. For this reason, around the Galactic center, 30 observations in each of the LSST's filters are scheduled with a roughly logarithmic distribution in time.

The expected airmass of 2.2 or less will result in a 0.6 mag loss of sensitivity at 500 nm compared to an observation in the zenith (LSST Science Collaborations and LSST Project 2009). Sky regions with $-75^\circ < \delta < 15^\circ$ can be observed with an airmass of 1.4 or smaller, providing especially good image quality for weak lensing.

During the survey mission of 10 years, each patch of the sky will be visited 1000 times (where a "visit" is defined as a pair of 15-second exposures, performed back-to-back in a given filter, and separated by a four-second interval for readout and opening and closing of the shutter). 1000 visits implies that a single-visit depth is $r{\sim}24.5$ mag, which is consistent with the coadded depth constrain of $r{\sim}27.5$. This will produce data with time-resolved astrometric and photometric data for 20 billion objects.

For scheduling the observations, preference is given to the r and i band observations in the presence of good seeing and low airmass. LSST will visit each field as often as possible twice with visits in a 15–60 min time interval. This part of the observing strategy enables linking detection of moving objects in order to provide motion vectors. Additionally, this time sampling enables the measurement of short-period variability. Planning of observations is also done by ensuring that the visits to each field are widely distributed in both position angle on the sky and rotation angle of the camera. This should minimize systematics in the observations.

3.2.4 Photometry

The LSST filter set (*ugrizy*) is modeled on the system used for SDSS, covering the available wavelength range with roughly logarithmic spacing and at the same time avoiding the strongest tellurium features and sampling the Balmer break (LSST Science Collaborations and LSST Project 2009). The system is extended to the y band, in comparison to the SDSS filter set ending with the z band, because the deep-depletion CCDs offer high sensitivity to 1 μm. The transmission curves of LSST, in comparison to PS1, are given in Fig. 3.5 at the end of the Section.

In the following, photometric system capabilities are given as stated in LSST Science Collaborations and LSST Project (2009):

(i) Single-visit depths (point sources, 5σ): $u < 23.9$, $g < 25.0$, $r < 24.7$, $i < 24.0$, $z < 23.3$, $y < 22.1$ in AB mag

(ii) Baseline number of visits over 10 years: u: 70, g: 100, r: 230, i: 230, z: 200, y: 200

(iii) Coadded depths (point sources, 5σ): $u < 26.3$, $g < 27.5$, $r < 27.7$, $i < 27.0$, $z < 26.2$, $y < 24.9$ in AB mag

(iv) Photometry accuracy (rms mag): repeatability 0.005, zeropoints 0.01.

Fig. 3.4 gives the coadded r band depth over the survey lifetime.

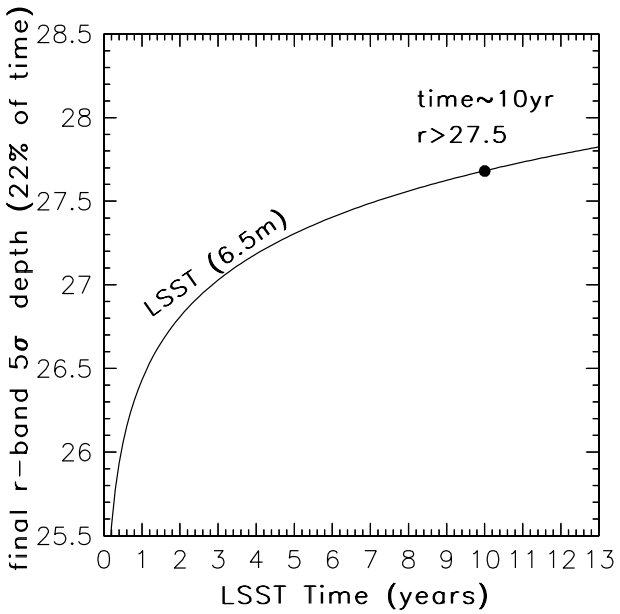


Figure 3.4 The coadded depth of LSST in the r band (AB magnitudes) vs. the survey lifetime. 22% of the total observing time (corrected for weather and other losses) will be allocated for the r band. The ratio of the surveyed sky area to the field-of-view area will be 2,000. Adapted from LSST Science Collaborations and LSST Project (2009).

3.2.5 Data releases

Each night, LSST will generate about 15 TB of data. The total amount of data collected over the ten years of operation will be 60 petabytes (PB), and processing this data will produce a 15 PB catalog database.

To handle this tremendous amount of data, the LSST data management system (DMS) (LSST Science Collaborations and LSST Project 2009) will reduce the raw data to generate data products and to make them available to scientists and the public. It will continuously process the incoming stream of images in order to produce real-time transient alerts and to archive the raw images. About once a year, a Data Release (DR) will be produced, being a stable self-consistent collection of data products taken from the beginning of the survey mission to the cutoff date set for the DR in case. In the end, there will be eleven data releases. The DMS also produces periodically calibration data products such as flat fields.

Real-time alerts of LSST discoveries will be available on a webpage, additionally an auto email alert service will be provided. This will permit users to custom filter alerts based on a number of parameters. LSST alerts as well as educational programs will be available world-wide. Catalogs and images itself will be available to scientists in the US and Chile, as well as to international institutions that are supporting LSST operations.

The underlying DMS is developed as a new, general-purpose, high-performance, scaleable, well documented, open source data processing software stack for O/IR surveys in general. Prototypes

of this stack form the basis of the Hyper-Supreme Cam (HSC) Survey data reduction pipeline (LSST Science Collaborations and LSST Project 2009).

3.3 The Capabilities of PS1 3π as LSST Pilot Survey

Based on the properties of PS1 3π and LSST, together with analysis carried out during this work (see Chapter 5), it turns out that PS1 3π can serve as a valid pilot survey for the upcoming LSST.

In the following, indicators for why PS1 3π can serve as a pilot survey are summarized:

purpose:
Both PS1 3π and LSST have a wide variety of science drivers, many of them overlapping and the ones new with LSST are possible mostly due to higher cadence and deeper magnitude limits. For two of the main LSST goals – mapping the Milky Way and especially its halo, observing AGN –, methodology for analyzing time-domain data was developed for and tested with PS1 3π as part of this work.

sky coverage:
Despite observing different parts of the sky, both surveys have about the same size of sky coverage and have an overlapping region. The high sky coverage of 30,000 deg^2 for PS1 3π, and of 20,000 deg^2 for LSST, puts constraints on the methods suitable for analyzing the data. Methods not needing much intervention after a test phase, methods being reliable, delivering automatically stochastic constraints and being able to deal with various kinds of variable sources and being fast are needed. These requirements are placed for data from PS1 3π, and even more from the upcoming LSST. Methods for feature extraction from light curves, as well as machine learning approaches that infer from such features were developed throughout this work and can be applied also to LSST data. Methods being capable of evaluating sparse data in a reliable manner can help especially during the first stages of LSST.

bands and magnitude limit:
Both surveys have similar bandpasses, $g_{\rm P1}$,$r_{\rm P1}$,$i_{\rm P1}$,$z_{\rm P1}$,$y_{\rm P1}$ for PS1 3π, *ugrizy* for LSST. The LSST's additional u band, known from SDSS, is extremely powerful for separating low-redshift QSOs from hot stars. However, variability selection is crucial for low-luminosity AGN, which cannot be selected by color as they would be swamped by their host galaxies. Variability and variability-color selection was tested extensive within PS1 3π.
LSST will look much deeper than PS1 3π. Especially, it will be able to map the Milky Way to greater distances and detail. This will provide a vast amount of sources, again requiring methods being reliable, highly automated and fast.

cadence:
Both surveys are dedicated to time-domain science. LSST will have a much higher cadence and will be almost simultaneous, while PS1 3π has a relatively sparse sampling and is not simultaneous.

For this reason, methods capable of dealing with LSST may be allowed to comply with somewhat less stringent requirements than methods for PS1 3π data. However, all methods developed here for the non-simultaneous PS1 3π data are very general, so they can be very easily used for simultaneous time-series data. Additionally, such methods can help for the case of not perfectly simultaneous data, as LSST will observe through 5 out of 6 filters coincident. Methods for dealing with non-simultaneous data are also necessary for incorporating data of other surveys, either being non-simultaneous itself or not meeting the LSST cadence.

duration:
LSST will have an observational baseline more than twice as long as PS1 3π. Methods for PS1 3π are developed under the aspect of dealing with such a relatively short baseline in combination with the sparse cadence. For this reason, methods developed for PS1 3π can be helpful especially during the first months and years of LSST.

data storage and management:
For both PS1 3π and LSST, data storage, calibration and management is a huge challenge due to the amount of data. Of course, this task will be ways more challenging for LSST. PS1 3π uses PSPS (Public Science Product Subsystem) based on Large Survey Database (LSD Jurić et al. 2011) for data storage, enabling standard queries for position and cone search, as well as complex queries in order to search for sources with very specific characteristics. LSD is a Python framework and DBMS for distributed storage, cross-matching and querying of large survey catalogs ($>10^9$ rows, >1 TB), optimized for fast queries and parallelization for typical requirements on queries in astronomy.
Using such an environment is essential as for working with local flat files, as often done for SDSS, 2MASS and other surveys, is not longer feasible for PS1 3π, LSST and other upcoming synoptic sky surveys. Data from such surveys are also unsuitable for processing on a single machine, so parallelization is crucial for processing and calibrating, providing, and evaluating data.

The experiences gathered from processing large surveys such as PS1 3π were used in order to develop a software stack for LSST. LSST will use a comprehensive Data Management system (Jurić et al. 2015) to process the amount of about 15 TB per night. It will produce data products at several "levels", including real-time alerts as well as data releases and added value catalogs. To carry out this task, a new general-purpose, high-performance open source data processing software stack was developed. Prototypes were tested with processing data from existing surveys such as SDSS.

In the following, a summary on various science metrics for PS1 3π, LSST are given in Table 3.1, with information from Metcalfe et al. (2013), Schlafly et al. (2014) and LSST Science Collaborations and LSST Project (2009).

Additionally, the curves for PS1 3π and LSST are given in Fig. 3.5.

Table 3.1. Comparison of Different Surveys

	PS1 3π	LSST
sky area	30,000 deg^2	20,000 deg^2 for main survey, up to 31,000°
sky region	$\delta > -30°$	$\delta < 10°$ for main survey, up to $\delta < 35.5°$
filters	$g_{\rm P1}$,$i_{\rm P1}$,$r_{\rm P1}$,$z_{\rm P1}$,$y_{\rm P1}$	*ugrizy*
single exposure depths (point sources, 5σ)	$g_{\rm P1} < 22.0$, $r_{\rm P1} < 22.0$, $i_{\rm P1} < 21.9$, $z_{\rm P1} < 21.0$, $y_{\rm P1} < 19.8$	single-visit depths $u < 23.9$, $g < 25.0$, $r < 24.7$, $i < 24.0$, $z < 23.3$, $y < 22.1$ mag
coadded depths (point sources, 5σ)	$g_{\rm P1} < 23.4$, $r_{\rm P1} < 23.4$, $i_{\rm P1} < 23.2$, $z_{\rm P1} < 22.4$, $y_{\rm P1} < 21.3$	$u < 26.3$, $g < 27.5$, $r < 27.7$, $i < 27.0$, $z < 26.2$, $y < 24.9$ mag
median seeing FWHM	1.1"	1.0"
cadence	67 epochs over 5.5 years	1000 visits over 10 years
nightly data volume	1 TB	30 TB
catalog data volume	$\sim$100 TB	30 PB

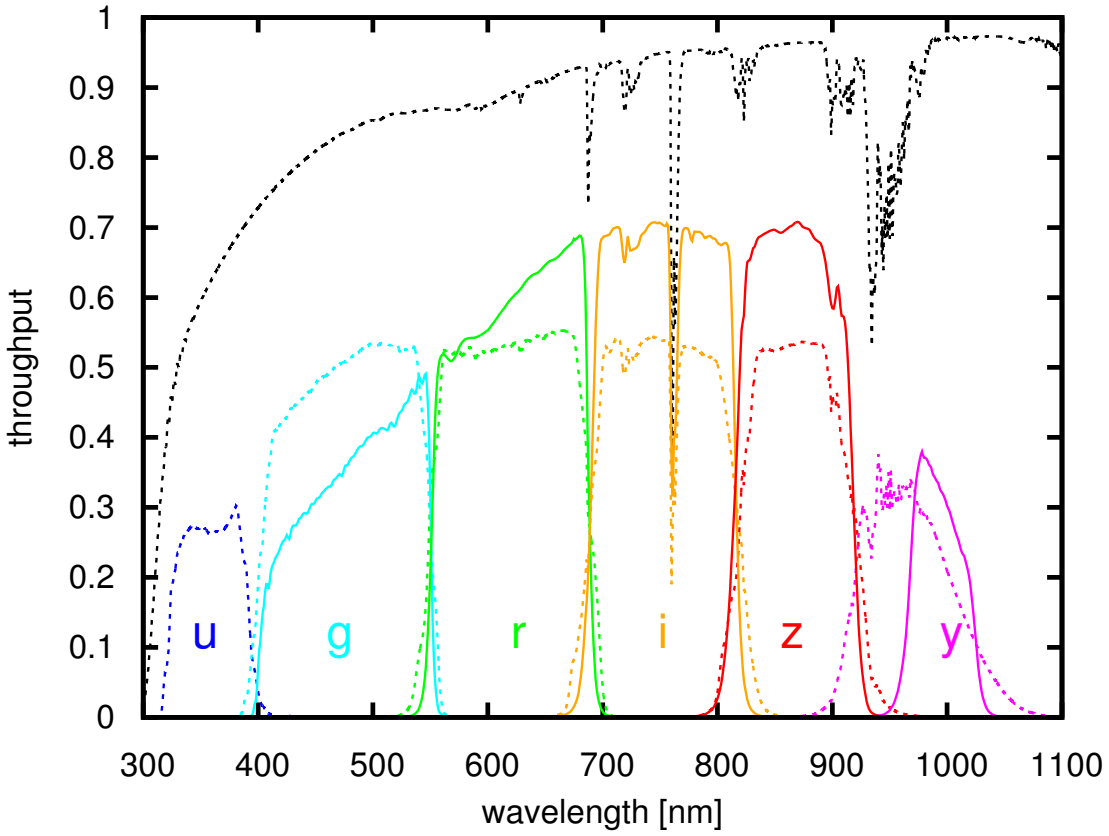


Figure 3.5 Comparison of the total throughput of Pan-STARRS1 bandpasses $g_{\rm P1}$,$i_{\rm P1}$,$r_{\rm P1}$,$z_{\rm P1}$,$y_{\rm P1}$ (solid lines) and LSST bandpasses u,g,r,i,z,y (dashed lines). The dashed black line gives the throughput of a standard atmosphere. The primary differences are the additional u band in LSST, the greater near-IR sensitivity thanks to the wider y band.
Pan-STARRS1 bandpasses from Tonry et al. (2012), standard atmosphere and LSST bandpasses from the LSST project (https://github.com/lsst/throughputs/blob/master/baseline/README.md)

Chapter 4

Classifying Variable Sources in Non-Simultaneous Multi-Color Surveys

Classification of variable sources relies fundamentally on algorithms quantifying different aspects of variability found in light curves. Since light curve data are, in general, sampled at irregular intervals and span different bandpasses, classifying variable sources directly based on their light curves would be both too challenging and too erroneous. For this reason, light curves are transformed into a set of numbers (or higher-dimension analogs) describing their variability characteristics, so-called *features*. This process, *feature extraction*, uses various methods generally known from signal processing, as well as methods tailored to astronomical time series.

The purpose of the first part of this work is to build a many-class classification framework by proper feature extraction and selection in the presence of noise and spurious data, and fast and reliable classification based on those features.

The main challenges ahead of nowadays synoptic time-domain surveys are the timely identification of interesting transients in the vast amount of photometric data for maximizing the utility of the follow-up observations, as well as identification and classification of variable sources used mainly for cosmological studies and studies of the Milky Way and Local Group.

Some methods draw on the methods developed earlier on for characterizing variability in single-band data, so the description of possible methods applied to multi-band time domain surveys starts with these methods.

Within Chapter 4, the methodological concepts of finding variable sources, quantifying their astrostatistical properties as well as automated source classification by machine-learning methods are introduced. Here a focus is given to the question how to deal with the challenges that come up when developing such methodology for non-simultaneous multi-band surveys. Additionally to presenting known variability measures, such as single-band structure functions, in this Chapter also the newly developed multi-band structure function fitting is outlined in great detail.

This Chapter also deals with automated classification of sources by machine-learning classifiers and with the question how to test and quantify the reliability of such methods. Here, known machine-learning concepts from the literature are given that are tailored to the specific science cases of this thesis and applied in the following chapters to Pan-STARRS1 3π data.

4.1 Identifying Significantly Varying Sources in Single-Band Light Curves

In contrast to targeted observations carried out for a specific science case, survey data have often drawbacks like a scheduling of observations that doesn't match exactly the requirements of analysis. To alleviate this disadvantages and make full usage of the advantages of survey data, carefully chosen analysis methods are necessary.

In general, some of the following properties may be found in time series data (Falk 2012):

- the data is not generated independently
- their dispersion varies in time
- they are often governed by a trend and/or have periodic components.

In the last decades, several time-series analysis methods have been developed to study the properties of variable sources while trying to overcome the limitations of the data. Most of such analysis methods work on single bandpasses, and their ability to deal with measurement errors as well as data gaps differs. Application of single-band methods to multi-band time-domain surveys is possible, if the light curve has high enough sampling to refer to only one bandpass, one is interested in independent light curves in different bands, or used to construct multi-band methods based on them.

Features should be chosen in a way that involves less computation effort – all-sky time-domain surveys come with a tremendous number of sources – and should be as informative and discriminative as possible, thus allowing machine learning to use them to distinguish between classes of light curves. Such features can range from basic statistical properties such as the mean or standard deviation, to more complex time series characteristics such as the autocorrelation function or the structure function.

For variable stars, features fall into two categories: those that are related to the period of a source, ant those that are not. In the following, an overview of common methods for feature extraction from periodic as well as aperiodic light curves is given. Methods being related to such used later on in the analysis of PS1 3π light curves are highlighted in greater detail.

4.1.1 Single-Band Periodic Light Curve Features

Periodicity is the most prominent appearance of light curves. Periodicity can be found e.g. in light curves of RR Lyrae and Cepheids. However, periodicity is not always present, and if the source would be theoretically variable, this can be masked due to the cadence of the survey. For this reason, it is important to apply feature-extraction methods that can deal with sparse and unevenly sampled data in order to detect periodicity.

Lomb-Scargle Periodogram

The Lomb-Scargle periodogram (Scargle 1982) is a common tool applied to time series for period finding and frequency analysis. As it is able to handle unevenly spaced data points – as typically found in time-domain surveys, at least after outlier cleaning –, preference is given over the Discrete Fourier Transform (DFT).
The algorithm decomposes time series into linear composition of cosine and sine waves of the form $y = a\cos\omega t + b\sin\omega t$, carrying out a transformation from the time domain to the frequency domain. The Lomb-Scargle periodogram is defined as

$$P(\omega) = \frac{1}{2\sigma^2} + \left\{ \frac{\left[\sum_{n=1}^{N}(m_n - \bar{m})\cos\left[\omega(t_n - \tau)\right]\right]^2}{\sum_{n=1}^{N}\cos^2\left[\omega(t_n - \tau)\right]} + \frac{\left[\sum_{n=1}^{N}(m_n - \bar{m})\sin\left[\omega(t_n - \tau)\right]\right]^2}{\sum_{n=1}^{N}\sin^2\left[\omega(t_n - \tau)\right]} \right\} \tag{4.1}$$

where $\omega = 2\pi T$, T being the period, and the time offset τ is defined by:

$$\tan(2\omega\tau) = \sum_{n=1}^{N}\sin(2\omega t_n) / \sum_{n=1}^{N}\cos(2\omega t_n). \tag{4.2}$$

Once the period is known, periodic light curves can be transformed so that each period is mapped onto the same time axis, known as *phase folding*.

Given the period $\mathcal{P}$ for a light curve with observations at time t_i, phase-folding replaces the time axis by a phase $\phi = \left(\frac{t-t_0}{\mathcal{P}}\right) - E(t)$ (Hoffmeister et al. 1985), where t is the time of an observation, t_0 is some reference time, $\mathcal{P}$ is the period, and $E(t)$ indicates the integer part of $(t - t_0)/\mathcal{P}$. This results in replacement of the time axis by a phase axis, ranging $[0, 1[$. An example is shown in Fig. 2.2.

4.1.2 Single-Band Non-Periodic Light Curve Features

In seeking to classify variable source light curves, it is not always possible to characterize flux variations by detecting and characterizing periodicity. Reasons for this are both the non-periodic nature of various variable sources (such as quasars), as well as time series who lack of periodic information due to their sampling. Also, they are helpful for data sets assumed to be composed of non-variable as well as variable sources of different classes: Non-periodic light curve features can be helpful in determining e.g. variability amplitude and time scale for all light curves in a given survey, in order to identify candidates for specific classes of variable sources, among them periodic as well as aperiodic variables.

A summary on various features for single-band time series data is given in Nun et al. (2015). Despite not used here, it provides a range of tools for single-band data, of which some might be generalizeable to the multi-band case.

Single-Band Structure Functions

Beyond simply establishing variability (by rejecting a null hypothesis of time-independent fluxes), variable sources can and should be characterized by their variability amplitude and the timescales over which they vary.

A useful and well-established tool in the field of variability is the structure function (Hughes et al. 1992; Collier and Peterson 2001; Kozłowski et al. 2010) which measures the mean squared magnitude difference for pairs of observations m_i, m_j that are separated by a given time lag, t_{ij}, where $V(t_{ij}) = \langle [m_i - m_j]^2 \rangle$. The structure function is commonly characterized in terms of a Damped Random Walk (DRW, see A.1.5) or a power law.

For a DRW, the structure function is specified by two parameters, ω and τ, and is given by

$$V(t_{ij}|\tau,\omega) = \omega^2(1 - e^{-|t_{ij}|/\tau})\,. \tag{4.3}$$

In this notation, ω^2 reflects the expectation value for the squared magnitude difference, m_{ij}^2, among measurements separated in time by t_{ij}. τ is called the decorrelation time of the DRW.

When parameterizing single-band variability using a power-law model for the structure function, the structure function is instead specified by two parameters, then amplitude A and the power law index γ as

$$V(t_{ij}|A,\gamma) = A\left(\frac{t_{ij}}{1\ \mathrm{yr}}\right)^{\gamma}. \tag{4.4}$$

The source variability is then characterized by two structure function parameters, (ω,τ) for the DRW or (A,γ) for the power law, usually estimated by examining a likelihood function on a parameter grid or by using MCMC (see A.2).

Objects of different classes typically occupy different regions in structure function parameter space. The fitted structure function parameters can be used to select remarkably pure and complete samples of variable sources of certain classes, which makes selection by structure function parameters an efficient approach for both selecting stochastically varying and periodic variable objects. Single-band structure function fitting in order to select samples of QSOs and RR Lyrae was carried out e.g. by Schmidt et al. (2010). They selected complete and pure samples based on the intrinsic SDSS S82 r-band light curves, characterized by a power-law structure function.

The structure function is considered by several researchers (e.g. Collier and Peterson 2001; Zu et al. 2011; Hernitschek et al. 2015) to be an ideal method for studying the time-domain properties of samples consisting irregularly sampled light curves of various classes. They can also used for other light-curve related tasks, such as reverberation mapping, a technique to estimate a Black Hole's mass by measuring its broad-line region (Zu et al. 2011; Hernitschek et al. 2015).

Historically, the structure function has been used in the study of turbulent plasmas (Kolmogorov 1941a,b). It was later introduced to astrophysics by radio astronomers studying slow oscillations

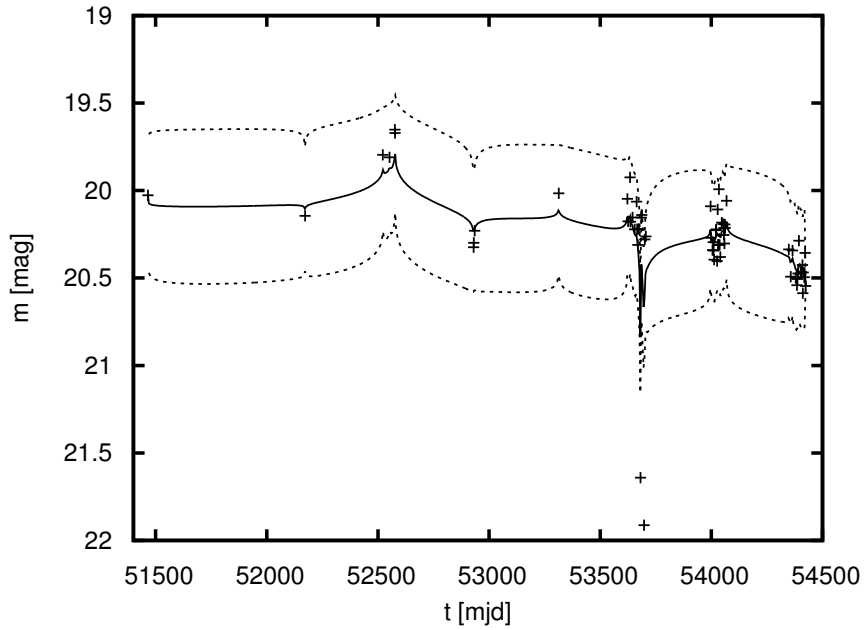


Figure 4.1 Example structure function fit for a SDSS S82 QSO in the r band. The solid black line shows the expected mean light curve and the dashed lines shows the spread of light curve realizations about the mean being consistent with the measurements.

in the interstellar medium (Rickett et al. 1984). The first systematic description of the structure function methodology adjusted to the needs of astronomical data sets was made by Simoneti et al. (1985) to demonstrate that the time-series of flat- and deep-spectrum radio sources differ qualitatively. During the same period, Cordes and Downs (1985) estimate the structure function of 21 pulsars, and Hjellming and Narayan (1986) derived the first quantitative structure function results for the compact galactic radio source 1741-038. Subsequently, the structure function has been employed for the study of the timing properties of higher energy bands.

Fig. 4.1 shows an example structure function fit, applied to the r band of a SDSS S82 QSO. The expected mean light curve shape, as well as the spread of possible realizations being consistent with the measurements are shown.

4.1.3 Multi-Band Periodic Light Curve Features

Multi-band all-sky surveys have the power to detect a huge amount of variable sources, among them periodic variables such as RR Lyrae and Cepheids. During source classification, candidates for such sources can be found by applying non-periodic methods. However, it is very important to find out their period to make use of e.g. period-luminosity relations to infer their distance, and also to get cleaner candidate samples by using the estimated period as feature in subsequent classification.

The multi-band nature of such data makes period estimation challenging, especially when the light curves are also non-simultaneous. There are a couple of methods dealing with this issue. In this section, two methods that are also applied to PS1 3π light curves are outlined.

Multi-Band Periodogram

The multi-band periodogram (VanderPlas and Ivezić 2015) is a generalization of the Lomb-Scargle approach (see Section 4.1.1).

The light curves in each band are modeled as arbitrary truncated Fourier series, with the period and phase shared across all bands. For this purpose, the model is composed of a N_{base}-term truncated Fourier "base model" that models the overall variability shared among all K, and a set of N_{band}-term truncated Fourier fits so that each of it models the residual within a single band from the shared variability accounted for in the base model.

The total number of parameters used for K filters is then $M_K = (2N_{base} + 1) + K(2N_{band} + 1)$. The model of the observed magnitudes is then:

$$\begin{aligned} y_k(t|\omega,\theta) = \quad & \theta_0 + \textstyle\sum_{n=1}^{N_{base}} \left[\theta_{2n-1}\sin(n\omega t) + \theta_{2n}\cos(n\omega t)\right] \quad + \\ & \theta_0^{(k)} + \textstyle\sum_{n=1}^{N_{band}} \left[\theta_{2n-1}^{(k)}\sin(n\omega t) + \theta_{2n}^{(k)}\cos(n\omega t)\right]. \end{aligned} \tag{4.5}$$

An important property of this model is that the base parameter θ is shared among all bands, whereas the offsets $\theta^{(k)}$ are determined individually.

As an approach for period finding in multi-band light curves, VanderPlas and Ivezić (2015) suggest a *hybrid strategy* composed of the multi-band periodogram and template fits (see below):

(i) Apply the multi-band periodogram to find candidate periods. This algorithm is relatively fast and can be parallelized.
(ii) Apply a template-fitting algorithm to each candidate period. This step is more computationally intensive, so the previous step is required for pre-selection.
(iii) Evaluate the fits found by the template fitting using a goodness-of-fit statistic. If none of the candidate periods is suitable, the template-fitting algorithm should be applied across the full period range.

Template Fitting

Template fitting uses light-curve templates – either synthetic or from other surveys – in order to fit them to light curves of presumably variable sources.

Such methods are specifically tailored to the class of variable sources one is looking for, and often also to sub-classes. An example on template fitting is given in Sesar et al. (2010), who build a system of SDSS *ugriz* RR Lyrae templates, containing both RRab and RRc templates. They make use of the relatively large sample of RR Lyrae within SDSS S82, having densely sampled

light curves as needed for making templates. Within S82, Sesar et al. (2010) pick light curves with high S/N, apply period folding and subsequent B-spline interpolation. Among them, a set of light curves smoothly covering the parameter space is chosen.

The usage of synthetic light curves can be very helpful if it is not possible to get real sample light curves, but if possible, preference is given to real observational light curves of the desired class.

4.1.4 Multi-Band Non-Periodic Light Curve Features

To characterize variability sufficiently in order to identify and classify variable sources, features being capable of describing multi-band light curves and not demanding periodicity are needed. In the work at hand, the following non-periodic features are used: a generic and non-parametric measure derived from χ^2 statistics, and a novel generalization of structure functions to multi-band light curves.

χ^2-based Variability Quantity

As a very generic and non-parametric measure to characterize variability, the significance of variability of a light curve can be defined by

$$\hat{\chi}^2 = \frac{\chi^2_{\text{source}} - N_{\text{d.o.f}}}{\sqrt{2\,N_{\text{d.o.f}}}}, \tag{4.6}$$

with

$$\chi^2_{\text{source}} = \sum_{\lambda} \sum_{i=1}^{N} \frac{(m_{\lambda,i} - \langle m_\lambda \rangle)^2}{\sigma^2_{\lambda,i}} \tag{4.7}$$

where N is the total number of photometric points for one object across all n bands, $m_{\lambda,i}$ denotes a magnitude measured in band λ, $\langle m_\lambda \rangle$ denotes the mean magnitude in band λ, the sum over λ is over the PS1 bands g_{P1},r_{P1},i_{P1},z_{P1},y_{P1}, and $N_{\text{d.o.f}} = N - n$ is the number of degrees of freedom. Assuming that most of the sources are not variable, the distribution of $\hat{\chi}^2$ is expected to be a unit Gaussian distribution. In contrast, varying sources should form a "tail" of higher $\hat{\chi}^2$.

This is applied in Section 5.4.1 to PS1 3π data.

Multi-Band Structure Functions

The cadence of surveys like the SDSS provides data that allow application of the usual single-band formulation of structure functions. However, the cadence of PS1 3π data, which observes non-simultaneous in different bands with a small number of epochs per band (see Sec. 5.2), makes it necessary to extend this approach for multi-band fitting. This approach, as developed within this thesis, will turn the light curves in each band into an overall light curve that pools all the information while keeping track of possibly different variability amplitudes in different bands.

The model outlined here is applied in Chapter 5 to PS1 3π data and published in Hernitschek et al. (2016).

If objects would show the same kind of variability in all observed bands, implementing such an approach would simply entail determining the (time-averaged) mean color of the object and shifting the light curves in the different bands to a common magnitude. However, most astrophysical objects show wavelength depending variability, i.e. vary more at shorter wavelengths. To account for this, the new multi-band model presented here has, beyond ω and τ, a set of temporal mean magnitude parameters in each PS1 band, $\vec{\mu}$, and it links the variability *amplitudes* $\omega(b)$ in different bands b by a power law with exponent α. Specifically,

$$\alpha = \frac{\log(\omega(b)/\omega(r))}{\log(\lambda_b/\lambda_r)}\,, \tag{4.8}$$

where λ_b is the effective wavelength of the band b.

To assign a likelihood to an object's photometry, given a structure function model, this model makes use of a Gaussian Process formulation (see A.1) for stochastic source variability. In contrast to single-band structure function models (e.g. Rybicki and Press 1992; Zu et al. 2011; Hernitschek et al. 2015), the Gaussian Process is not applied to any particular band but instead to an arbitrarily constructed fiducial band which can be scaled and shifted onto the particular bands. This permits simultaneous treatment of multiple bands, without requiring any simultaneous or near-simultaneous observations. This makes the method ideal for application to surveys such as PS1 3π, and is also helpful in cases where a survey with in principle simultaneous observations becomes non-simultaneous due to outlier cleaning.
It is key in this context to realize that the fiducial band is a latent variable – it is never directly observed; only the scaled and shifted versions are observed, where substantial measurement noise is present.

The fiducial light curve can be described with a Gaussian process having zero mean and unit characteristic variance, as done in the case of the single-band DRW model by Zu et al. (2011). That is, the prior probability distribution function (pdf) for a set of N fiducial "magnitudes" $\vec{q}$ that are instantiated at observed times t_n is a multivariate normal distribution:

$$p(\vec{q}) = \mathcal{N}(\vec{q}\,|\,0, C^q)\,, \tag{4.9}$$

where C^q is a $N \times N$ symmetric positive definite covariance matrix. In the case of a DRW model, C^q is given by

$$C^q_{nn'} = \exp\left[-\frac{|t_n - t_{n'}|}{\tau}\right]\,. \tag{4.10}$$

This is identical to the usual single band DRW covariance matrix, except for dropping a scale factor ω^2 from Equ. (4.10), because the fiducial band q was defined to have unit variance. This factor reappears in our multi-band structure function through the scale factors that link the fiducial band to observed bands.

Consider now a given source having N observations across $N_{\rm band}$ different bands. The data consist of the magnitude and uncertainty vectors $\vec{m}$ and $\vec{\sigma}$, the times of observation t_n, and the corresponding bands b_n. The source also has $N_{\rm band}$ temporal mean magnitudes $\vec{\mu}$. The $N \times N_{\rm band}$ matrix $\mathbb{M}$ is defined so that

$$\mathbb{M}\vec{\mu} = [\mu(b_1), \mu(b_2), \cdots, \mu(b_N)]\,. \tag{4.11}$$

The likelihood of an individual measurement m_n, given its observational uncertainty σ_n and a value for the corresponding fiducial magnitude q_n, is found by shifting and scaling the fiducial magnitude and adding Gaussian noise. This makes the single-datum likelihood

$$p(m_n \,|\, q_n, b_n, \sigma_n^2) = \mathcal{N}(m_n \,|\, \omega(b_n) q_n + \mu(b_n), \sigma_n^2), \tag{4.12}$$

where $\omega(b_n)$ is the variability in bandpass b_n relative to the unit variability of the unobserved fiducial band.

Introducing the diagonal $N \times N$ matrix Ω, defined by $\Omega_{ii} = \omega(b_i)$, the full likelihood is given by

$$p(\vec{m} \,|\, \vec{q}, \Sigma) = \mathcal{N}(\vec{m} \,|\, \Omega\vec{q} + \mathbb{M}\vec{\mu}, \Sigma^2)\,, \tag{4.13}$$

where Σ is a diagonal matrix with $\Sigma_{ii} = \sigma_i$. Because everything is Gaussian, the latent fiducial magnitudes never have to be explicitly inferred; they can all be marginalized out analytically. This marginalization leads to the likelihood given the model, and the covariance matrix of the data:

$$p(\vec{m} \,|\, \text{structure function parameters}, \vec{\mu}) = \mathcal{N}(\vec{m} \,|\, \mathbb{M}\vec{\mu}, C) \tag{4.14}$$

$$C = \Omega C^q \Omega + \Sigma^2\,. \tag{4.15}$$

This is identical to the case of a single-band DRW model, except the rows and columns of C^q are scaled by amplitudes $\omega(b_n)$, $\omega(b_{n'})$ for the bands b_n and $b_{n'}$, and a contribution from the photometric uncertainties is added to the diagonal:

$$C_{nn'} = \omega(b_n)\,\omega(b_{n'})\,\exp\left[-\frac{|t_n - t_{n'}|}{\tau}\right] + \sigma_n^2\,\delta_{nn'}. \tag{4.16}$$

Equations (4.14) through (4.16) provide a method for computing the probability of any set of observed magnitudes m, given their meta data (t_n, b_n, σ_n^2) and their structure function parameters $\omega(b)$, τ.

The most interesting result here will be the structure function parameters and are relatively uninterested in the exact mean magnitudes $\vec{\mu}$. This is exactly the same situation as in Zu et al. (2011). Following that work, the likelihood of the structure function parameters, given the multi-band data, marginalized over $\vec{\mu}$, is given by:

$$p(\vec{m} \,|\, \text{structure function parameters}) = \mathcal{L} \propto |C|^{-1/2} |C_\mu|^{1/2} \exp\left(-\chi^2/2\right) \tag{4.17}$$

where

$$\begin{aligned} C_\mu &= (\mathbb{M}^{\mathrm{T}} C^{-1} \mathbb{M})^{-1} \\ \chi^2 &= (\vec{m} - \mathbb{M}\vec{\mu})^{\mathsf{T}} C^{-1} (\vec{m} - \mathbb{M}\vec{\mu})\,. \end{aligned} \tag{4.18}$$

It is important to note that the factor of $|C_\mu|^{1/2}$ in Equation (4.17) comes from the marginalization over $\vec{\mu}$. Maximization over $p(\vec{m}\,|\,\mathrm{SF\ parameters})$ is done to obtain best fit values of the structure function parameters. Thus, $\vec{\mu}$ is obtained as the maximum likelihood values of $\vec{\mu}$ given the structure function parameters. That is, the mean magnitudes are given by

$$\vec{\mu} = (\mathbb{M}^T C^{-1} \mathbb{M})^{-1} \mathbb{M}^{\mathrm{T}} C^{-1} \vec{m}\,,$$

and have variance C_μ.

Interpolating Multi-Band Light Curves with Uncertainties

One advantage of this approach is that it can be used to predict unobserved data based on observed data. Because both the process is Gaussian and the noise is assumed to be Gaussian, conditional predictions of the magnitudes can be made given the observed data and the structure function. The analysis is exactly the same as in Rybicki and Press (1992), with the exception of adopting the multi-band structure function C of Equation (4.16). The magnitudes $\tilde{m}_k$ at K unmeasured times t_k, taken through bandpasses b_k, conditioned on the data in hand, are given by:

$$p(\tilde{m}|\vec{m}) = \mathcal{N}(\tilde{m}|\tilde{\mu}, \tilde{C}) \tag{4.19}$$

$$\tilde{\mu} = \vec{\nu} + X \cdot C^{-1} \cdot [\vec{m} - \mathbb{M}\vec{\mu}] \tag{4.20}$$

$$\tilde{C} = Y - X \cdot C^{-1} \cdot X^{\mathsf{T}} \tag{4.21}$$

$$\vec{\nu} = [\mu(b_1), \mu(b_2), \cdots, \mu(b_K)]. \tag{4.22}$$

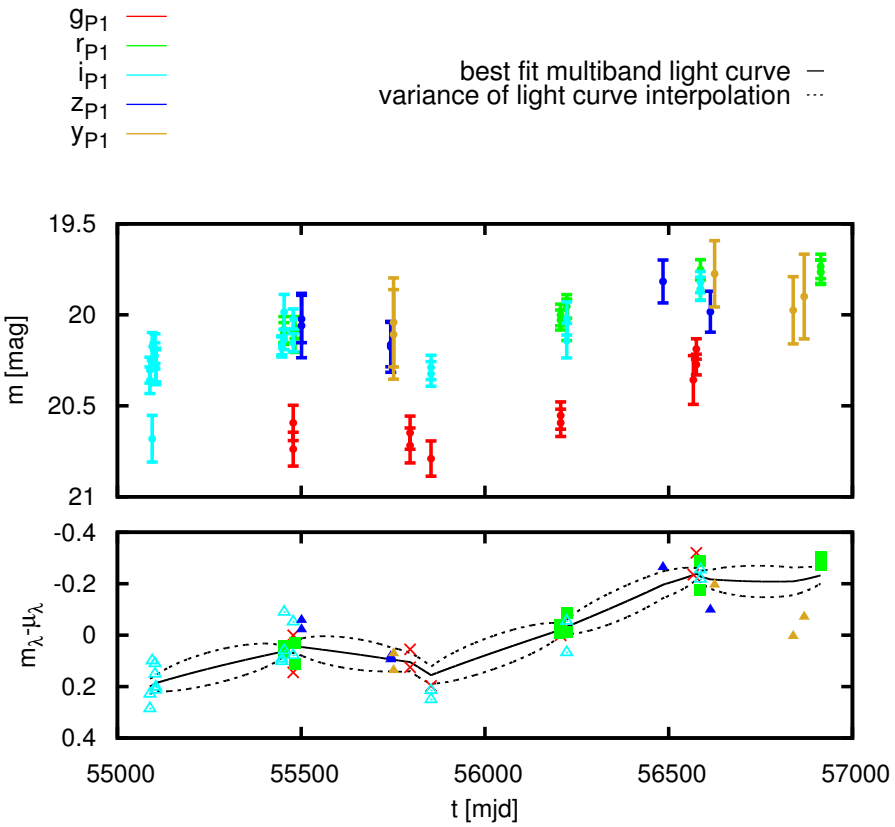
In the case of a multi-band DRW model,

$$X_{kn} = \omega(b_k)\,\omega(b_n)\,\exp\left[-\frac{|t_k - t_n|}{\tau}\right] \tag{4.23}$$

$$Y_{kk'} = \omega(b_k)\,\omega(b_{k'})\,\exp\left[-\frac{|t_k - t_{k'}|}{\tau}\right]. \tag{4.24}$$

Here $\tilde{m}$ is the column vector of conditional predictions, $\tilde{\mu}$ and $\tilde{C}$ are a conditional mean vector and a conditional variance matrix, (temporary) mean vector ν is K-dimensional, and the matrices $\tilde{C}$, X, and Y are $N \times N$, $K \times N$, and $K \times K$ respectively.

Fig. 4.2 shows an example structure function fit, applied to a PS1 3π PV3 light curve. The expected mean light curve shape, as well as the spread of possible realizations being consistent with the measurements are shown.



(a)

Figure 4.2 Example structure function fit for a PS1 3π light curve. The upper panel gives the light curve, whereas the lower panel shows the light curve fitted by a multi-band DRW structure function. The solid black line shows the expected mean light curve Equ. (4.20), and the dashed lines shows the spread of light curve realizations about the mean being consistent with the measurements, i.e. the variance Equ. (4.21) for the r band.

4.2 Classifying Variable Sources Using Machine-Learning Classifiers

Classification of variable stars – the identification of a certain source with a previously identified class – presents several challenges. First, observational data of probably variable sources represents not only a picture of the source itself, but is always influenced by noise, foreground effects (source confusion, dust-caused reddening) as well as time-sampling effects acting as a window function that may hide aspects as variability. Second, time-series data of a given survey alone provide an incomplete picture of a given source as observations at certain wavelength ranges are missing - typically, time-domain surveys observe in the optical and near-IR - and also, in most cases, spectroscopic information is missing.

In order to overcome these effects, it is necessary to carefully carry out feature extraction including correction for reddening and removing of unreliable measurements, as well as check whether it is

required to add information from other surveys by cross-matching, as well as carefully choose a classification algorithm that fits the need for accuracy and speed.

Several authors have used machine-learning methods to classify variable sources using their light curves: Eyer and Blake (2004) use a Bayesian mixture-model classifier, and Debosscher et al. (2009) experiment with several methods, including Gaussian mixture models, Bayesian networks, and support vector machines (SVM), and Hernitschek et al. (2016) apply a Random Forest Classifier to preliminary PS1 3π data.
Providing a machine-learned classification that is accurate and fast is a challenging task on many frontiers (see e.g. Eyer and Mowlavi 2008). In many cases, there may be only a light curves in a given class to build the training set, making training and validation difficult. Even with many labeled light curves provided, in the force of noisy, spurious and sparsely sampled data, there is a limit to the statistical inferences that can be gained.

Whereas the previous section described methods for analyzing time-series data (such as light curves) in order to provide so-called features, this section will now cover the topic of machine-learning algorithms that can handle these features in order to produce statements about the source classification.
One purpose of this work is to build a many-class classification framework by proper feature creation and selection in the presence of such noise and spurious data, and fast and reliable classification based on those features. Also, a formalism for evaluating the results of the classification is presented. The work makes use of features derivable from time-domain data in multiple bandpasses; in addition to innovative multi-band light curve features such as generalized multi-band structure functions, color information is used.

Classification fundamentally relies upon the ability to recognize and quantify the differences between light curves. To build a *supervised machine-learning classifier*, many light curves are required for each class of interest. Given a set of sources whose class is already known and thus having a class assigned (being *labeled* and make up the *training set*), done by e.g. methods that rely on data not present in the survey that should be examined, a classifier learns a model that describes each source's class probability as function of its features. The *training set* can be built by cross-matching sources of the survey of interest to already classified sources in other survey. The other survey's classification can rely on e.g. additional bands, complementary spectroscopic information, better light-curve sampling or higher S/N. These members of the training set are then used in the training and validation process, in order to both build a classifier that is capable of classifying new sources, as well as estimating the strengths and weaknesses of the classifier. This model is then used to automatically predict the class probabilities of new sources.

The work in hand deals only with supervised machine-learning approaches. However, there are also *unsupervised machine-learning classifiers*.

Supervised learning is the machine learning task of inferring a function from labeled training data, whereas unsupervised methods cannot depend on any labels. Unsupervised learning applies therefore methods being related to density estimation to find structure in parameter space.

In the following, only supervised machine-learning classifiers are considered, and an overview of decision tree-based classifiers will be given.

Decision tree-based classifier are a a popular method for classification and regression in statistics and machine learning since the 1980s (e.g. Breiman et al. 1984). Since about a decade, the astronomical community is using tree-based techniques for several problems. For example, tree-based classifiers have be used by Ball et al. (2006) for star-galaxy separation, by Bailey et al. (2006) to identify supernova candidates, and are an important component of books dealing with machine learning.

Decision-tree classifiers have been studied extensively in the past two decades and used heavily; for an overview, see Richards et al. (2011). Many of these studies propose heuristics to construct a tree either for optimal classification accuracy or to minimize its size.

There exists a huge range of tree-based classifiers, among them *classification and regression trees* (CART Breiman et al. 1983), *Random Forest Classifiers* (RFC Breiman 1999, 2001) and *Gradient Tree Boosting* (Friedman 2001). Tree-based classifiers are powerful because they are able to capture complicated interaction structures within the feature space, are robust to outliers, are resilient to irrelevant features and offer feature importance ranking important for improving the classifier and understanding its results. Also, they can cope with missing feature values, and are computationally efficient and scaleable for large problems (Richards et al. 2011) as present in the age of all-sky time-domain surveys.

In this section, an overview of three selected tree-based classification methods is given: classification trees, Gradient Tree Boosting Classifier and Random Forest Classifier (RFC). This section is mainly based on the introduction in classification trees by Richards et al. (2011). Additionally, methods for measuring performance of classifiers and to chose the optimal set of features are given.

4.2.1 Classification Trees

In a decision tree, an input is entered at the top and as it transverses down the tree the data is divided into smaller and smaller subsets.
Tree-based machine learning algorithms use recursively binary partitioning to split the feature space, $\mathscr{R}$, into disjoint regions, $R_1, ..., R_M$. Each split is performed with respect to a single feature, producing a partitioning of R into a set of disjoint "rectangles" in the feature space (represented by the nodes of the tree). At each step, the algorithm selects both the feature and split point that produces the smallest impurity in the two resultant nodes. The splitting process is recursively repeated in order to build a tree with multiple levels (Richards et al. 2011).

To build a classification tree, begin with a training set of (feature, class) pairs $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{N}$ where $\mathbf{X}_i$ denotes the vector of features of the i-th source in the training set, and Y_i denotes the class label of this i-th source. Y_i can take any value in $\{1, ..., C\}$ where C is the number of classes.

Following Richards et al. (2011), at node m of the tree – representing a region R_m of the feature space R – the probability that a source with features in R_m belongs to class c is estimated by

$$\hat{p}_{mc} = \frac{1}{N} \sum_{\mathbf{X}_i \in R_m} I(Y_i = c). \tag{4.25}$$

This is the proportion of the N_m training set objects in node m whose class is c. The indicator function $I(Y_i = c)$ is defined to be 1 if $Y_i = c$ and 0 else. During the tree-building process, each subsequent split is chosen among all possible features and split points so that it minimizes a measure of the resultant node impurity. Measures of the node impurity are e.g. the Gini index (Gini 1913) $\sum_{c \neq c'} \hat{p}_{mc}\hat{p}_{mc'}$ or the entropy $-\sum_{c=1}^{C} \hat{p}_{mc} \log_2 \hat{p}_{mc}$. This splitting process is repeated recursively until some pre-defined stopping criterion (such as the relative improvement in the objective function) is reached. Once a classification tree is trained on a training set $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{N}$, it is straightforward to predict the class of unseen data sets $\mathbf{X}_{\text{new}}$. Specifically, the algorithm identifies the part of the decision tree $\mathbf{X}_{\text{new}}$ resides in and then assigns a class according to that node's estimated probabilities given in Equ. (4.25). For example, if $\mathbf{X}_{\text{new}} \in R_m$, then the assigned probability that the source is of class c is

$$\hat{p}_c(\mathbf{X}_{\text{new}}) = \hat{p}_{mc}, \tag{4.26}$$

where $\hat{p}_{mc}$ is defined in Equ. (4.25). Using Equ. (4.26), the predicted class is the class for which the highest value of $\hat{p}_c(\mathbf{X}_{\text{new}})$ is reported, $\hat{p}(\mathbf{X}_{\text{new}}) = \arg\max_c \hat{p}_c(\mathbf{X}_{\text{new}})$.
The classification output for each new source can then be described either as a vector of class probabilities (giving the probability for each of the C classes) or as its predicted class (with the highest probability).

Decision trees, automatically constructed by machine learning algorithms, can generate powerful classifiers due to both their conditional structure and their high execution speed. The method shown so far tempts to construct very large trees, as they will indeed fit the training set well. However, decision trees often cannot be grown to the desired complexity because of loss of generalization accuracy on new ("unseen") data occurs. Another problem is that trees can be prone to be overly adapted to the training data, or being too complex and thus overfit data. On the other hand, constructing a very lean tree will likely not be sufficient to capture the complexity of the underlying process that led to the different classes well, and thus will be not sufficient for classifying. In the end, the appropriate size of a classification tree depends on the complexity of model necessary for the particular application at hand and hence should be determined by the data.
The standard approach to this problem is to build a large tree and then to prune this tree to find the sub-tree that performs best in verification methods like the approaches shown in Sec. 4.2.4.

Pruning back a fully-grown tree may increase the generalization accuracy at unseen data, often at the expense of the accuracy on the training data. Probabilistic methods that allow descent through multiple branches with different confidence measures also do not guarantee optimization

of the training set accuracy. Apparently there is a fundamental limitation on the complexity of tree classifiers – they should not be grown too complex to overfit the training data.

The development of *ensemble methods* has led to significant improvements in classification accuracy. Such methods grow many trees, forming an ensemble, and letting the trees "vote" for the most probable class. Carrying out such a divide-and-conquer approach improves the classification performance. The main principle behind ensemble methods is that a group of "weak" classifiers can form a "strong" one. An example of such a method is *bagging* (Breiman 1996), where for the construction of each tree a bootstrap sample (a random selection without replacement) is made from the sources in the training set: given a specific training set T, form bootstrap training sets T_k, construct classifiers $h(\mathbf{x}, T_k)$ and let these vote to form the bagged predictor.
Another example is *random split selection* (Dietterich 2000), selecting at each node a split at random from among the K best splits. *Randomized outputs* (Breiman 1998, 1999) grows trees on training sets with randomly perturbing the output of the original training set: For a fixed number s, at each node, s best splits (in terms of minimizing deviance) are found and the actual split is randomly uniformly selected from them. *Random feature selection* (Amit and Geman 1997; Breiman 1999) looks for the best split over a random subset of the features. The *random subspace* method by Ho (1998) does a random selection of a subset of features to grow each tree. *Perfect Random Trees Ensembles* Cutler and Zhao (2001) uses an extreme randomness: at each node, randomly choose a variable to split on, and on the chosen variable choose randomly uniformly a split point between two randomly chosen points coming from different classes. The *Random Forest Classifier* (Breiman (2001), see also Sec. 4.2.3 for a detailed description) is an ensemble method where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.

4.2.2 Gradient Tree Boosting Classifier

The Gradient Tree Boosting was first introduced as "Gradient Boosting Machine" by J.H. Friedmann (published later as Friedman 2001). The generalized class of algorithms as was described as "functional gradient boosting" by Mason et al. (2009). The idea was originally brought up by Breiman as "gradient boosting". Breiman who showed that boosting can be interpreted as an optimization algorithm over a suitable cost function.

Here, initially the *Gradient Boosting* as introduced in Friedman (2001) is described; this method is more general and is not mandatory using trees.

Algorithm 1: Gradient Boosting

Input: $\{(\mathbf{X}, Y_i)\}_{i=1}^{N}$: training set
$L(Y, F(\mathbf{X}))$: differentiable loss function
M: number of iterations

Output: $F_m(\mathbf{X})$

begin

initialize model with a constant value $F_0(\mathbf{X}) = \arg\min_\rho \sum_{i=1}^{N} L(Y_i, \rho)$

for $m = 1, ..., M$ **do**

(i) compute pseudo-residuals

$$\tilde{Y}_i = -\left[\frac{\partial L(\mathbf{X}_i, F(x_i))}{\partial F(\mathbf{X}_i)}\right]_{F(\mathbf{X})=F_{m-1}(\mathbf{X})} \quad \text{for } i = 1, .., N \tag{4.27}$$

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} \left[\tilde{Y}_i - \beta h(\mathbf{X}_i; \mathbf{a})\right]^2 \tag{4.28}$$

where the function $h(\mathbf{X}_i; \mathbf{a})$ is a simple parameterized function of the input variables $\mathbf{X}$, characterized by parameters $\mathbf{a} = \{a_1, a_2, ...\}$. $h(\mathbf{X}_i; \mathbf{a})$ is called the *base learner*.

(ii)

$$\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(Y_i, F_{m-1}(\mathbf{X}_i) + \rho h(\mathbf{X}_i; \mathbf{a}_m)) \tag{4.29}$$

(iii) update model

$$F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + \rho_m h(\mathbf{X}; \mathbf{a}_m) \tag{4.30}$$

output $F_M(\mathbf{X})$

Gradient boosting is typically used with decision trees (especially CART trees) of a fixed size as base learners, leading to *Gradient Tree Boosting*. For this case, Friedman (2001) gives the following modifications:

Gradient boosting at the m-th step would fit a decision tree as base learner to pseudo-residuals. For a tree with J terminal nodes, the tree partitions the feature space $\{R_j\}_1^J$ (that covers the space of all joint values of features $\mathbf{X}$) into J disjoint regions $R_{1m}, ..., R_{Jm}$. Each regression tree model has then the additive form

$$h(\mathbf{X}; \{b_j, R_j\} = \sum_{j=1}^{J} b_j 1(\mathbf{x} \in R_j) \tag{4.31}$$

The indicator function $1(\cdot)$ takes the value 1 if its argument is true, and 0 otherwise.
The model update in Algorithm (1) becomes now for a regression tree:

$$F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + \rho_m \sum_{j=1}^{J} b_{jm} 1(\mathbf{X} \in R_{jm}) \tag{4.32}$$

with $\{R_{jm}\}_1^J$ being the regions defined by the terminal nodes of the tree at the m-th iteration.

J can be adjusted for the problem at hand. It controls the maximum allowed level of interaction between variables in the model. Setting $J = 2$, will allow no interaction between variables, $J = 3$ allows the interaction between up to two variables, and so on.

4.2.3 Random Forest Classifier

Random Forests are among the recent additions to the ensemble methods and machine learning toolbox. Classifier based on Random Forests are ensemble methods such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. See Breiman (1999) and Breiman (2001) for a overview.

Classification trees, as described in Sec. 4.2.1, can work reliable in many cases. However, one of their drawbacks is that such models tend to have high variance. Small changes in the composition of the training set can led to very different tree structures. This drawback just follows from the hierarchical nature of the tree model: small differences in the top few nodes can produce highly different structure as those perburbations are propagated down the tree. To reduce the variance of tree estimates, Random Forest Classifiers (RFC, Breiman 1999, 2001) uses an ensemble of trees – a forest – and attempt to de-correlate the T trees by selecting a random subset $\mathbf{X}_{\text{try}}$ of the input features as candidates for splitting at each node during the tree-building process. The result is that the final model has lower variance than a single tree. For a source to classify, the class probabilities are estimated as the proportion of the T trees that predict each class. Again, as in classification trees, classification output for each new source can then be described either as a vector of class probabilities (giving the probability for each of the C classes) or as its predicted class (with the highest probability).

To give the definition of Breiman (2001):
"A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{X}, \Theta_k)_{k=1,...}\}$

where $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{X}$."

Breiman (1999) (see also Breiman 2001) formalized the concept of a Random Forest with M trees as an estimatior using an ensemble of randomized trees $\{h(\mathbf{X}, \Theta_m), m = 1, ..., M\}$ where the $\{\Theta_m\}$ are independent identically distributed random vectors, and the m-th randomized tree is an estimator $h(\mathbf{X}, \Theta_m)$, where x is a feature vector. The predictions of the T randomized trees are averaged to give the final prediction. (Other possible options are e.g. an average or weighted average of all terminal nodes reached, or, in the case of categorial variables c, a voting majority.) In Random Forest Classifiers, randomized trees are typically built without any pruning (Breiman 1999). The tree building continues until either the terminal node is pure in its classification (i.e.: no further split can be done), or each terminal node contains no more than a pre-defined number of training sample points to split on.

Practical Aspects of Random Forest Classifiers

In the following, the more practical aspects of using a Random Forest Classifier are discussed.

A Random Forest Classifier is trained by executing the steps as described in Algorithm (2).

Algorithm 2: Random Forest Classifier training

Input: $\{(\mathbf{X}, Y_i)\}_{i=1}^{N}$: training set
M: number of trees
k: number of features to split on

Output: $\{h(\mathbf{X}, \Theta_m), m = 1, ..., M\}$: ensemble of randomized trees

begin

for $m = 1, ..., M$ **do**
$\tilde{\mathbf{X}} \subsetneq \mathbf{X}$: sample from $\mathbf{X}$ with replacement, $|\tilde{\mathbf{X}}| > 0.5|\mathbf{X}|$
select k features at random from all features
feature providing the best split, according to some objective function, is used for a binary split on that node
output $\{h(\mathbf{X}, \Theta_m), m = 1, ..., M\}$

Depending on the value of k, there are three different systems:

- Random splitter selection: $k = 1$
- Breiman's bagger: $k =$ total number of predictor variables
- RFC: $k \ll K$ where K is the number of features. Breiman suggests three possible values for k: $1/2\sqrt{K}$, $\sqrt{K}$, $2\sqrt{K}$.

It is important to note that (Breiman 2001): Having a large number of features, the eligible feature set will be quite different from node to node. The greater the inter-tree correlation, the greater the error rate of the Random Forest Classifier. For this reason, the trees must be as uncorrelated as possible. With decreasing m, both the inter-tree correlation and the strength of individual trees are decreasing. For this reason, so some optimal value of m must be discovered.

When applying Random Forest Classifiers, and tree-based classifiers in general, one should not neglect the **importance of hyperparameters**. In contrast to parameters found within the estimators, hyperparameters describe the execution of the algorithm itself. Usually they are fixed before the training process begins.

Hyperparameters of Random Forest Classifiers are (Bernard et al. 2009):

- k, the subset of feature randomly drawn without replacement; this number allows to introduce more or less randomization in the split selection, in such a in such a way that the smaller the value of m, the stronger the randomization,
- M, the number of trees,
- the maximum depth.

In the context of machine learning, hyperparameter optimization is the problem of choosing a set of hyperparameters, usually with the goal of optimizing a measure of the algorithm's performance. Often, hyperparameter tuning is carried out by a grid search, an approach that will methodically build and evaluate a model for each combination of hyperparameters specified in a grid.

When using a Random Forest Classifier, one must be aware of their **strengths and weaknesses**.

Random Forest Classifier are superior to many other methods in terms of accuracy and efficiency, and they are able to deal with unbalanced and missing data. Because the method averages the predictions over multiple trees, the estimated classification probabilities are much more robust to imbalanced training sets than methods using a single tree. They can be parallelized. The feature importance in classification can be easily estimated.
Weaknesses of Random Forest Classifiers are that when used for regression, they are not able to predict beyond the range in the training data. Additionally, they may over-fit data sets that are particularly noisy for small number of trees.

4.2.4 Verification of Classification Results

Despite classifiers like the ones described above are robust, one should not apply them as a "black box", nor use the results without further verification. Also, classification probabilities are not exactly what they might look at first glance.

There exist several concepts on how to test classifiers and verify their results.

Precision and Recall, Purity and Completeness

The classifier predicts whether the input source would be a member of a certain class. The classifier's output is a number $p_{\text{class}} \in [0, 1]$, often called "class probability", but this value should not be used directly in the sense of a probability. Instead a threshold on p_{class} is needed, and the classification quality will then be calculated for sources above this threshold.
To quantify the quality of a classifier (see e.g. Fawcett 2006), consider a two-class problem (binary classification), in which the outcomes are either positive (p), which means being of the desired class, or negative (n). If the classification outcome is p and the actual value also p, this is called a *true positive* (TP). If the outcome is p, but the actual value is n, then it is a *false positive* (FP). Conversely, a *true negative* (TN) means that both the classification outcome and the actual value are n, and a *false negative* means that the classification result is n for an actual value.
Precision P is then defined as

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{4.33}$$

Recall R is defined as

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{4.34}$$

Precision is a measure of result accuracy, while recall is a measure of how many of the truly relevant sources are found.
A classifier with high recall but low precision finds many of the truly relevant sources (positives), but also produces a lot of false positives. A classifier with high precision but low recall is highly accurate, as it rarely produces false positives, but finds only a small fraction of the positives.
High scores for both indicate that the classifier is accurate (high precision) as well as able to find a high fraction of the relevant sources (high recall). Typically, precision and recall are inversely related. As precision increases, recall decreases, and vice versa.

In order to illustrate the performance of a classifier as its threshold on p_{class} is varied, *Precision-Recall curves* (see e.g. Fawcett 2006; Davis and Goadrich 2006) are used. Such curves are created by plotting precision vs. recall while the threshold on p_{class} is varied. An example is shown in Fig. 4.3.

The diagonal from (0,1) to (1,0) in Fig. 4.3 divides the recall-precision space. Points above the diagonal indicate good classification results (better than random), whereas points below the diagonal indicate poor classification results (worse than random).
A classifier with high precision and recall, which would be the ideal case, would return all true positives, while returning no false positives or false negatives, though having all sources classified correctly. Thus, the Precision-Recall curve of the best possible classifier would be represented by a point at (1,1). In realistic classifiers, having recall varying with the threshold on p_{class}, the ideal curve would be precision(recall)=1.
The area under the curve can also be used as a measure for the quality of the classifier, as a high area represents both high recall and high precision.

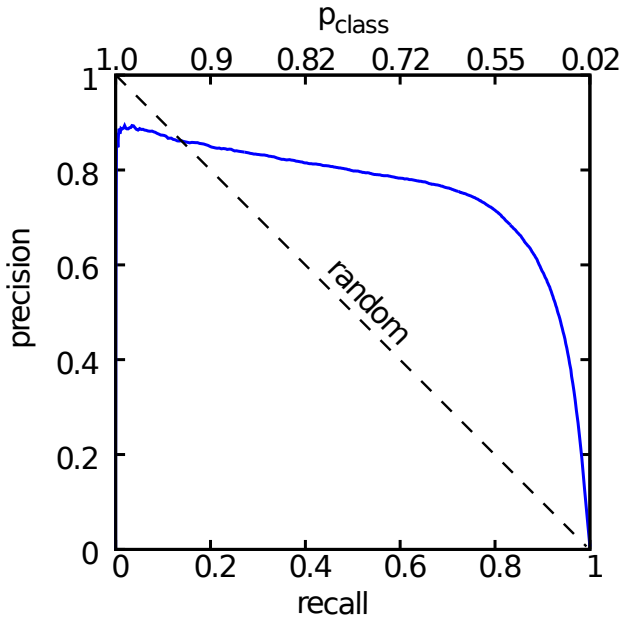


Figure 4.3 Example Precision-Recall curve. The solid line indicates the Precision-Recall curve from some model, the dashed line indicates the outcome of a random process. Points above the diagonal indicate good classification results, whereas points below the diagonal indicate poor classification results.

Thresholds on p_{class} can be used to select samples of a desired purity or completeness. The above defined precision corresponds to the purity of the sample, whereas the recall corresponds to the completeness. As recall is defined as in Equ. (4.34), where TP+FN is independent of the threshold on p_{class}, lowering the threshold may increase the completeness of the sample, as the number of true positives increases.

Cross-Validation

Tests on classifiers should be done under conditions as close as possible to their desired application. Cross-validation does this by training the classifier on all sources of the training set except a small number, or even one ("leave one out"), of held-out sources. As belonging to the training set, the class is also known for these held-out sources, so the classification accuracy can be examined on them. As the held-out fraction is small, its influence on the training set can be selected. This procedure is applied in turn for all sources of the training set.
A typical case is the "10-fold cross-validation" where in turn 10% of the training set's objects are held out. Finally, the results from each cross-validation run are collected for statistical analysis.

Measuring and Analyzing Feature Importance in RFC

An advantage of tree-based classifiers is that they allow to estimate the importance of each feature in the model by construction. As trees are constructed by splitting on one feature at a given time, a feature's importance can be estimated by i.e. counting how often that feature is split, or the decrease in node impurity for splitting on this feature. Another measure used quite

often is referred to as the *variable importance*. This indicates roughly what the decrease in overall classification accuracy would be if a feature were replaced by random permutation of its values, i.e. if this feature would be useless or not exiting.

Analyzing the feature importance is a critical step in building a classification model. This helps in eliminating useless features, and also incorporating additional features in a new version of the model. Also, determining feature importance gives insight into the differences between particular classes, in the cases described here different classes of of variable sources.

Chapter 5

Finding, Characterizing and Classifying Variable Sources in Multi-Epoch Sky Surveys: QSOs and RR Lyrae in PS1 3π Data

The methodology of how to classify variable sources in non-simultaneous multi-color surveys described in the previous Chapter 4, among them the newly developed multi-band structure function fitting, was applied to PS1 3π data in order to identify and classify variable sources.

In this chapter, detailed adaptions of the methodology, containing mainly structure function fitting and machine-learning classification, that was laid out more general to PS1 3π data, as well as results and an outlook for further applications are given.

The details of applying the general approach for classifying variable sources in non-simultaneous, multi-color surveys to PS1 3π are shown. This new approach for quantifying statistical properties of non-simultaneous, multi-color surveys through light curve structure functions turns PS1 3π PV3 effectively into a $\sim$67 epoch survey.

A subsequent machine-learning classifier then assigns probabilities to each source whether it is a QSO, RR Lyrae or, in later work, Cepheid. This approach is used to estimate variability amplitudes and time-scales as well as mean colors and source types for almost all point-sources in the survey.

Using PV3, aside from the Galactic plane, QSO and RR Lyrae samples of purity $\sim$80% and completeness $\sim$80% can be selected. On this basis, a sample of 6.1×10^5 QSO candidates, as well as an unprecedentedly large and deep sample of 4.8×10^4 RR Lyrae candidates spanning distances from $\sim$10 kpc to $\sim$130 kpc was selected for $|b| > 20°$.

Using the RR Lyrae candidate sample, a distance precision of 4% within the Draco dwarf spheroidal can be reached. Additionally, the extent of the Sagittarius stream is visible up to 130 kpc.

The work presented in the following is based on PS1 3π PV3 (see Section 3.1.5).
Before PV3 came out, most parts of the analysis shown here were done for the previous processing version of PS1 3π, PV2, and published (Hernitschek et al. 2016, containing only PV2 results). The work done on PV2 lead to a catalog of all likely variable point sources and QSOs within the

survey, a total of 25.8×10^6 sources, that was already published (Hernitschek et al. 2016).
A comparison between results from PV2 and PV3, as well as details of the methodology differing between both versions is given throughout this Chapter.

The work presented here makes use of two methods developed by Branimir Sesar, namely the outlier detection and cleaning (see Section 5.3.2) and methods on how to enhance the RR Lyre sample purity even more as described in Section 5.6.

The RR Lyrae candidates found by the methods described here are used in the subsequent Chapter 6 to infer the geometry of Sagittarius stream out to 120 kpc, using a combined halo and stream model. The approach described in Section 5.6 was developed by Branimir Sesar with contribution from the author, and will is submitted as (Sesar, Hernitschek et al. 2016).

5.1 Introduction

In the context of time-domain astronomy, the Pan-STARRS1 survey (PS1) 3π (Chambers 2011) offers a unique combination of area, time sampling and depth. PS1 data have been extensively used to find and study transient sources, such as supernovae (Rest et al. 2014) or episodic black hole accretion (Gezari et al. 2012), focusing mostly on the many-epoch coverage in the medium-deep fields. It lends itself also to finding and characterizing sources of less ephemeral variability, and can do so across most of the sky. Such sources of interest are, for example, QSOs and variable stars, such as RR Lyrae.

PS1 3π is a multi-epoch survey that covered three quarters of the sky at typically 72 epochs between mid 2009 and the end of 2014. Yet, in any one of its five bands ($g_{\rm P1}, r_{\rm P1}, i_{\rm P1}, z_{\rm P1}, y_{\rm P1}$), it is only a few-epoch survey, and the observations in different bands are not taken simultaneously. (For further details on PS1 3π, see 3.1.)

Though there are approaches for finding RR Lyrae in PS1 based on their variability properties (e.g. Abbas et al. 2014a,b), there are no readily available approaches to exploit the full information content of the data, e.g. to find, identify, and characterize variable sources generically.

In this Chapter, an approach to characterize variable sources in a survey such as PS1 3π is layed out, developed, tested and finally applied to the full survey. The basic approach should also be very relevant to the Large Synoptic Survey Telescope (LSST)[4], which will also collect non-simultaneous multi-band time-domain data. The methodology encompasses three basic steps: first, identifying sources that clearly vary; second, characterizing their lightcurves with a multi-band structure function; finally, using the identification of variable sources to train the classifier. The last step is carried out using a Random Forest Classifier that takes the classification available for the Sloan Digital Sky Survey (SDSS) Stripe 82 (S82) (Schneider et al. 2007; Schmidt et al. 2010; Sesar et al.

[4]LSST Science Collaborations and LSST Project 2009, LSST Science Book, Version 2.0, arXiv:0912.0201, `http://www.lsst.org/lsst/scibook`

2010) to classify variable sources within PS1 3π. Throughout this analysis, Stripe 82, which was fully observed by the PS1 survey, serves as a testbed for many aspects of the analysis.

In the classification analysis, this work focuses on two classes of astrophysical objects: QSOs and RR Lyrae. These objects have numerous applications. For example, the RR Lyrae can act as tracers of the Milky Way's stellar outskirts (Sesar et al. 2010, 2013a,b) with high distance precision. Variability of QSOs is astrophysically interesting for a variety of reasons (Schmidt et al. 2010; Morganson et al. 2014; Hernitschek et al. 2015), but QSO candidates may also serve as reference sources for calibrating the astrometry of sources near the Galactic plane.

This Chapter is organized as follows. In Section 5.2, a brief description of the data used for the analysis is provided. Beside PS1 3π light curves, this section also describes complementary WISE data that prove important for QSO/RR Lyrae discrimination, as well as the existing QSO and RR Lyrae classification in SDSS S82, which is central for training and validating a Random Forest Classifier. Section 5.3 describes outlier cleaning for PS1 3π, where the approaches differ for PV2 and PV3.

In Section 5.4, the methodology is described that lead from PS1 3π lightcurves to QSO and RR Lyrae candidates. Methods described previously in Chapter 4 are now tailored especially to the needs of PS1 3π data. This section gives also information on how the classification available for SDSS Stripe 82 helps in classifying variable objects in PS1 3π. Results are given in Section 5.5. Here it is demonstrated, relying on Stripe 82 data and faint RR Lyrae in Draco dSph as ground truth, how well the identification and classification of variables with PS1 data works. In particular, the purity and completeness of various QSO and RR Lyrae samples, e.g. at high latitude and around the Galactic anticenter, are quantified and discussed. Finally, in this section, the result on full PS1 3π is given, resulting in a catalog of QSO and RR Lyrae candidates across three quarters of the sky.

Section 5.6 describes further approaches and results in the context of this work, showing how period folding can lead to a cleaner RR Lyrae candidate sample as well as precise distance estimation. Results of this Chapter are discussed in Section 5.7.

Larger figures are given in the Figure section of this Chapter, Section 5.8.

5.2 Data

The approach for calculating variability features and using them to detect and classify variable sources is based on PS1 3π data, supported by time-averaged photometry from the Wide-field Infrared Survey Explorer (WISE) survey. Sources from SDSS S82 as well as sources at Draco dSph (Kinemuchi et al. 2008) serve as ground truth. In this section, the pertinent properties of these surveys are described.

5.2.1 PS1 3π Data

The PS1 survey (Kaiser et al. 2010) is collecting multi-epoch, multi-color observations undertaking a number of surveys, among which the PS1 3π survey (Chambers 2011) is the largest. It has observed the entire sky north of declination $-30°$ in five filter bands ($g_{\rm P1}, r_{\rm P1}, i_{\rm P1}, z_{\rm P1}, y_{\rm P1}$) with average wavelengths of 481, 617, 752, 866, and 962 nm, respectively (Stubbs et al. 2010; Tonry et al. 2012) with a 5σ single-epoch depth of about $g_{\rm P1} < 22.0$, $r_{\rm P1} < 22.0$, $i_{\rm P1} < 21.9$, $z_{\rm P1} < 21.0$, $y_{\rm P1} < 19.8$ magnitudes, respectively.

In contrast to the SDSS filters, the $g_{\rm P1}$ filter extends 20 nm redwards of $g_{\rm SDSS}$, and the $z_{\rm P1}$ filter reaches only to 920 nm. PS1 has no u band. In the near-IR, $y_{\rm P1}$ covers the region from 920 nm to 1030 nm. A more detailed descriptions of PS1 3π is given in Section 3.1.

In the following, single-epoch photometry resulting in light curves from PS1 3π will be used in order to specify variability, as well as to give near-IR and optical colors. A total of 1.1×10^9 sources within PS1 3π were selected for analysis.

All data processing shown in this work is carried out under PS1 catalog processing version PV3. For comparison, attempts on PV2 are partially shown in order to illustrate both the effect of available cadence as well as the power of certain methods even in the low-cadence domain.

5.2.2 WISE Data

Quasars are one of the biggest sources of contamination when selecting RR Lyrae stars, especially at faint magnitudes. They overlap with RR Lyrae in $g - r$ and redder optical colors (e.g. Sesar et al. 2007), and may look as variable as RR Lyrae when observed in spase datasets such as PS1 (see e.g. Fig. 5.3). To better separate QSOs and RR Lyrae stars, PS1 3π data is supplemented with the $W12$ color provided by the all-sky WISE mission.

WISE (Wide-field Infrared Survey Explorer) is a NASA infrared-wavelength astronomical space telescope providing mid-infrared data with far greater sensitivity than any previous survey. It performed an all-sky survey with imaging in four photometric bands over ten months (Wright et al. 2010). Nikutta et al. (2014) have shown that the mid-infrared color $W12 = W1 - W2 > 0.5$ is an excellent criterion to isolate QSOs, because $W12$ is an indicator of the hot dust torus in AGN. To aid in the QSO identification, it is reasonable to find objects with these unusual $W12$ colors. It is necessary to make sure that these colors are not merely a consequence of poor WISE photometry.
Cross-matching between WISE and PS1 3π is done using the nearest source within a 1" radius. If a PS1 3π source does not have a WISE $W1$ or $W2$ measurement, or these measurements have uncertainties ≥ 0.3 mag (i.e. the WISE detection is less than 5σ above the background), the $W12$ color is not used.

For objects with good measurements ($\sigma_{W1} < 0.3$, $\sigma_{W2} < 0.3$), $W12$ is used as feature for classification. Also, if both magnitudes are available, the $i_{\rm P1} - W1$ color is used as feature for classification.

2.6×10^8 out of the 1.1×10^9 selected objects from Sec. 5.2.1 have reliable $W12$ ($\sigma_{W1} < 0.3$, $\sigma_{W2} < 0.3$, where σ_{W1}, σ_{W2} are the errors given on the WISE magnitudes).

5.2.3 SDSS S82 Sources

The Sloan Digital Sky Survey (SDSS, York et al. 2000) is a major multi-filter imaging and spectroscopic survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. The Sloan Legacy Survey covers about 7,500 degrees of the Northern Galactic Cap in optical *ugriz* filters with average wavelengths of 355.1, 468.6, 616.5, 748.1 and 893.1 nm. In typical seeing, it has a 95% completeness down to magnitudes of 22.0, 22.2, 22.2, 21.3, and 20.5, for u, g, r, i, z, respectively. Additionally, the Sloan Legacy Survey contains three stripes in the South Galactic Cap totaling 740 square degrees. The central stripe in the South Galactic Cap, Stripe 82 (S82), was scanned multiple times to enable a deep co-addition of the data and to enable discovery of variable objects.

Most of the observations of the SDSS S82 were obtained primarily for a supernova search, but S82 has provided a wealth of information about transients and variable sources of many kinds.

S82 has $\sim$60 epochs of imaging data in *ugriz*, taken over $\sim$5 years, where extensive spectroscopy provides a reference sample of nearly 10,000 spectroscopically confirmed quasars (Schneider et al. 2007; Schmidt et al. 2010). For S82, there is also a sample of 483 identified RR Lyrae available (Sesar et al. 2010). The classification of QSOs and RR Lyrae in SDSS S82 will be used as a ground truth. This means, they will be used as training set for classification as well as for testing how well the classification method works (see Section 5.4).

5.3 PS1 Object Selection and Outlier Cleaning

Outlier detection and cleaning – the process of removing non-astrophysical photometric outliers from light curves – as well as object selection – excluding some objects from processing – is crucial in order to prepare for reliable determination of variability features. Outliers are prone to cause spurious variability, leading to wrong variability estimates for the underlying source.

The method of outlier cleaning applied during this work differs between PS1 3π PV2 and PV3.
For PV2, a outlier cleaning based on hard detection cuts, mostly motivated by flags, was applied. This method, developed by the author, was also published in Hernitschek et al. (2016).
Later on, for PV3, a machine-learning based outlier cleaning developed by Branimir Sesar was applied.

5.3.1 PV2

For outlier cleaning on PV2 data, a number of cuts on the PS1 data to remove outliers and unreliable data were applied. These cuts fall into two categories: *detection* cuts that remove individual detections, and *object* cuts that remove all detections of a source from the analysis.

Detection Cuts

The most important detection applied to PV2 removes data taken in non-photometric conditions, according to Schlafly et al. (2012), and data from any Orthogonal Transfer Array (OTA) where the detections of bright stars on that chip are on average over 0.02 mag too faint. These cuts remove about 30% of detections.

The second most important detection applied removes observations which land on bad parts of the detector, as indicated by having `psf_qf_perfect` < 0.95. This removes about 10% of detections. Similarly importantly, any observation were excluded where the PSF magnitude is inconsistent with the aperture magnitude by more than 0.1 mag or four times the estimated uncertainties, removing 10% of detections.

Furthermore, any detections with problematic conditions noted by the PS1 pipeline are removed, according to the detections' flags. For the cleaning flags used, see Table 5.1 and also Magnier et al. (2012). This eliminates only about 2% of detections.

Finally, an outlier cleaning based on the z-score of the individual measurements $z_i = (m_i - \mu(b_i))/\sigma_i$ is applied, where m_i is a given magnitude measurement, σ_i is its uncertainty, and $\mu(b_i)$ is the error-weighted mean magnitude of all measurements of that source in its band b_i. This is limited to eliminate at most 10% of the detections of any individual source.

Fig. 5.1 gives the number of PV2 epochs, as well as their cadence, in each band after all of these cuts have been applied. The average number of surviving epochs per source is 35 rather than the total 55 observations shown in Fig. 3.2.

The detection cuts done for PV2 are summarized in Table 5.2. If a detection has one problematic condition, it is likely also affected by other problematic conditions.

Object Cuts

Additionally to individual epochs, all detections of some objects are excluded from consieration. When only a small number of epochs are sampled, tests had shown that structure function estimation becomes unreliable.

To ensure that only objects are considered having enough epochs and high enough signal to noise to be appropriate for variability studies and in particular having enough epochs for multi-band structure function fitting, for PV2 only objects were selected having

(i) $15 < \langle g_{\rm P1} \rangle, \langle r_{\rm P1} \rangle, \langle i_{\rm P1} \rangle < 21.5$, where $\langle \cdot \rangle$ is the error-weighted mean magnitude after applying detection cuts

(ii) at least 10 epochs remaining after after applying detection cuts

Two additional criteria remove extended objects, as well as objects thought to have problematic PS1 detections:

(iii) fewer than 25% of epochs eliminated by `psf_qf_perfect`$\leq$0.95

(iv) fewer than 25% of epochs eliminated by |`ap_mag` - `psf_inst_mag`| $\geq \max(4\sigma, 0.1)$.

Among sources within a magnitude range of 15 to 21.5, these two criteria each remove about 5% of PV2 sources. This was significantly more than expected. However, visual inspection of a selection of affected sources indicates that these cuts were unnecessarily restrictive. These sources could have in fact been included in the analysis without difficulty, but for PV2 this loss was accepted. This loss was called a "selection loss" in Hernitschek et al. (2016), and it means that all samples (QSOs, RR Lyrae, and variable objects in general) will be missing 10% of the objects. For PV2, more than 3.88×10^8 objects across three quarters of the sky survived the cuts, and were therefore processed in order to analyze the variability of them.

Table 5.1. Bit-flags used to exclude bad or low-quality detections in PV2

FLAG NAME	Hex Value	Description
PM_SOURCE_MODE_FAIL	0x00000008	Fit (non-linear) failed (non-converge, off-edge, run to zero)
PM_SOURCE_MODE_POOR	0x00000010	Fit succeeds, but low-SN or high-Chisq
PM_SOURCE_MODE_SATSTAR	0x00000080	Source model peak is above saturation
PM_SOURCE_MODE_BLEND	0x00000100	Source is a blend with other sources
PM_SOURCE_MODE_BADPSF	0x00000400	Failed to get good estimate of object's PSF
PM_SOURCE_MODE_DEFECT	0x00000800	Source is thought to be a defect
PM_SOURCE_MODE_SATURATED	0x00001000	Source is thought to be saturated pixels (bleed trail)
PM_SOURCE_MODE_CR_LIMIT	0x00002000	Source has `crNsigma` above limit
PM_SOURCE_MODE_MOMENTS_FAILURE	0x00008000	could not measure the moments
PM_SOURCE_MODE_SKY_FAILURE	0x00010000	could not measure the local sky
PM_SOURCE_MODE_SKYVAR_FAILURE	0x00020000	could not measure the local sky variance
PM_SOURCE_MODE_BIG_RADIUS	0x00100000	poor moments for small radius, try large radius
PM_SOURCE_MODE_SIZE_SKIPPED	0x10000000	size could not be determined
PM_SOURCE_MODE_ON_SPIKE	0x20000000	peak lands on diffraction spike
PM_SOURCE_MODE_ON_GHOST	0x40000000	peak lands on ghost or glint
PM_SOURCE_MODE_OFF_CHIP	0x80000000	peak lands off edge of chip

Table 5.2. Cuts used to exclude bad detections in PV2

Condition	Fraction of detections removed
Photometric conditions	0.29
$\lvert$`ap_mag - psf_inst_mag`$\rvert < \max(4\times\sigma_m, 0.1)$	0.10
`psf_qf_perfect` > 0.95	0.11
Pipeline flags (Tab. 5.1)	0.017
$\lvert z_i - z_{\rm median}\rvert < 5\sigma$	0.02

5.3.2 PV3

Machine learning can be a valuable tool in outlier detection to get precise, but not too restrictive cuts. When processing light curves from PV3, instead of the detection cuts described above, a machine-learning based outlier cleaning developed by Branimir Sesar was applied. In the following, this method is described, together with changed object cuts being more sensible in the case of PV3.

Detection Cuts by Machine Learning

This method uses not bit-flags or other hard cuts to exclude detections that are possible photometric outliers in PS1 3π light curves, but a machine-learning algorithm that more efficiently identifies bad photometric data.

A non-astrophysical outlier (thus, caused not by variability of the physical source but by effects related to the observational conditions such as instrumentation and atmosphere) is defined as a photometric measurement deviating by more than 2.5σ from its "expected" value. The expected value is calculated via a model, and σ gives the total photometrc uncertainty of the detection in case.

In order to identify and subsequent remove such outliers, a machine-learning model was developed that predicts whether a detection will be a photometric outlier or not. For doing so, properties associated with a detection, e.g. position on the chip, bandpass, level of agreement with a PSF model, seeing are investigated by the model being trained on a set of non-varying sources (bright K and G stars).

Validation tests by Branimir Sesar have shown that the machine-learned outlier model is able to identify 80% of all 2.5σ outliers, while misclassifying only 1 good observation for every found true 2.5σ outlier. For comparison, the outlier cleaning applied to PV2 as described in Section 5.3.1 also identifies almost all of the 2.5σ outliers, but has the problem of misclassifying 8 good observations as outliers for every found true 2.5σ outlier.
Another advantage of the new method is that feature importance can be used for understanding what causes outliers.

After removing photometric outliers from PV3 light curves using this machine-learned outlier model, the average number of observations per source is 67 (out of the initial 72).

The method described here is part of Sesar, Hernitschek et al. (2016) (submitted).

Object Cuts

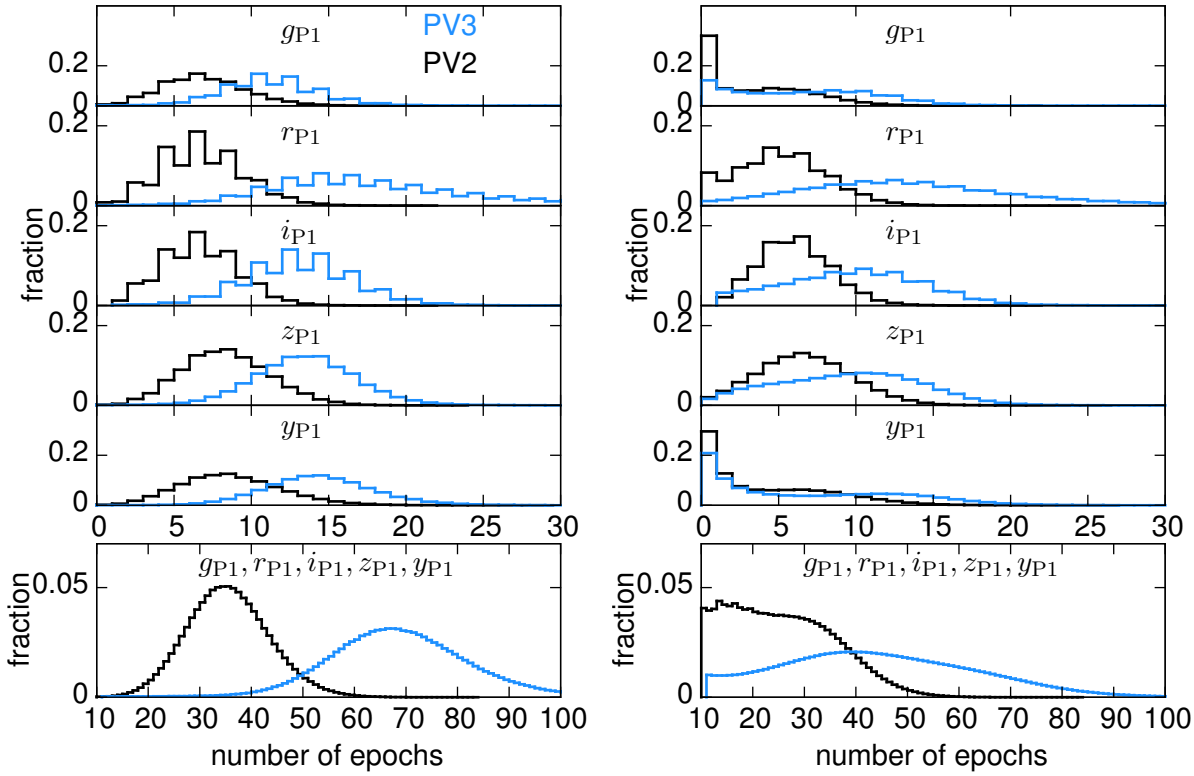
Because of the higher number of epochs in PV3 at general, together with the improved outlier cleaning by machine learning, the object cuts have changed from PV2 to PV3. To ensure reliable

structure function estimation, the requirement on the total number of epochs remains, namely at minimum 10 epochs over all bands.

No additional cuts are applied at this stage, to not restrict the number of sources more than mandatory. Possibly magnitude cuts are applied after the light curves are processed, to select samples tailored to the specific application and evolution.

5.3.3 Comparison PV2 vs. PV3

Fig. 5.1 depicts the total number of bright ($15 < i_{\rm P1} < 18$) and faint ($18 < i_{\rm P1} < 21.5$) PS1 3π PV2 and PV3 epochs after outlier cleaning. In contrast, the total number of epochs before outlier cleaning is given in Fig. 3.2.



(a) bright limit: $15 < i_{\rm P1} < 18$

(b) faint limit: $18 < i_{\rm P1} < 21.5$

Figure 5.1 The typical number of observations for (a) bright and (b) faint sources in PV2 and PV3 after source and detection outlier cleaning. For bright sources, the average number of epochs after outlier cleaning in PV3 is 67, in contrast to only 35 in PV2. For faint sources, the average number of epochs in PV3 is 40 in contrast to only 30 in PV2. This is an effect of having more epochs per source in PV3 than in PV2, as shown in Fig. 3.2, but also an effect of the more sensible outlier cleaning provided for PV3.
A minimum number of 10 epochs after cleaning was enforced for further processing.

The work at hand is carried out with the current internal data release of the PV1 3π survey, PV3, for final results, and uses PV2 for pre-analysis and during the methodology was designed. Results

from PV2 are published in Hernitschek et al. (2016) and are given for comparison as far as it is reasonable. Results on PV2 can show how much is possible with the methodology developed within this work on even more sparse light curves.

5.4 Methodology

In this section, the three steps are described that are taken to identify and characterize variable point sources: first, determine whether sources are variable; second, characterize their variability with a structure function; and third, attribute classifications. Classification is carried out using a Random Forest Classifier that utilizes a training set from SDSS S82 and Draco dSph. Throughout the following steps, all data are conform to the selection requirements described in Section 5.2. Fig. 5.2 illustrates the logical flow of the methodology that is detailed in the following subsections.

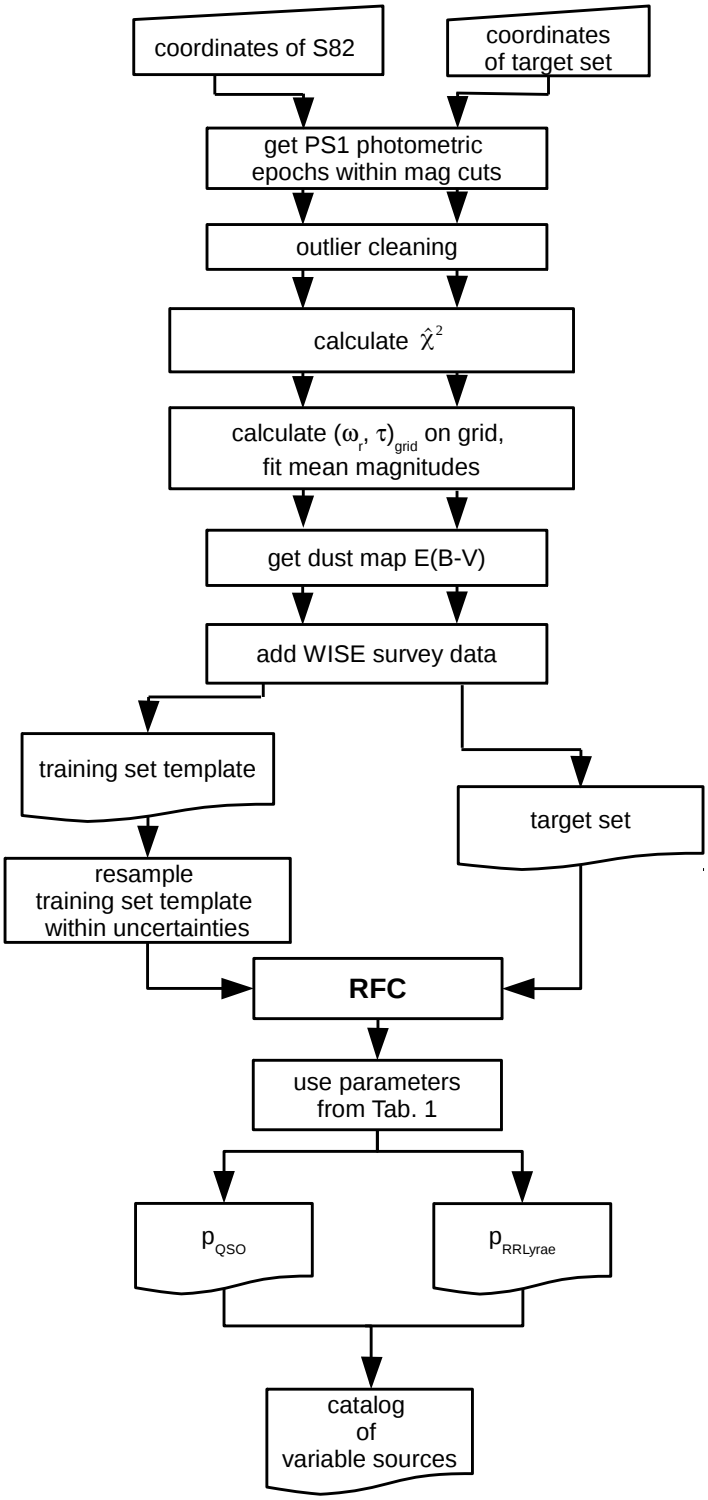


Figure 5.2 Logic flowchart for finding and classifying variable sources as set out in Section 5.4.

5.4.1 Identifying Significantly Varying Sources

The description of the methodology starts by laying out a very generic and non-parametric measure for variability, simply to characterize the significance of variability by a scalar quantity. Specifically, it is defined as

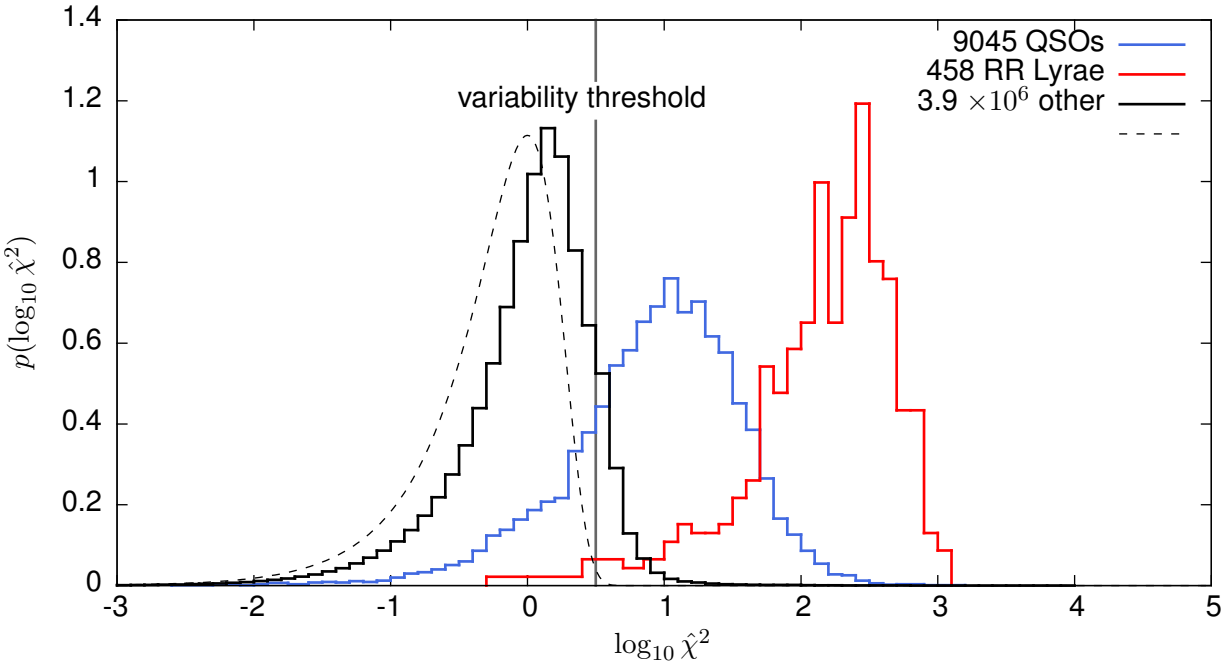
$$\hat{\chi}^2 = \frac{\chi^2_{\text{source}} - N_{d.o.f}}{\sqrt{2\,N_{d.o.f}}}, \tag{5.1}$$

with

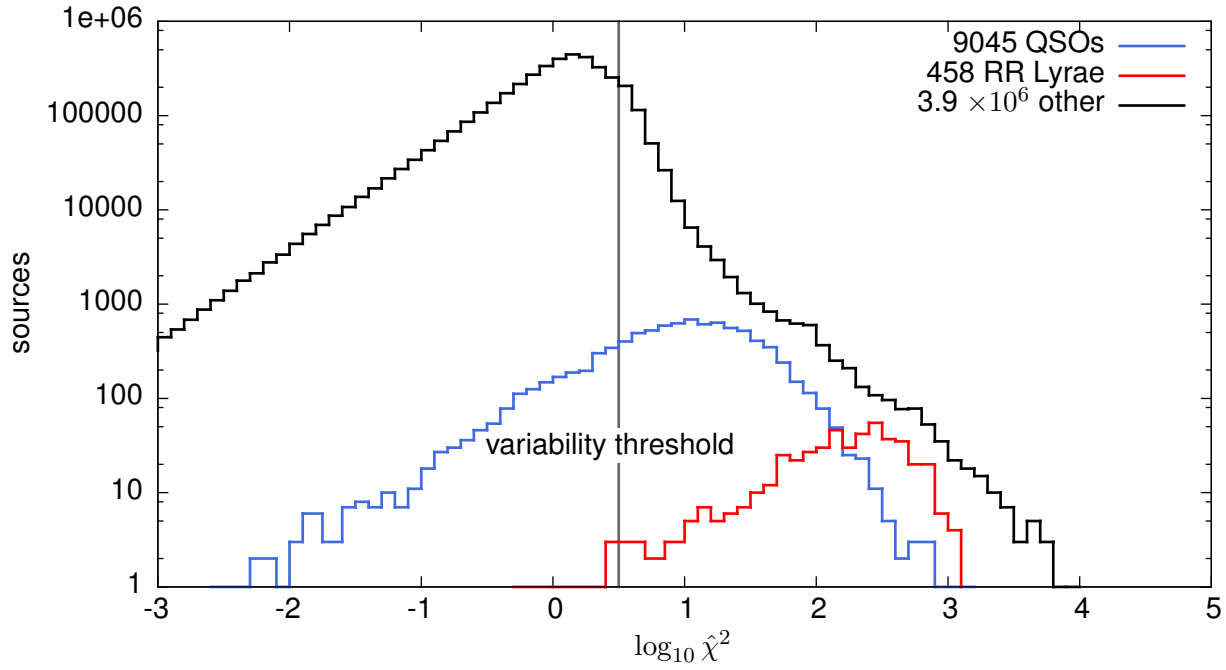
$$\chi^2_{\text{source}} = \sum_{\lambda} \sum_{i=1}^{N} \frac{(m_{\lambda,i} - \langle m_\lambda \rangle)^2}{\sigma^2_{\lambda,i}} \tag{5.2}$$

where N is the total number of photometric points for one object across all n bands for the source (for PS1 3π, $n \leq 5$), the sum over λ is over the available bands (for PS1 3 π, $\subseteq \{g_{\text{P1}}, r_{\text{P1}}, i_{\text{P1}}, z_{\text{P1}}, y_{\text{P1}}\}$), and $N_{d.o.f} = N - n$ is the number of degrees of freedom.
Assuming that most of the sources are not variable, the distribution of $\hat{\chi}^2$ is expected to be a unit Gaussian. In contrast, varying sources should form a "tail" of higher $\hat{\chi}^2$. Figure 5.3 shows the normalized distribution of $\hat{\chi}^2$, derived from the PS1 photometry of all selected objects in S82, with known QSOs (blue) and known RR Lyrae (red) shown in separate (normalized) distributions. The "other" objects have a $\hat{\chi}^2$-distribution close to that expected for non-varying sources (dashed line), confirming that most sources in the sky are non-varying (within a level of less than a few percent) and that the PS1 3π photometry is reliable. The QSOs and RR Lyrae appear well separated in the *normalized* distibutions. However, there are only 458 RR Lyrae and 9045 QSO, compared to $\sim 3.9 \times 10^6$ "other" objects in SDSS S82 cross-matched to PS1 3π and surviving the cuts of Sec. 5.3.2. Fig. 5.3 (b) shows how the distribution of "other" sources superimposes the distribution of QSOs and RR Lyrae due to the high number of "other" sources.

Therefore, a simple criterion such as $\hat{\chi}^2$ is insufficient to identify QSOs or RR Lyrae. In the subsequent analysis, all objects are used, as $\hat{\chi}^2$ serves only as a feature for the classifier and no cut is placed on $\hat{\chi}^2$. However, for RR Lyrae only, one could in principle restrict oneself to objects with $\hat{\chi}^2 > 10$ without losing completeness.



(a)



(b)

Figure 5.3 Histograms for $\hat{\chi}^2$ of the training set's sources after outlier cleaning; PS1 photometry in S82 region, type from SDSS.
(a) Normalized histogram, overplotted: theoretical expectation from unit Gaussian distribution ($\mu = 0, \sigma = 1$). The differences between the black histogram and the dashed line arises from a combination of noise-model imperfections and actual variability of objects. The cutoff for the variability criterion ($\log_{10} \hat{\chi}^2 > 0.5$ for the Catalog of Variable Sources, see Sec. 5.5.4) is given as a grey line.
(b) Full histogram showing how the distribution of "other" sources superimposes the distribution of QSOs and RR Lyrae due to the high number of "other" sources.

5.4.2 Application of Multi-Band Structure Function Fitting to PS1 Data

The general technique for determining structure functions for multi-band, non-simultaneous data – a novel method developed as part of this work – is already described in Section 4.1.4. The key ingredient is a description of the ratios of the variabilities in the different bands, which is characterized by a power law with exponent α (Equation (4.8)), which leads to a band-specific variability amplitude. The other elements of the structure function analysis – variability, time scale, and linear nuisance parameters (mean magnitudes) – are the same as in the single band case. For the case of the PS1 3π data at hand, it turns out that α is poorly constrained for individual objects, making it preferable to derive an external estimate of α from other data (SDSS S82), and fix it for the subsequent PS1 3π analysis. Assuming Equ. (4.8), data from SDSS S82 were used to derive characteristic values for α, leading to $\alpha \approx -0.65$ for QSOs and $\alpha \approx -1.3$ for RR Lyrae, both with an uncertainty of 0.01, which is in good agreement with Sesar (2012). Experiments of fitting PS1 3π data with both choices of α resulted in similarly good fits. Accordingly, a single fixed $\alpha = -0.65$ was chosen throughout this analysis. This choice of α corresponds to variability amplitude ratios $\omega(b)/\omega(r) = 1.175, 1.00, 0.88, 0.80, 0.75$, where b represents the PS1 bands $g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$, $z_{\rm P1}$, and $y_{\rm P1}$.
With α fixed, the fit to each source is described by the time scale τ, an overall variability scaled to the r band, ω_r, and the mean magnitudes $\vec{\mu}$.

Fig. 5.4 shows four example fits to the PS1 photometry of objects in SDSS S82: one QSO, one RR Lyrae, one "other" variable object and one seemingly non-varying object. For each object, the light curve is shown as observed in the five bands (top panel), and the combined light curve after shifting each band by the estimated $\mu(b)$; the structure function parameters ω_r and τ are listed for each case. Note that the QSO in Fig. 5.4 (a) has τ of over a year, while the RR Lyrae in panel (b) has a τ of about a day. The Figure also shows the interpolated light curves, given the observations and the structure function parameters, according to the technique of Rybicki and Press (1992).

One could sensibly derive the pdf's for the parameters ω_r, τ and $\vec{\mu}$ via MCMC (see A.2 in the Appendix); however, it proved computationally less expensive by a factor of $\gtrsim$100 to calculate $p(m|\omega_r, \tau)$ based on a reasonable parameter grid. The linear optimization of the $\vec{\mu}$ was computed for each grid-point on a log-spaced grid of $-2 < \log \omega_r/\mathrm{mag} < 0.5$, $0.04 < \tau/\mathrm{day} < 5000$ with 20 values evenly spaced in $\log \omega_r$ and 30 in $\log \tau$ to find the best-fit structure function parameters on the grid $\omega_{r,\rm grid}$ and $\tau_{\rm grid}$. For a small subsample, it was verified that this approximation agrees well with full MCMC runs. Fig. 5.5 shows the gridded log-likelihood estimates for the same four sources as in Fig. 5.4. The panels show the 68% CI of the $\log \mathcal{L}$ distribution and the maximum likelihood values of the parameters.

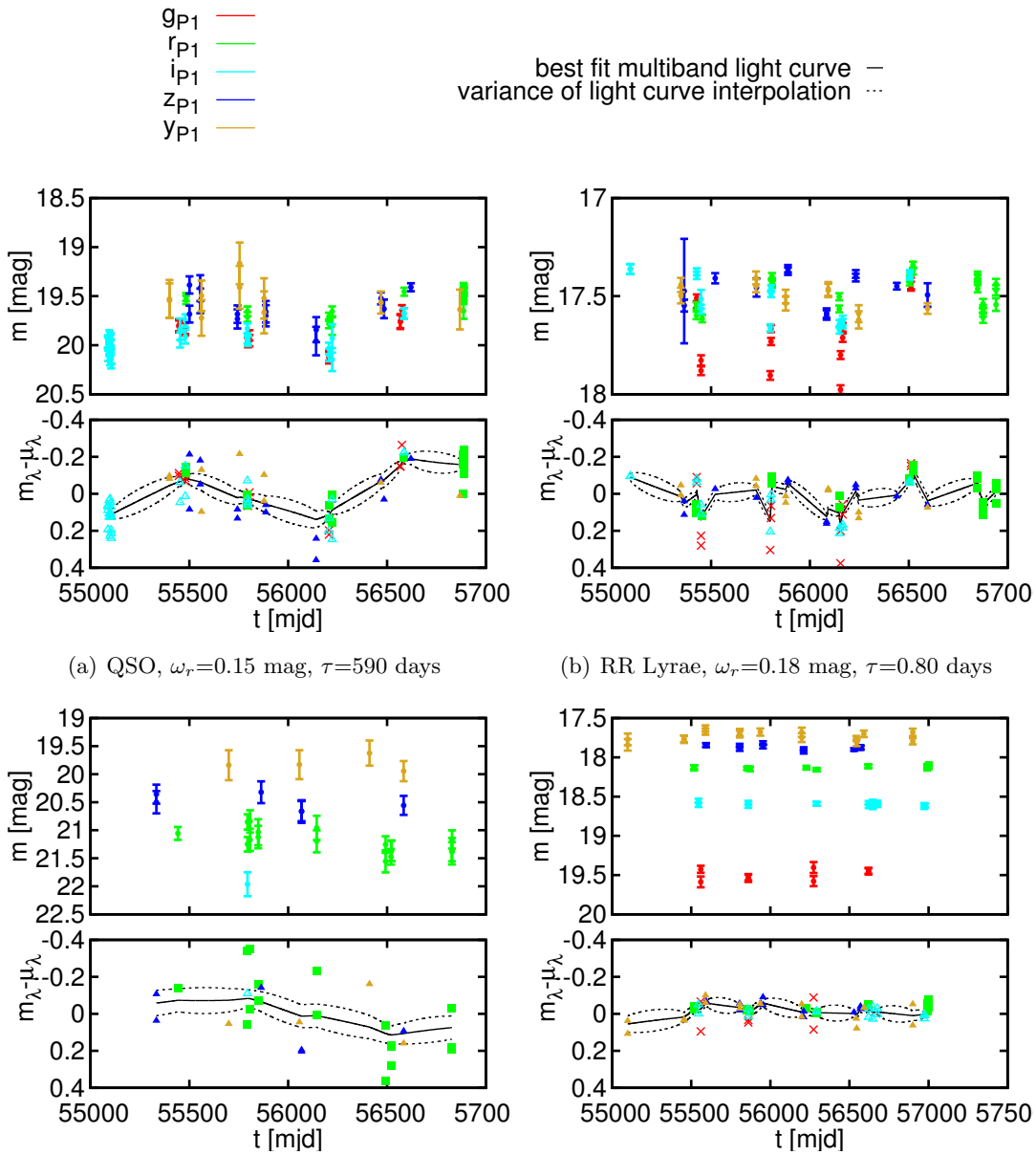
Fig. 5.6 shows the distribution of variability parameters ω_r and τ, for all PS1 objects in the SDSS S82 area that survive the magnitude cut and which have significant variability, either satisfying $\hat{\chi}^2 > 5$ or $\hat{\chi}^2 > 30$ for objects within the stellar locus. This Figure illustrates a number of points: first, and unrelated to variability, it shows the power of the WISE color $W1 - W2$ to separate

QSOs from other sources (Nikutta et al. 2014). Second, it shows that RR Lyrae and QSOs indeed populate different areas of (ω_r, τ) space. While they can only be roughly differentiated by their amplitudes ω_r, they have dramatically different time scales τ: RR Lyrae have typical $\tau \sim 1$ day and QSOs have $\tau \sim 100 - 1000$ days.

Additionally, a power law model for the structure function was tested. It provided worse separation between QSO, RR Lyrae and other objects in the structure function parameter plane. This can be explained by the cadence of the survey, as the definition of the power law makes the structure-function fitting more sensitive to the TTI pairs.

Figures 5.3 and 5.6 show that the light curve parameters will be very helpful in classifying variable sources. Yet, these figures also show that simple cuts on some parameters will not be optimal for differentiating object classes. A more sophisticated machine-learning method is needed here.

The distribution of the variability timescale τ is shown in detail in Fig. 5.7. The distribution is calculated from the PS1 3π photometry of a subsample of 2380 QSO, 362 RR Lyrae and 5196 "other" objects (black) surviving a magnitude cut of $r_{\mathrm{P1}} < 21.5$ mag and having $\hat{\chi}^2 > 5$, $\hat{\chi}^2 > 30$ in the stellar locus of S82. The values were estimated by using a MCMC. Whereas the distribution for the "other" sources shows white noise, RR Lyrae and QSO show distinct distributions. This makes τ a sensible feature for distinguishing between QSO, RR Lyrae and "other", presumably not significantly variable sources.



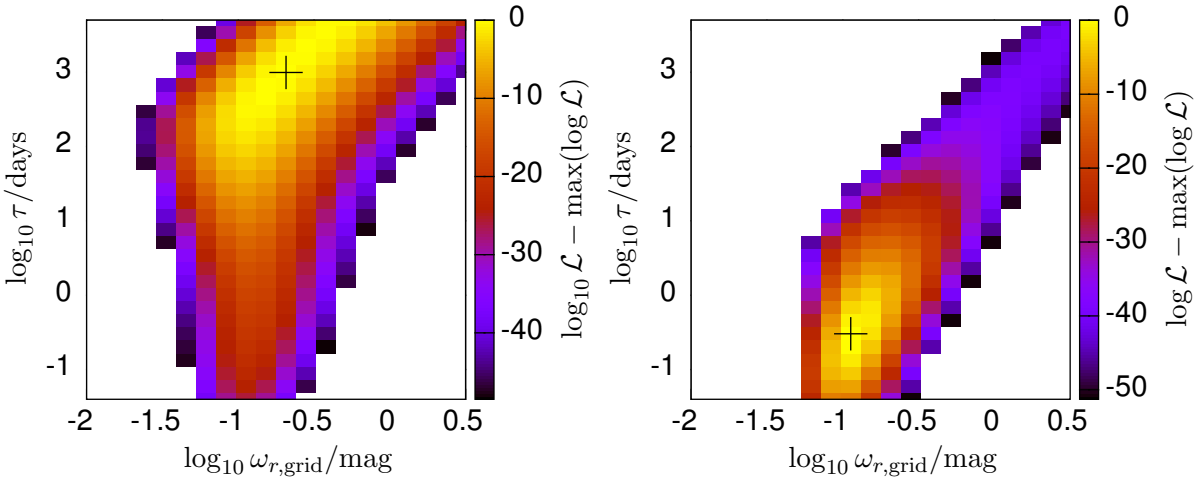
(a) QSO, ω_r=0.15 mag, τ=590 days

(b) RR Lyrae, ω_r=0.18 mag, τ=0.80 days

(c) other (variable?), ω_r=0.14 mag, τ=490 days

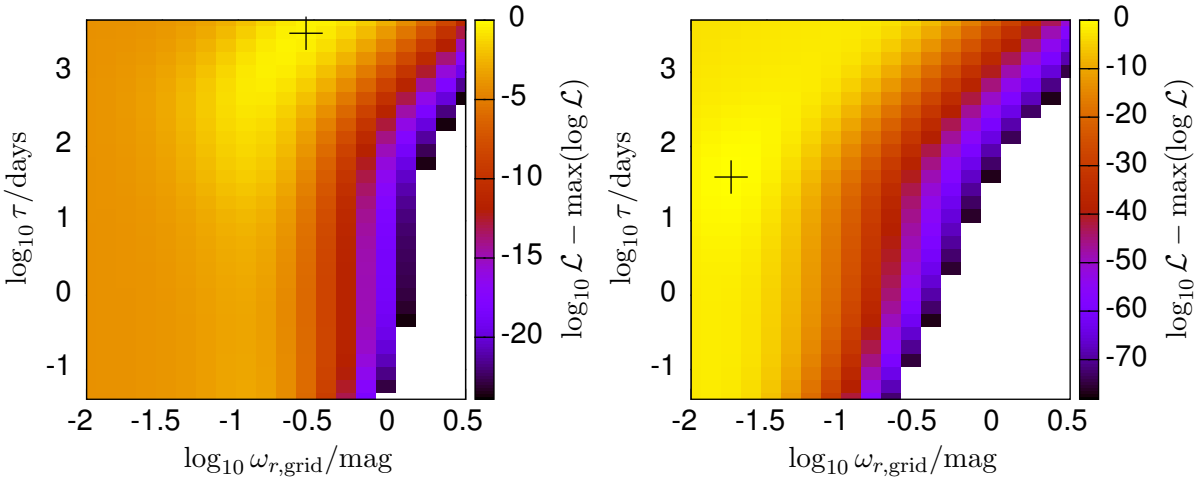
(d) other (nonvariable?), ω_r=0.021 mag, τ=54 days

Figure 5.4 Examples of multi-band lightcurve models for different types of sources. In each figure, the upper panel gives the PS1 lightcurve data points with error bars after outlier cleaning. The lower panel shows the lightcurve fit by a multi-band DRW structure function. The solid lines represent the best fit mean model lightcurve Equ. (4.20). The area between the dotted lines represents the variance Equ. (4.21) for the r band. For ω_r and τ, the best MCMC point-estimates of the parameters for each source are used.



(a) QSO, $\omega_{r,\mathrm{grid}}$=0.21 mag, τ_{grid}=990 days

(b) RR Lyrae, $\omega_{r,\mathrm{grid}}$=0.11 mag, τ_{grid}=0.30 days



(c) other (variable?), $\omega_{r,\mathrm{grid}}$=0.28 mag,τ_{grid}=3300 days

(d) other (nonvariable?), $\omega_{r,\mathrm{grid}}$=0.018 mag, τ_{grid}=39 days

Figure 5.5 Gridded log-likelihood estimates for the structure function parameters. The figures show the 68% CI of the of $\log \mathcal{L}$ evaluated on the log-spaced grid for the sources shown in Fig. 5.4. The maximum is marked with a cross, and the values of τ_{grid} and $\omega_{r,\mathrm{grid}}$ corresponding to the cross are given in the caption.

5.4.3 Classification of PS1 3π Sources Using a Random Forest Classifier

For classifying objects based on variability measures and mean magnitudes calculated during structure function fitting, as well as other features, a Random Forest Classifier (RFC) is used

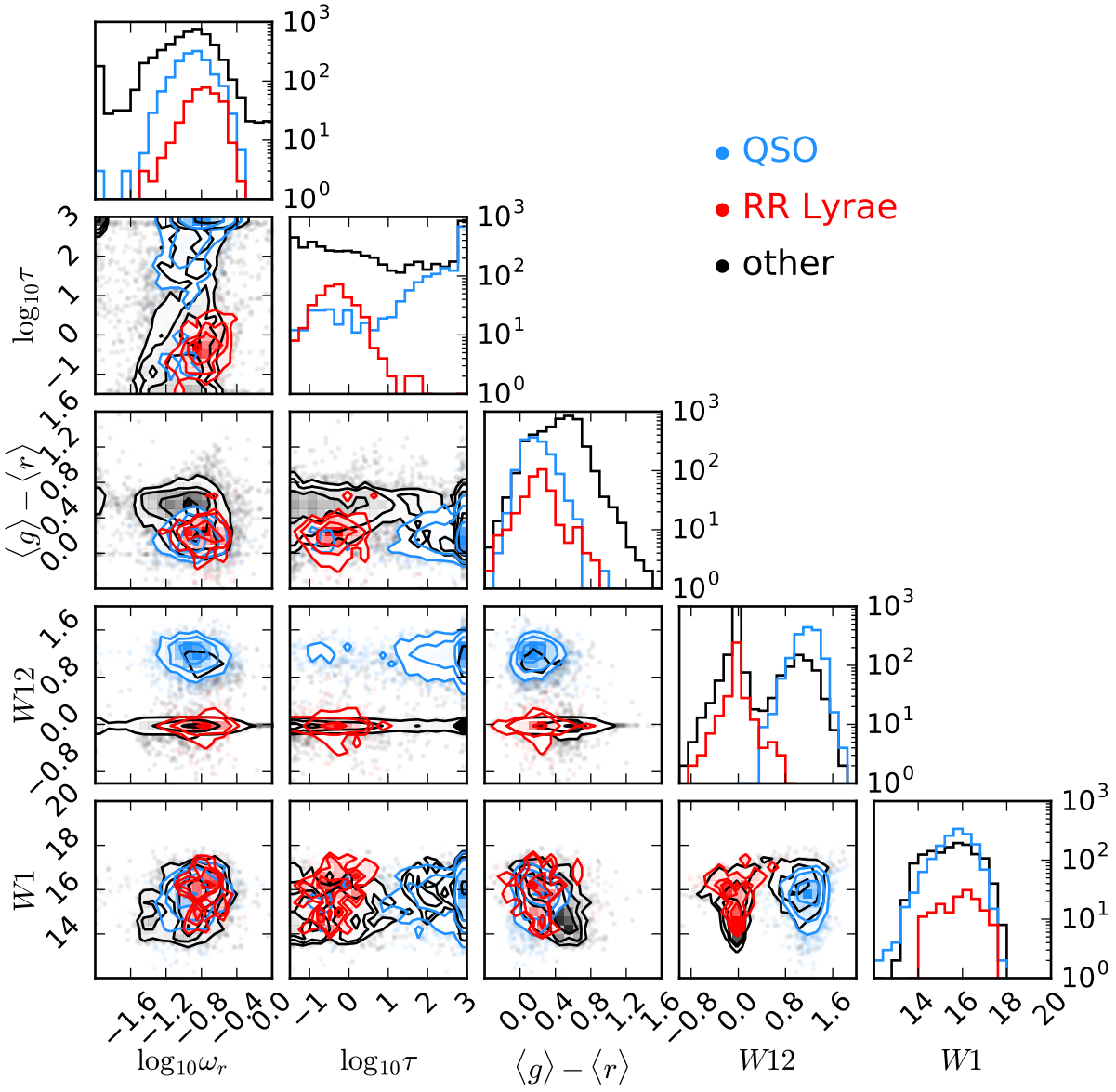


Figure 5.6 An extract of the feature space showing several features (for complete feature list, see Table 5.7) by source class.
Structure function parameters and colors calculated from PS1 3π photometry for a subsample of 2380 QSO (blue), 362 RR Lyrae (red) and 5196 "other" objects (black) surviving magnitude cut $r_{\rm P1} < 21.5$ and having $\hat{\chi}^2 > 5$, $\hat{\chi}^2 > 30$ in the stellar locus of S82. Note that for this Figure the structure function parameter (ω_r, τ) are obtained using a MCMC, as the discrete griding of ω_r and τ proved visually distracting. For ω_r and τ, the best MCMC point-estimates of the parameters for each source are used. The $W12$ color (bottom row) illustrates how powerful WISE data are in separating QSOs from other sources (Nikutta et al. 2014). It is presumed that most "other" sources with $W12 > 0.5$ are indeed QSOs missed by the SDSS classifiation.

(see Section 4.2.3). This machine-learning classifier is implemented in Python's `scikit_learn` package, an open-source Python package for machine learning (Pedregosa et al. 2011)[5]. Using a training set, it gives the probability of a target set's object being of a certain class (the *class*

[5]http://scikit-learn.org

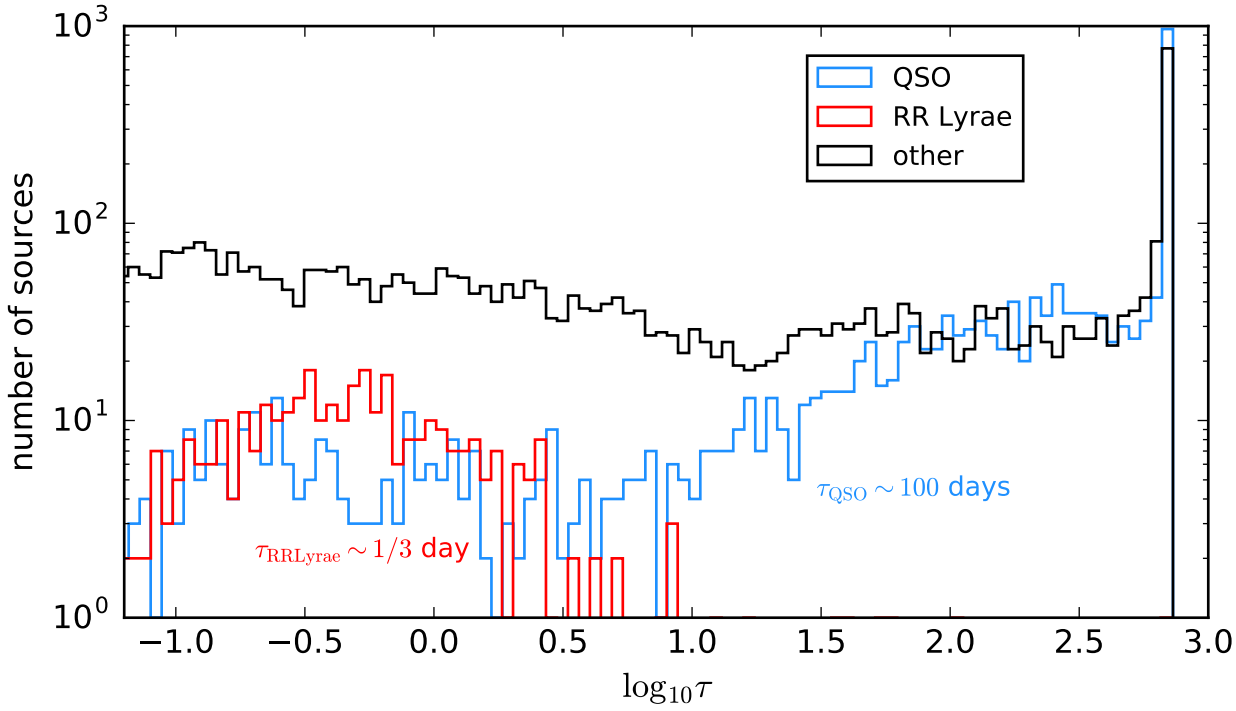


Figure 5.7 The variability timescale τ, calculated from PS1 3π photometry for a subsample of 2380 QSO (blue), 362 RR Lyrae (red) and 5196 "other" objects (black) surviving magnitude cut $r_{\mathrm{P1}} < 21.5$ and having $\hat{\chi}^2 > 5$, $\hat{\chi}^2 > 30$ in the stellar locus of S82. The values were estimated via MCMC. Whereas the distribution for the "other" sources is white noise, RR Lyrae and QSO show distinct distributions.

probability), p_{QSO} and p_{RRLyrae}. The class probability should not be used as a probability, but instead, purity and completeness should be calculated from the obtained the sample later on.

For using a RFC, a training set is needed, with observed object parameter values as well as classification labels. In the following, this training set – that will also serve as a ground truth in validation – is described.

The Training Set

The training set is built by cross-matching PS1 3π sources in dedicated regions to reliable classified sources from catalogs. If the position of a source in one of these three catalogs matches the position of the closest PS1 3π sources within 1", the PS1 3π source is labeled according to the catalog. The remaining PS1 3π sources in these regions are labeled as "other", being neither QSO nor RR Lyrae and being mostly non-variable.

The largest part of the training set consists of more than 3.9×10^6 PS1 3π sources located in the SDSS S82 region ($310° > \alpha < 59°$, $|\delta| < 1.25°$) that meet the conditions described in Section 5.3.2 and are at least $24'$ away from the center of globular cluster NGC 7089 ($24'$ is two times the tidal radius, see Harris (1996)). To label the objects in the training set, they are matched to the

Sesar et al. (2010) catalog of RR Lyrae as well as the spectroscopic QSOs (Schneider et al. 2007; Schmidt et al. 2010).

Within S82, 9045 QSO and 461 RR Lyrae are cross-matched to PS1 3pi. Out of the RR Lyrae, 458 are outside of NGC 7089. Additionally, more than 3.9×10^6 "other" objects are selected from S82.

Additionally, for RR Lyrae, sources within Draco dSph are used (Kinemuchi et al. 2008). In SDSS S82, the majority of RR Lyrae stars are located within 30 kpc of the Sun (83% of the sample, (see Sesar et al. 2010) and thus are bright ($r_{\rm P1} < 18.5$). To enhance the training set with fainter, and thus more distant RR Lyrae stars, 269 RR Lyrae in the Draco dwarf spheroidal galaxy are used, located at a heliocentric distance of $\sim$80 kpc (Kinemuchi et al. 2008).

Since these catalogs are based on observations that are slightly deeper and more numerous than PS1 3π, the catalogs are considered to be 100% pure and complete up to the adopted PS1 magnitude limit, and likely beyond. Consequently, the labels of the sources within the training set are considered as the "ground truth" when measuring the efficiency of our selection method (i.e., the selection completeness and purity). In the following, all numbers of purity and completeness are given with respect to the cross-matched sources, i.e. 9073 QSOs and 727 RR Lyrae (458 in S82, 269 in Draco dSph) in the training set.

The Feature Set

As a RFC cannot deal with measurement uncertainties by default, this issue is addressed by extending the training set by copies of itself, sampled within the assumed errors of the PS1 and WISE data. For each object in the training set, 5 samples are taken in addition to the original one.

Furthermore, the training set is extended to account for uncertainties in reddening. Correction for foregfround reddening is done by interpolating the extinction at the position in case using the Schlafly et al. (2014) dust map with the extinction coefficients of Schlafly and Finkbeiner (2011).

The training set is then extended by presenting additional QSOs, RR Lyrae, and other objects to the classifier, having artificially introduced a small dereddening error. This is done in the following way:

(i) make $E(B-V)_{\rm sample}$ drawn from Gaussian $G(E(B-V)_{\rm catalog}, \delta E(B-V) = 0.1E(B-V)_{\rm catalog})$ at the position of the training set source
(ii) 5% chance that $E(B-V)_{\rm sample} = 0$, irrespective of catalog entry
(iii) sample new mean magnitudes in bands $g_{\rm P1}, r_{\rm P1}, i_{\rm P1}, z_{\rm P1}, y_{\rm P1}$ for PS1, and $W1$, $W2$ from WISE within their errors
(iv) deredden them by $E(B-V)_{\rm sample}$
(v) brighten magnitudes so that $r_{\rm PS1}$ after dereddening by $E(B-V)_{\rm sample}$ is the same as after dereddening by $E(B-V)_{\rm catalog}$.

Features are derived from PS1 3π as well as WISE photometry, giving variability as well as color features. Dereddened optical colors are useful as they provide a rough estimate of the spectral type, and could help with identification of RR Lyre stars (which are A-F type stars). Thus, observed PS1 3π magnitudes are corrected for extinction using the extinction coefficient of Schlafly and Finkbeiner (2011) and the dust map by Schlafly et al. (2014) and calculate $(g-r)_{\rm P1}$, $(r-i)_{\rm P1}$, $(i-z)_{\rm P1}$, $(z-y)_{\rm P1}$ colors from the mean magnitudes calculated by structure function fitting.

The Table 5.7 summarizes the feature set being used for the RFC.

Though the mean r band magnitude is helpful in detecting RR Lyrae in general, it is not here as it introduces a too strong bias in distance, as the training set covers only the range $14.5 \lesssim r_{\rm P1} \lesssim 21.5$ and the aim is to identify candidates out to 22 mag. Among the colors, the dereddened $(i-z)_{\rm P1}$ is a helpful gravity indicator that helps to reduce contamination (Vickers et al. 2012).

When using a RFC, missing values have to be replaced by some dummy values ("imputation") in the training and target sets. A common solution is replacing missing values by the mean of the available ones. This can be done not only for missing values, but also for values considered as unreliable. As imputation of the median is impractical for the way the data are processed, an imputation of -9999.99 is used instead and tested to behave comparably without effects on the results. If for some reason an object is not observed in a particular PS1 3π band, the value of the color involving that band is reset to -9999.99. Accordingly, imputation of -9999.99 is also used in cases where $\sigma_{W1} > 0.3$, $\sigma_{W2} > 0.3$, or when magnitude errors are not available.

With the stellar locus defined as

$$\texttt{stellar\_locus} = \begin{cases} 1, & [(\langle r\rangle_{\rm P1} - \langle i\rangle_{\rm P1}) < 1.4(\langle g\rangle_{\rm P1} - \langle r\rangle_{\rm P1} + 0.05)] \,\& \\ & [(\langle r\rangle_{\rm P1} - \langle i\rangle_{\rm P1}) > 1.4(\langle g\rangle_{\rm P1} - \langle r\rangle_{\rm P1} - 0.05)] \\ 0 & \text{else.} \end{cases} \tag{5.3}$$

Table 5.3. Feature set for the Random Forest Classifier

feature	description
$\omega_{r,\mathrm{grid}}$,τ_{grid}	best fit structure function parameter on log-spaced grid
$\hat{\chi}^2$	normalized χ^2 statistic, see Equ. (5.1)
$(g-r)_{\mathrm{P1}}$, $(r-i)_{\mathrm{P1}}$, $(i-z)_{\mathrm{P1}}$, $(z-y)_{\mathrm{P1}}$	colors from dereddened PS1 mean magnitudes
`stellar_locus`	see Equ. (5.3)
$W12$	$W1-W2$, helps with QSO identification
$i_{\mathrm{P1}}-W1$	separates RR Lyrae from QSO

To find the optimal hyperparameters (see Section 4.2.3), the `GridSearchCV` function available in the `scikit_learn` package was used. `GridSearchCV` selects the test values of hyperparameters from a grid, and then measures the performance of the classification model (for the given hyperparameters) using a 10-fold stratified cross-validation (for cross-validation, see Section 4.2.4). In detail, the training set is split into 10 subsets using stratified splitting (i.e. making sure that the ratio of RR Lyrae and non-RR Lyrae sources, or QSO and non-QSO sources is equal in both sets). The model is then trained on 9 subsets, and the class probability is obtained from the model trained in this way for the tenth subset. The performance on the classification is then evaluated on this tenth set, and the whole procedure is repeated nine more times, each with a different held-out set. The average of 10 performance evaluations is stored, and the set of hyperparameters with the best average performance is finally used for training the classifier.

In order to rank the features by their importance (*feature importance*, see Section 4.2.3), the built-in functionality of `scikit_learn` was used. Given that all features are available, the feature importance order is as follows for QSOs:

$\omega_{r,\mathrm{grid}}$, τ_{grid}, $\hat{\chi}^2$, `stellar_locus`, $(g-r)_{\mathrm{P1}}$, $(r-i)_{\mathrm{P1}}$, $(i-z)_{\mathrm{P1}}$, $(z-y)_{\mathrm{P1}}$, $W12$, $i_{\mathrm{P1}}-W1$

and for RR Lyrae:

$\hat{\chi}^2$, τ_{grid}, $\omega_{r,\mathrm{grid}}$, $(i-z)_{\mathrm{P1}}$,$(g-r)_{\mathrm{P1}}$, $(r-i)_{\mathrm{P1}}$, $i_{\mathrm{P1}}-W1$, $(z-y)_{\mathrm{P1}}$, `stellar_locus`, $W12$.

This ranking can be understand when taking a look at the obtained purity-completeness curves later on which clearly indicate that variability is crucial (see Sec. 5.4.4). Also, when taking a look at Fig. 5.3, it is noticeable that $\hat{\chi}^2$, separates better separates RR Lyrae out of the full sample, than QSOs.

5.4.4 Verification of the Method Using SDSS S82 Classification Information

In order to test the efficacy of the selection and classification method, detailed testing was carried out on the training set and especially on the S82 area, using PS1 3π lightcurves, with the training set's labels from S82 (Schneider et al. 2007; Sesar et al. 2010) and Draco dSph (Kinemuchi et al.

2008) as the "ground truth" to quantify purity and completeness of the classifications. Purity and completeness are always quantified with respect to a threshold of the class probability p_{RRLyrae} or p_{QSO}, respectively. Given a threshold on the score and knowing the true class of each source in the training set, one can measure the fraction of recovered RR Lyrae (or QSO), the completeness, as well as the fraction of true RR Lyrae (or QSO) in the obtained sample, the purity.

For any one of the two categories, say RR Lyrae, one can define a candidate sample $\mathcal{S}$ by the choice of a minimum p_{RRLyrae}. On the basis of the S82 and Draco dSph ground truth, the completeness and the purity of this sample can be defined (see Section 4.2.4). Here, *purity* is defined as the fraction of all RR Lyrae stars in $\mathcal{S}$, and the "completeness" is the fraction of actual RR Lyrae stars contained in $\mathcal{S}$. In both instances, one would expect completeness to be monotonic and purity to be nearly monotonic in p_{RRLyrae}. For the QSOs and any other class, analogous definitions apply. Depending on context, a sample $\mathcal{S}$ is described either by a cut on $p_{\mathrm{RRLyrae/QSO}}$, or by the corresponding purity and completeness of this sample as determined on the training set.

In order to give estimates on purity and completeness, again a 10-fold stratified cross-validation is used. The model is trained on 9 subsets with a balanced ratio of sources from each class, and purity and completeness when applying this model to the tenth subset are calculated. The whole procedure is repeated 9 more times, each with a different held-out set. The average of 10 evaluations is finally used as purity and completeness. The spread of purity and completeness based on the chosen training set can be estimated from the spread of the purity and completeness obtained from the 10 individual runs.

For all purity-completeness plots in the following, a step size in p_{RRLyrae}, p_{QSO} of 0.001 was chosen. Tables of purity and completeness with a stepsize of 0.01 are given in the Table Appendix, Section B.1.

Fig. 5.8 shows purity-completeness curves (see Section 4.2.4) for the trade-off between purity and completeness with respect to the total cross-matched sources. These values are calculated for all sources in the training set, irrespectively of brightness. The purity-completeness curves are given for using not only the full feature set, but various subsamples of the features for classification, namely PS1 3π variability and color, PS1 3π variability only, PS1 3π and WISE color only. For comparison, the case of using all features in the training set is given for PV2 as dashed line.

The left column refers to QSO classification, the right one to RR Lyrae classification. This Figure shows that, as expected, for small completeness the purity is maximal, while the completeness is maximized with severe expense to the purity. What compromise needs to be made between completeness and purity in sample selection depends in detail on the science question, but the top panels of Fig. 5.8 suggests that the purity increases only little at the expense of completeness less than 80%. This may be a sensible threshold for an inclusive sample, whenever PS1 lightcurves and mean colors, as well as WISE colors are available. At the top of the horizontal axis, the relation between completeness and p_{RRLyrae}, p_{QSO} is indicated.

The different lines in the upper panels of Fig. 5.8 illustrate the relative importance of the different pieces of data that may enter the classification; the classification as not only carried out with the full feature set from Table 5.7, but also tested the cases where only color-related or variability-related information was used. Also the calculated feature importance, as given above, highlights the rigorous importance of the variability features.

Fig. 5.8 shows that the variability information is absolutely indispensable to define a sample with a sensible combination of purity and completeness for RR Lyrae as well as for QSO. These different purity-completeness curves also indicate what one might expect for purity and completeness, when a particular source lacks some information used as feature, for instance, a detection in WISE or particular PS1 3π bands.

Given that the training set is finite in size, the purity and completeness will depend in detail on the chosen training sample. The individual lines in the lower panels of Fig. 5.8 reflect different samplings of the training set. For a training set of the size available in S82, the effect is noticeable, but small.

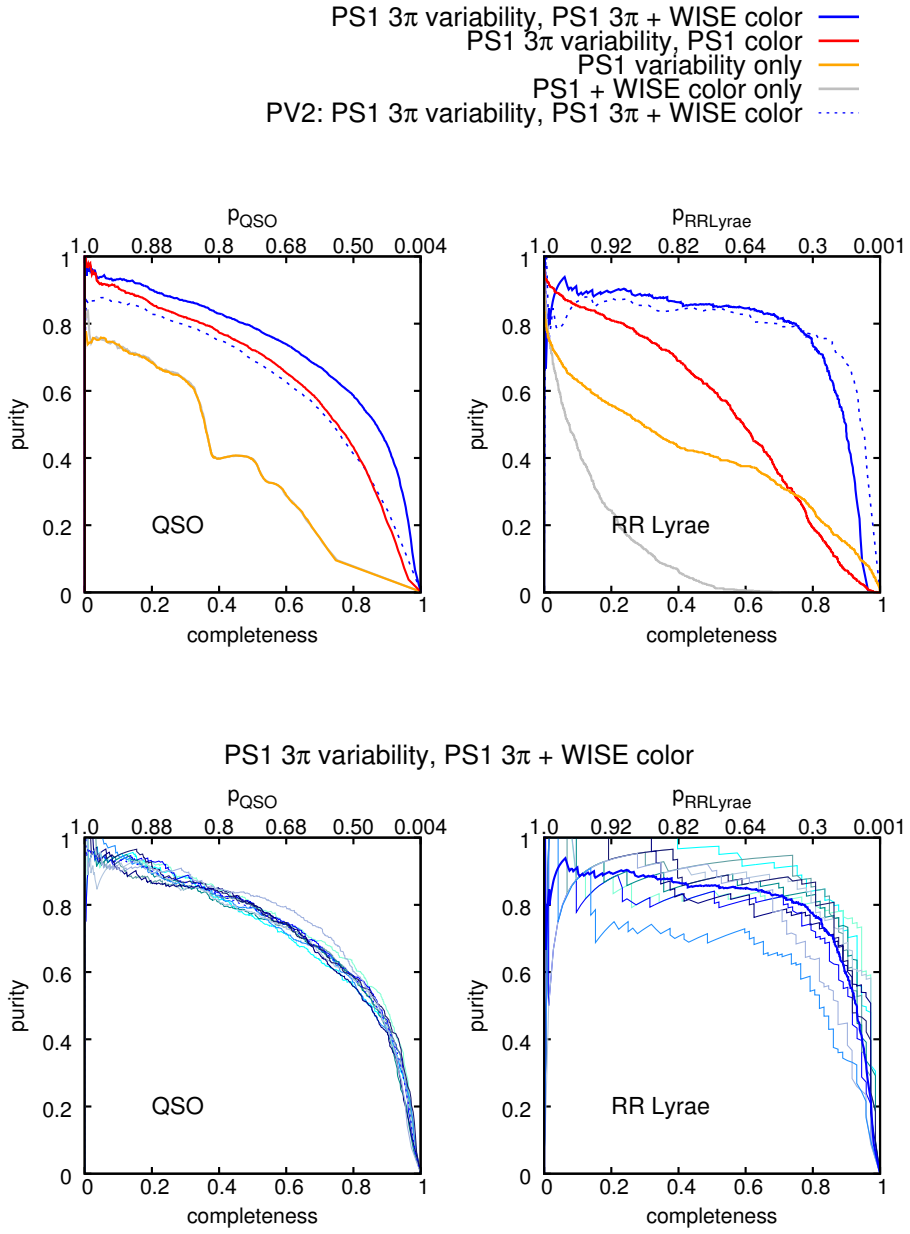


Figure 5.8 Trade-off between purity and completeness with respect to total cross-matched sources for different pieces of information provided to the RFC. The upper panels show purity-completeness curves when PS1 variability and PS1 + WISE colors, PS1 variability and colors only, PS1 variability only, PS1 + WISE colors are provided. There is a limited purity and completeness that can be achieved with variability only (yellow line). As expected, using all features (blue line) gives the best result quantified by purity and completeness.
The lower panel gives the impact of the training-set stochasticisty, illustrated by the dependence of purity and completeness on the chosen training set sources (presuming PS1 variability and PS1 + WISE colors are provided). The trade-off between purity and completeness is plotted from using 10 different randomly selected training sets, as well as their mean (thick dark blue line, the same as the blue line in the upper panel). At the top of the horizontal axis the relation between completeness, and p_{RRLyrae}, p_{QSO} is given. For RR Lyrae, with only 458 S82 RR Lyrae and 269 Draco dSph RR Lyrae in the training set, the stochasticisty is noticeable. In contrast, for QSO with 9045 sources in the training set, it is negligible.
Tables of purity and completeness for the case of using all features (blue line) can be found in the Table Appendix, Table B.1 for QSOs and B.3 for RR Lyrae, respectively.

Dependence of Purity and Completeness on Source Brightness

The result shown in Fig. 5.8 is integrated over a range of distances (roughly $14.5 < r_{\rm P1} < 22$, or $\sim 5-120$ kpc for RR Lyrae). Since it is reasonable to expect variations in purity and completeness as a function of distance (or magnitude), a more detailed analysis is needed. It is likely that the classification becomes more uncertain as sources get fainter and light curves become sparser and noisier.

To specify the heliocentric distance dependence for RR Lyrae classification, the training set's ground truth used for verification was split up into sources with ∼40 kpc ($14.5 < r_{\rm P1} < 18.5$) and ∼80 kpc ($19.7 < r_{\rm P1} < 20.7$). The obtained completeness and purity was then compared to the one from the full sample reaching $\sim 14.5 < r_{\rm P1} < 22$. The $r_{\rm P1} = 18.5$ mag brightness cut was used because the vast majority of halo RR Lyrae stars are located within that magnitude range (Sesar et al. 2010).

The resulting purity and completeness is given in Fig. 5.9 as well as in the Tables B.3 to B.5. At a heliocentric distance of ∼40 kpc, for a completeness of 0.8, a purity of 0.86 can be reached, using a $p_{\rm RRLyrae}$ threshold of 0.27. At ∼80 kpc, the same threshold results into a completeness of 0.8, purity of 0.8. For a threshold of 0.06, for sources at ∼40 kpc, the sample completeness is 0.98, the purity 0.62, and for sources at ∼80 kpc, the sample completeness is 0.88, the purity 0.52.

As being interested in distant sources, for further analysis the following thresholds are used:
Sources of the sample that can be selected using $p_{\rm RRLyrae} > 0.27$ are referred to as "likely RR Lyrae", whereas those selected using $p_{\rm RRLyrae} \geq 0.06$ are referred to as "possibly RR Lyrae".

To specify the magnitude dependence of QSO classification, a subsample of the training set's QSOs were used, selected by $14.5 < r_{\rm P1} < 20$. The obtained completeness and purity was then compared to the one obtained from the full sample. As for RR Lyrae, also for QSO a higher purity at the same level of completeness can be reached for less faint sources. The resulting purity and completeness is given in Fig. 5.10, as well as in the Tables B.1 and B.2.
Using again a threshold resulting into a completeness of 0.8, a purity of 0.8 can be reached using $p_{\rm QSO} \geq 0.56$. Using $p_{\rm QSO} \geq 0.31$ instead, this results into a sample purity of 0.75, completeness of 0.88.

For further analysis, the following thresholds are used:
Sources of the sample that can be selected using $p_{\rm QSO} \geq 0.56$ are referred to as "likely QSO", whereas those selected using $p_{\rm QSO} > 0.31$ are referred to as "possibly QSO".

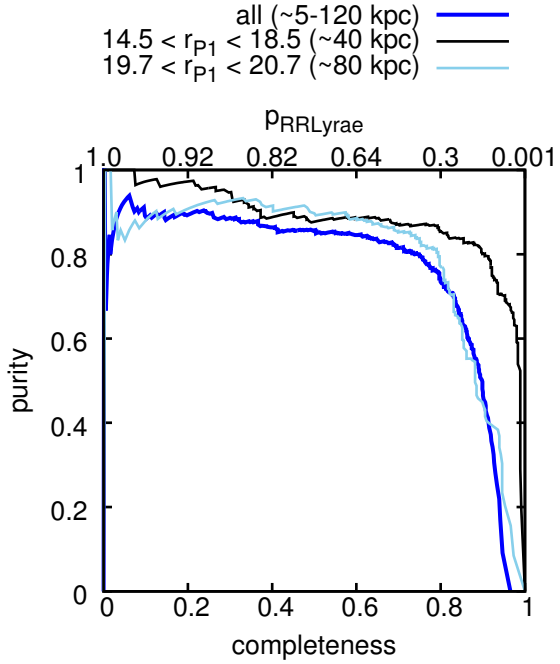


Figure 5.9 Trade-off between RR Lyrae purity and completeness with respect to total cross-matched sources for different brightness limits. The equivalent heliocentric distance for each range of appparent magnitude is indicated. At the top of the horizontal axis the relation between completeness and $p_{\rm RRLyrae}$ is given. At the bright end, $14.5 < r_{\rm P1} < 18.5$, a significantly higher purity at the same completeness can be reached than for fainter sources. The purity-completeness curve integrated over the full magnitude range is indicates as thick line. Tables of purity and completeness for the different brightness limits can be found in the Table Appendix, Tables B.3 to B.5.

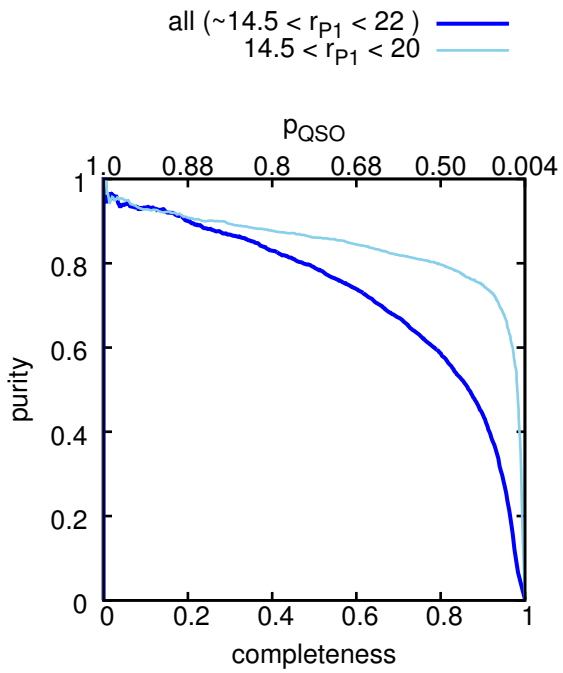


Figure 5.10 Trade-off between RR Lyrae purity and completeness with respect to total cross-matched sources for different brightness limits. At the top of the horizontal axis the relation between completeness and p_{RRLyrae} is given. As for RR Lyrae, also for QSO a higher purity at the same level of completeness can be reached for brighter sources. The purity-completeness curve integrated over the full magnitude range is indicates as thick line. Tables of purity and completeness for the different brightness limits can be found in the Table Appendix, Tables **??** to B.2.

5.4.5 Limitations of the Method

The method of automatic source classification is subject to several limitations. The most important of these are:

(i) mismatch between assumptions on the ground-truth training set and other regions of sky,
(ii) incompleteness of the training set, and
(iii) the inhomogeneity of the available data over the sky.

These limitations are addressed in the following.

The classifier is trained using data on SDSS S82, supplemented by Draco dSph, where existing large catalogs of RR Lyrae and QSOs provide an almost complete sample of objects in these regions. After training the classifier, it is applied to other regions, where no similar classifications already exist. In general, however, the application of the classifier to regions other than S82 and Draco dSph is only justified when the region has distributions of RR Lyrae, QSOs, and potential contaminants similar to that in S82. Over most of the high latitude sky, this is the case, and so the method can be applied without difficulty.

However, at low latitudes the number of contaminants is relatively much larger than the number of RR Lyrae and QSOs in S82 and Draco dSph, since in these regions the data include very large numbers of metal rich disk stars. Additionally, the data itself is qualitatively different: the presence of reddening influences the observed colors of sources, and variation in reddening as a function of distance means that even with a perfect 2D reddening map, dereddened colors may no longer match the true colors of objects. Accordingly, at low latitudes one should not expect such a classifier to perform with the same purity and completeness as at high latitudes, and our S82-based estimates of purity and completeness will no longer apply.

The second problem with the technique is that even in high latitude regions, the adopted training set is imperfect. This is especially the case for the adopted QSO training set. The method uses spectroscopically selected QSOs from Schmidt et al. (2010), which are complete only down to roughly an $i_{\rm P1}$-band magnitude limit of 21.25. Therefore, in the training set, fainter objects are marked as non-QSOs, so the classifier learns to discard these objects – even when they are, in fact, QSOs, as indicated by their WISE $W1-W2$ color and variability. This results in a quasar sample whose purity and completeness is really only relative to S82 spectroscopic quasars, rather than the underlying population of QSOs falling in our magnitude range.

A final concern with the method is that the ability to determine if an object is in fact a QSO or RR Lyrae depends on what information is available for it. The purities and completenesses computed are properties of the entire sample of selected objects. The assignment to classes of individual objects within that sample may be relatively uncertain, if, for instance, those objects lack specific PS1 colors or detections in WISE. Figure 5.8 serves to show what may happen to the purity and completeness of subsamples of objects, for which only limited information is available.

5.5 Results

The method of variability characterization and subsequent Random Forest classification was then applied to all sources in PS1 3π, with the selection criteria discussed in Section 5.3, resulting in a total of more than 1.1×10^9 classified sources. Fig. 5.21 shows the all-sky projection of PV3 source density within the cuts from Sec. 5.3. Here, the results of this classification are presented and discussed. Throughout, the discussion focuses on two illustrative regimes of Galactic latitude, the North Galactic cap and the Galactic anticenter region. Specifically these regions are selected:

- $0<l<360$, $60<b<90$ (around the Galactic north pole), about 2800 deg^2, about 3.1×10^7 classified sources, source density 1.1×10^4/deg^2
- $165<l<195$, $-15<b<15$ (around Galactic anticentre), about 900 deg^2, about 3.9×10^7 classified sources, source density 4.4×10^4/deg^2.

As the analysis considers RR Lyrae, but also QSOs, at low Galactic latitudes, a number of effects in the candidate selection are likely to become important: first dust extinction at low latitudes will push faint sources below the detection limit; imperfect dereddening may lead to differing

de-reddened colors; and the training set, S82, is mostly at high Galactic latitudes with low dust, leaving the classifier imperfectly prepared for very high level of Galactic disk star contaminants.

The coverage of the obtained sources is illustrated by selecting a sample of 1.5×10^5 highly probable RR Lyrae ($p_{\rm RRLyrae} > 0.27$ expected purity of 0.8 and completeness of 0.8 at 80 kpc) and plotting their angular distribution as Mollweide projection in Fig. 5.23. An analogous figure was made for QSO ($p_{\rm QSO} \geq 0.56$, expected purity of 0.8 and completeness of 0.8), as shown in Fig. 5.22 containing 3.7×10^5 sources.

From these plots – discussed in greater detail later on – a highly structured distribution of RR Lyrae candidates and a homogeneous distribution of QSO candidates is visible.

Other large area maps of QSO candidates are shown in Fig. 5.11 for the North Galactic cap, in Fig. 5.12 for the Galactic anticenter, and in Fig. 5.22 for the entire PS1 3π region. The analogous maps for these three areas, but shown in RR Lyrae candidates are shown on Figures 5.14, 5.15 and 5.23.

For both QSOs and RR Lyrae stars the obtained samples of candidates constitute by far the largest sets of high-quality candidates, both in terms of imaging depth, sky area and consequently sample size, e.g. compared to Morganson et al. (2014), who found a QSO purity of 48% and completeness of 67% for PS1-SDSS data.

In the following, all "purity" and "completeness" given for a threshold on $p_{\rm QSO}$, $p_{\rm RRLyrae}$ refer to the case having the full feature set from Table 5.7 available and referring to $19.7 < r_{\rm P1} < 20.7$ ($\sim$80 kpc heliocentric distance), or $14.5 < r_{\rm P1} < 20$ for QSO.

5.5.1 QSO Candidates

QSOs should be distributed isotropically across the sky, with a mean number density of candidates, of about 20 objects per deg^2 in the magnitude range $15 <$ mag < 21.5 (Hartwick and Schade 1990; Schneider et al. 2007; Schmidt et al. 2010). This allows for testing the large scale homogeneity of our classification in areas of high Galactic latitude, and it allows to look at the changing completeness and purity towards low latitudes. As contaminants are expected to increase at low latitudes, many more candidates with low $p_{\rm QSO}$ are expected towards the Galactic plane. Until dust extinction and disk star contamination become severe, an approximately uniform density of objects with high $p_{\rm QSO}$ is feasible.

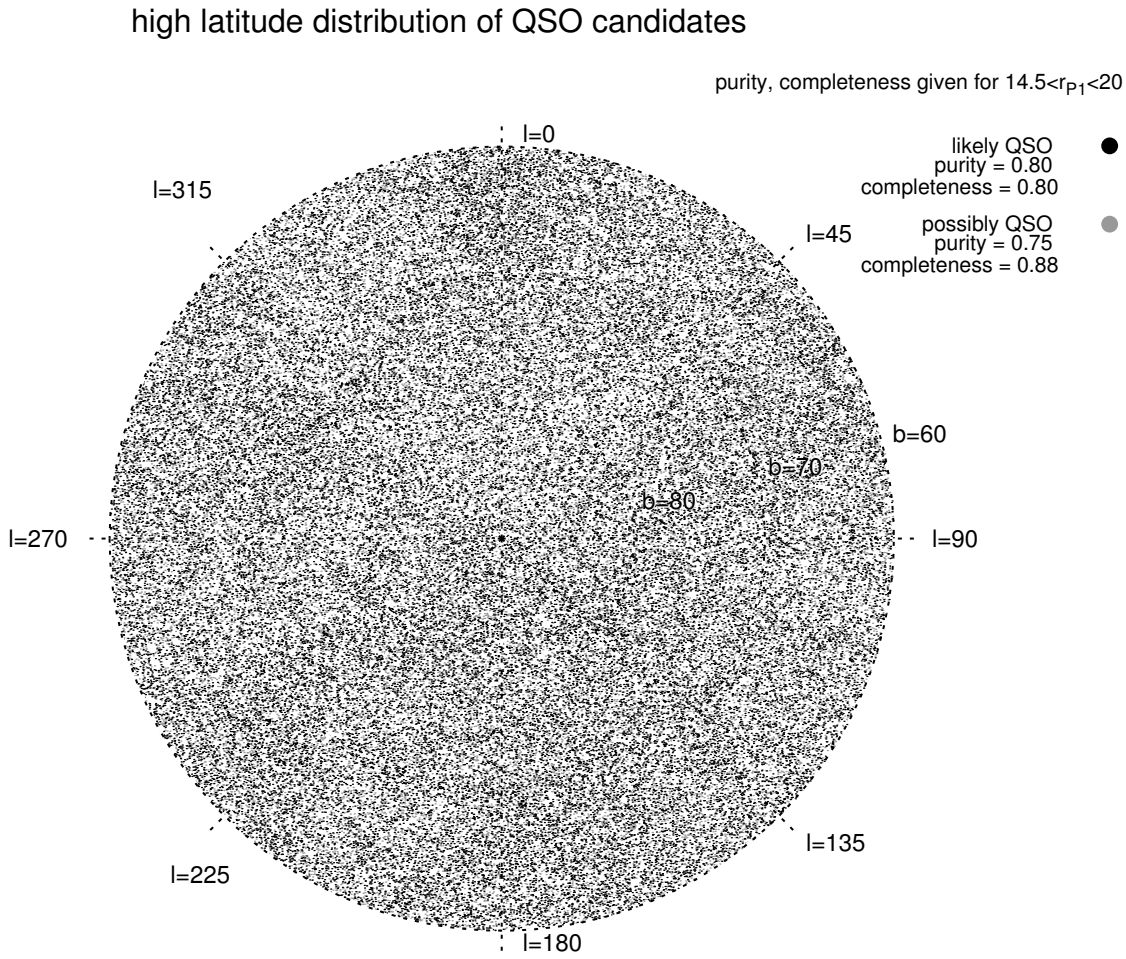
Some of these expectations are borne out in the candidate selection near the Galactic North pole: as shown in Figure 5.11 the selection of candidates with $p_{\rm QSO} > 0.56$, accounting for a purity in S82 of 0.8 and a completeness of 0.8, is uniform to a high degree.

In regions away from the Galactic plane, a homogeneous distribution of the QSO candidates is found. Homogeneity is obvious in Fig. 5.11 as well as in Fig. 5.12 down to $|b| \sim 10^\circ$. For $p_{\rm QSO} \geq 0.56$, the source density is $\sim$20 sources per deg^2, in good agreement with the assumption.

As depicted in Fig. 5.13, the number of sources per deg^2 at a given minimum $p_{\rm QSO}$ is comparable for all $|b| > 20°$, and comparable to S82. At high latitudes, the increase of candidates with $p_{\rm QSO}$ is similar on and off S82, as illustrated in Fig. 5.11. A sample selected using a lower threshold of $p_{\rm QSO}$ shows inhomogeneities caused by contamination at almost all Galactic latitudes.

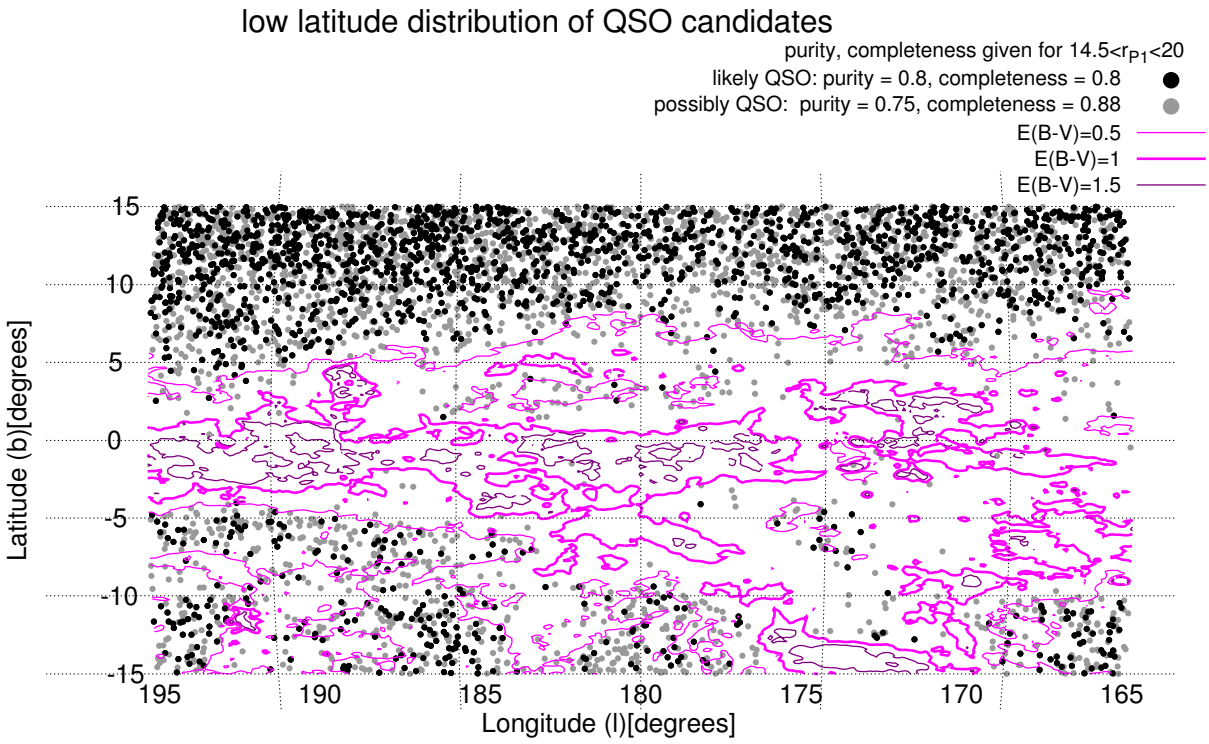
Around the Galactic anticenter (see Fig. 5.12) for $|b| \lesssim 10°$, the number of sources with high $p_{\rm QSO}$ per deg^2 is much lower than around the Galactic north pole, by a factor of $\gtrsim$10. This means that higher overall source density does not lead to an (presumably erroneous) increase of the number of candidate objects with a high $p_{\rm QSO}$. Indeed, the number of candidates per deg^2 decreases, caused by dust or varying WISE depth, to less than 10% of the sources at higher latitudes, and even vanishes for high $p_{\rm QSO}$ (see Fig. 5.13).

Across PS1's entire 3π area, there are 3.7×10^5 likely QSO candidates with $p_{\rm QSO} \geq 0.56$ (with an expected high-latitude purity of 0.8, completeness of 0.8), and 6.9×10^5 possible candidates (purity = 0.75, completeness=0.88) with $p_{\rm QSO} \geq 0.31$.



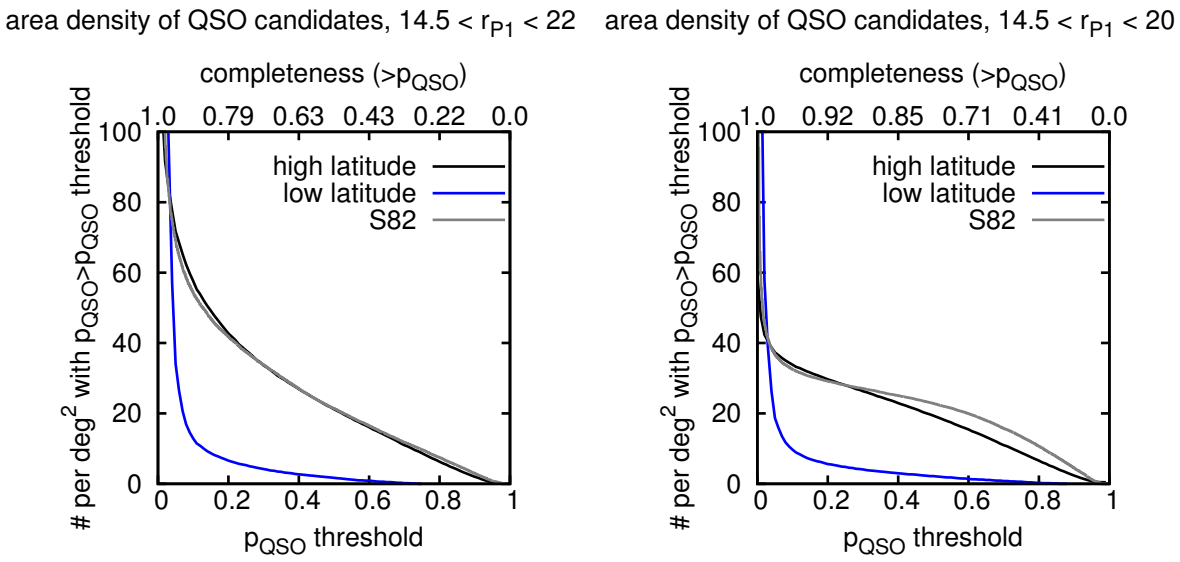
(a)

Figure 5.11 High latitude angular distribution of likely and possibly QSO candidates. This Lambert's Azimuthal Equal-Area Projection with north polar aspect shows well the uniformity of the 5.1×10^4 likely and 9.3×10^4 possibly QSO candidates for $b > 60°$.



(a)

Figure 5.12 Low latitude angular distribution of likely and possibly QSO candidates around the Galactic anticenter. Within the region shown here, there are 4.5×10^3 likely and 1.8×10^3 possibly QSO candidates. This Mollweide projection shows how the area density for both likely and possibly candidates drops towards the Galactic plane, caused by dust. A a contour plot of the reddening-based $E(B-V)$ dust map (Schlafly et al. 2014) overlayed.



(a) (b)

Figure 5.13 Area density of QSO candidates as function of the $p_{\rm QSO}$ threshold. The area density of likely QSO candidates at high latitude ($b > 60°$) is in very good agreement with the one found on S82 for all $p_{\rm QSO}$. At low latitudes around the Galactic center, as shown in Fig. 5.12, the area density drops depending on the $p_{\rm QSO}$ threshold to less than 10% of the high-latitude density. A small, but noticeable difference between S82 and the high-latitude area is found for the brighter sources, see right panel of the Figure.

5.5.2 RR Lyrae Candidates

In this section, the properties of the resulting RR Lyrae candidate sample are presented and discussed. In particular, it is tested whether the completeness and purity of the selection obtained by the method is good enough to recover known halo substructure, as well as whether it can compete with the classification from other surveys the method is not trained on.

Figures 5.14, 5.15 and 5.23 present the diagnostics of our RR Lyrae candidate identification, analogous to the Figures for the QSO candidates. Because the angular and 3D distribution of RR Lyrae is highly structured, diagnosing the quality of the candidate identification across PS1 3π is more complex than for the QSOs. Even Figure 5.14, showing the distribution of likely RR Lyrae candidates around the Galactic north pole, shows gradients and structure; the overdensity seen between $l = 220°$ and $315°$ is the Sagittarius (Sgr) tidal stream. The area density of the likely candidates ($p_{\rm RRLyrae} \geq 0.27$) fits with the expectation of about 1-2 RR Lyrae per deg^2 from SDSS S82 (Sesar et al. 2010).

At low latitudes, around the Galactic anticenter (see Fig. 5.15) where the total source density is 4 times times higher, the number of RR Lyrae candidates with $p_{\rm RRLyrae} \geq 0.27$ drops. The area density of RR Lyrae candidates as function of the $p_{\rm RRLyrae}$ threshold is shown in Fig. 5.16 for different magnitude and therefore assumed distance ranges. Whereas for high latitudes, the

number of candidates is fewer by about $0.5 - 1$ sources per deg^2 than in S82, the density drops more for the low latitudes ($|b| < 20°$), and especially for the bright sources.

This may reflect the combination of higher contamination (reducing the number of $p_{\rm RRLyrae} \geq 0.27$ candidates), with actual RR Lyrae in the Galactic disk. The density of possible RR Lyrae candidates, with $p_{\rm RRLyrae} \geq 0.06$ is much higher than around the Galactic north pole, by a factor of ∼6, which must reflect, foremost, increased contamination. In detail, the source densities are as follows: at $b > 60°$, the density of likely RR Lyrae candidates is 1.7/deg^2, and the density of possibly RR Lyrae candidates 2.9/deg^2. At low latitudes, $|b| < 20°$, the numbers are 3.1/deg^2 and 10.1/deg^2, respectively.

The panoptic view of the PS1-selected RR Lyrae candidates (Fig. 5.23) is quite striking, as it reveals how prominent the Galactic disk and bulge are in the map of likely RR Lyrae candidates. Note that this is in prominent contrast to the large-scale distribution of probable QSOs, whose density drops towards the Galactic plane. Therefore, these data may, in addition to contaminants, be revealing enormous numbers of RR Lyrae candidates throughout the disk and the bulge. Bulge RR Lyrae have been surveyed extensively, e.g. by OGLE (Udalski 2003); yet, to date there have been very few studies of RR Lyrae throughout the Galactic disk (Mateu et al. 2012). This survey therefore represents the largest sample of Galactic disk RR Lyrae candidates, by a wide margin. Of course, they require extensive verification and follow-up.

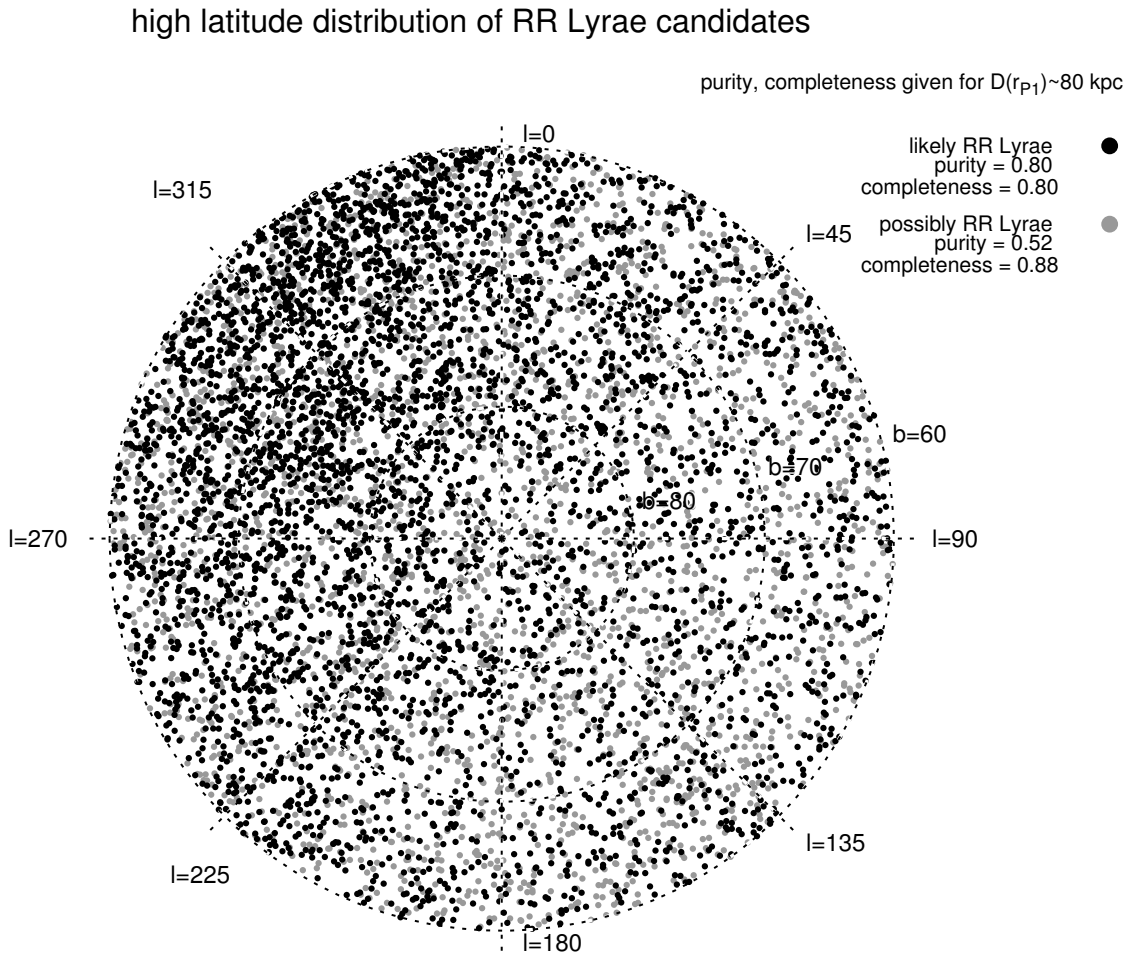
The obtained sample contains 1.5×10^5 likely candidates with $p_{\rm RRLyrae} \geq 0.27$ (purity=0.8, completeness=0.8). The sample contains furthermore 9.0×10^5 possibly halo RR Lyrae candidates at Galactic latitudes of $|b| > 20°$ outside of the bulge, having $p_{\rm RRLyrae} \geq 0.06$ (purity=0.52, completeness=0.88).

Within $|b| < 20°$, where reddening and contamination make the method less likely to be reliable (Section 5.4), the sample contains 1.0×10^5likely RR Lyrae candidates with $p_{\rm RRLyrae} \geq 0.27$ and 3.8×10^5 possibly RR Lyrae candidates with $p_{\rm RRLyrae} \geq 0.06$. Out of them, $3.2 \times 10^4$with $p_{\rm RRLyrae} \geq 0.27$ and 1.4×10^5 with $p_{\rm RRLyrae} \geq 0.06$ belong to the bulge as being in a radius of 20° around the Galactic center. Within the complete area covered by PS1 3π, the sample contains 1.5×10^5 likely and 4.7×10^5 possibly RR Lyrae candidates.

At higher Galactic latitudes, the PS1 3π includes sky regions with known halo substructures or satellite galaxies that contain RR Lyrae, and this can be used to verify our candidate selection. Known structures, clusters and satellite galaxies are labelled[6] in Fig. 5.23. Many of them show up in the map of likely RR Lyrae. Note that M31 and M33 appear in these maps, presumably because their (unreddened) Cepheids get misintepreted as RR Lyrae by our classifier.

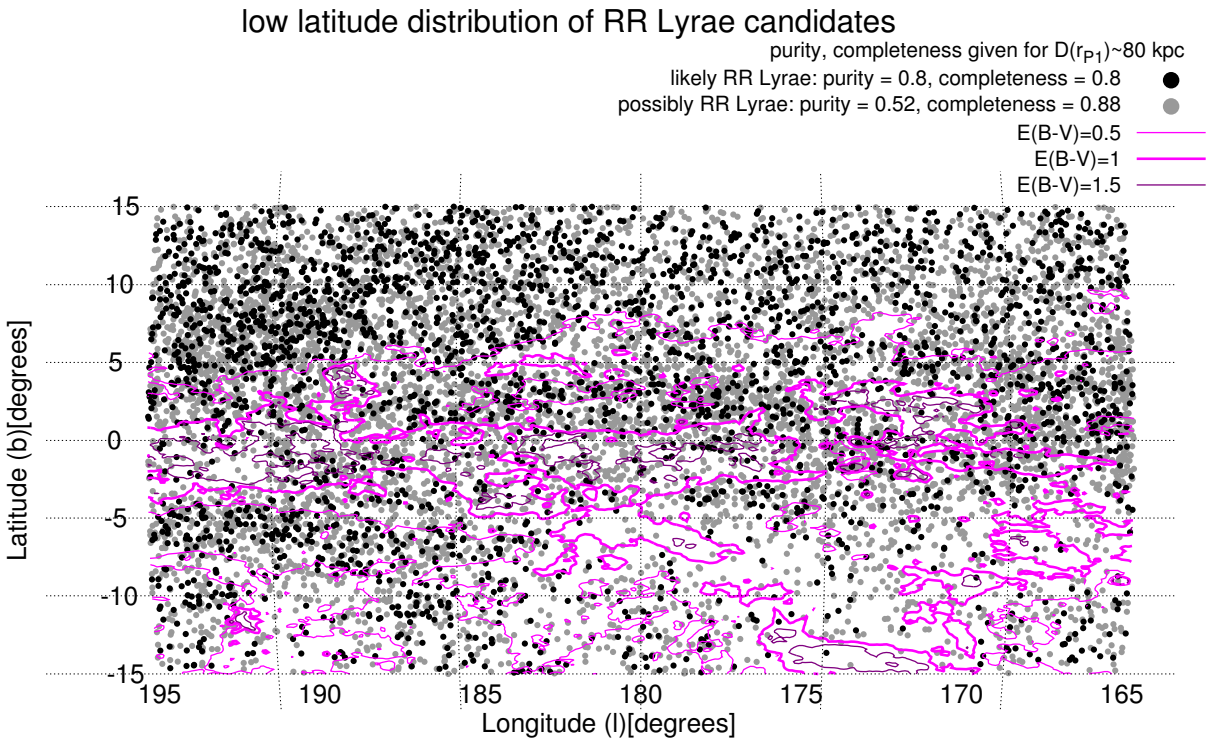
In detail, the Galactic halo substructure as seen by RR Lyrae candidates is depicted in Section 5.5.3

[6] http://homepages.rpi.edu/ newbeh/mwstructure/MilkyWaySpheroidSubstructure.html



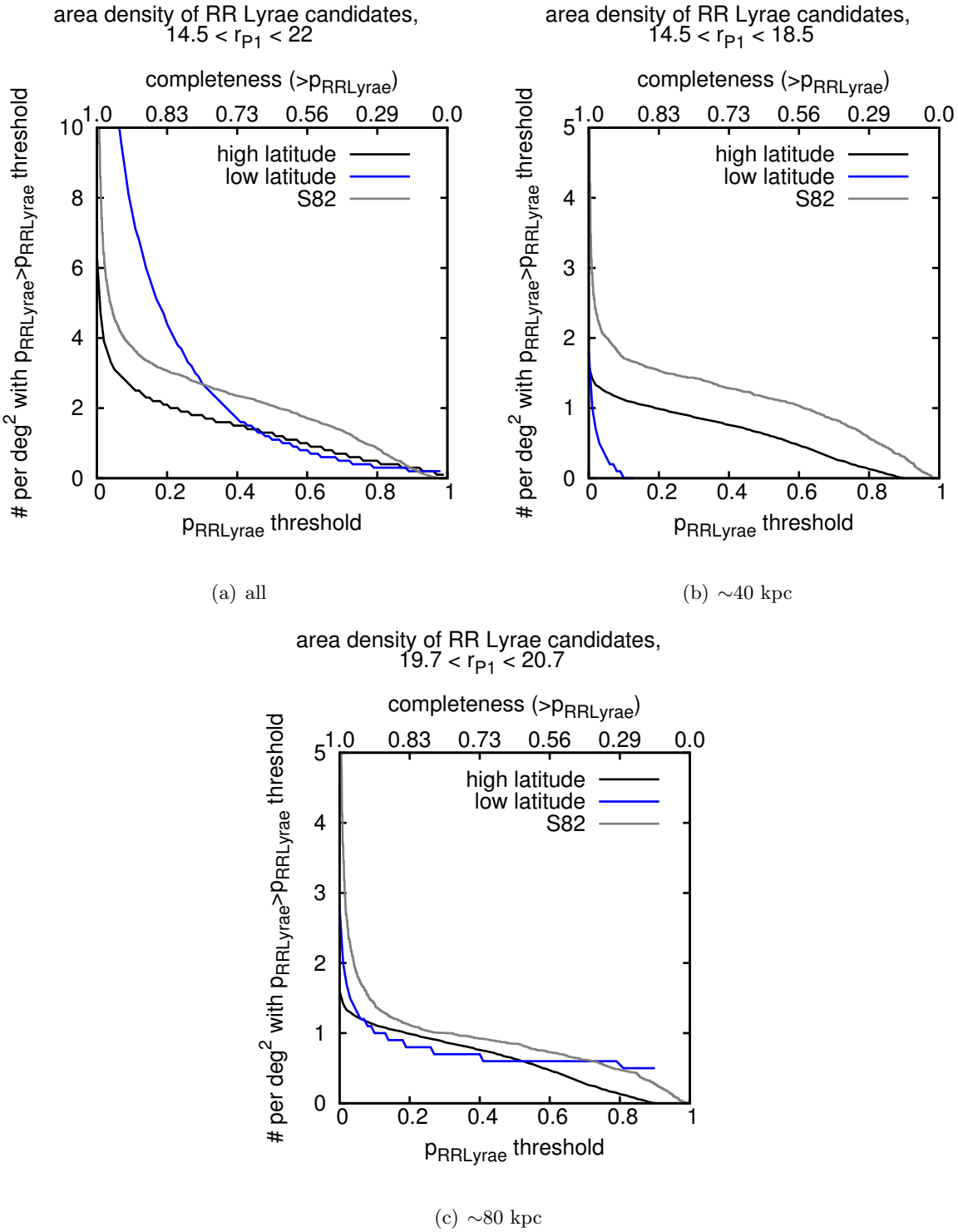
(a)

Figure 5.14 High latitude angular distribution of likely and possibly RR Lyrae candidates. This Lambert's Azimuthal Equal-Area Projection with north polar aspect shows well the inhomogeneity of the 4.8×10^3 likely candidates for $b > 60^\circ$, caused by the Sagittarius stream.



(a)

Figure 5.15 Low latitude angular distribution of likely and possibly RR Lyrae candidates around the Galactic anticenter. This Mollweide projection shows how the area density for both likely and possibly candidates drops towards the Galactic plane, caused by dust. A a contour plot of the reddening-based $E(B-V)$ dust map (Schlafly et al. 2014) overlayed.



(a) all

(b) ∼40 kpc

(c) ∼80 kpc

Figure 5.16 Area density of RR Lyrae candidates as function of the $p_{\rm RRLyrae}$ threshold. The area density of likely RR Lyrae candidates at high latitude ($b > 60\,$deg) is in very good agreement with the one found on S82 for all $p_{\rm RRLyrae}$. At low latitudes around the Galactic center, as shown in Fig. 5.15, the area density of likely RR Lyrae drops depending on the $p_{\rm RRLyrae}$ threshold.

Comparison to the Catalina Survey

Of course, PS1 3π is not the first large-area RR Lyrae survey at high Galactic latitudes; so in selected areas, comparison is possible to previous surveys, e.g. SDSS (York et al. 2000), Catalina (Drake et al. 2009), QUEST (Mateu et al. 2012), and PTF (Rau et al. 2009). Having used SDSS S82 and Draco dSph in the training of the classifier, the analysis here focuses on the Catalina Sky Survey (CSS Drake et al. 2009), which has covered the region around the Galactic north pole down to $b = 30°$, but only for bright sources ≤ 19 mag. CSS is a survey program for finding new near-Earth objects, composed of the original Catalina Sky Survey (CSS), the Siding Spring Survey (SSS) and the Mt. Lemmon Survey (MLS). Catalina photometry covers objects in the range -75° $< \delta <$ 70° and $|b| \gtrsim 15°$. In addition to asteroid search, the complete Catalina data is analyzed for transient sources by the Catalina Real-time Transient Survey (CRTS), resulting in catalogs of RR Lyrae (Drake et al. 2009, 2013a,b). This is used to verify the RR Lyrae candidate identification, by cross-matching in this region with respect to CSS and SSS. The following analysis focuses on the magnitude range in common between both surveys $\sim 15 - 18.5$ mag in order to compare the RR Lyrae candidate sample obtained within this work to the RR Lyrae identified by CSS.

The total number of CSS RR Lyrae with $b > 30°$ is 6855. For 6825 of them, cross-matching finds a PS1 3π source within 5" with $p_{\mathrm{RRLyrae}} \geq 0.27$. The faintest 15 CSS sources, with $V < 12.4$, never enter our analysis.

With respect to CSS, a completeness is obtained of 99% (i.e. finds 99% of their RR Lyrae), and a cross-identified fraction of 42% (i.e. they find 42% of the RR Lyrae candidate sample from PS1 3π), if adopting the above magnitude cuts and p_{RRLyrae} threshold.

When comparing to the SSS RR Lyrae, again the nearest match within a matching tolerance of 5" is used. Restricting to $15 < V < 18.5$, there are 3148 RR Lyrae in the region covered both by PS1 and SSS with $-30 < \delta < -15°$. Out of these, 3115 have $p_{\mathrm{RRLyrae}} \geq 0.27$, resulting in a completeness of 98%. To assess the cross-identified fraction, the area to consider is $|b| > 15°$, as SSS roughly misses $|b| < 15°$. The number of PS1 RR Lyrae candidates in the overlapping region and magnitude range and $p_{\mathrm{RRLyrae}} \geq 0.27$ not cross-matched to SSS is 7539 The number of SSS RR Lyrae within these boundaries is 2725.

In total, this results into a completeness with regard to SSS of 98%, and a cross-identified fraction of 36%.

The sample of likely RR Lyrae candidates from PS1 3π contains $\sim$3 times more RR Lyrae candidates with $p_{\mathrm{RRLyrae}} \geq 0.27$ than the pure samples of CSS and SSS RR Lyrae in the same area. Taken together, CSS and SSS's claim of 70% completeness (Torrealba et al. 2009) and the purity of the PS1 3π RR Lyrae sample being 0.8 at $p_{\mathrm{RRLyrae}} \geq 0.27$, about 56% of the candidates are expected to be cross-identified in CSS or SSS; this is close to the actual fraction of 42% for CSS within 5". in the SSS, a lower cross-identified fraction of 36% is obtained; this suggests that the completeness of the SSS is in fact lower than that of the CSS.

5.5.3 Halo Substructure by RR Lyrae Candidates

Two decades ago, the Milky Way was not thought as an isolated system any longer, as a tidally disrupted dwarf galaxy was found in a stream in the constellation Sagittarius - later called the Sagittarius stream (Ibata et al. 1994). This started the discovery of a number of streams, and highlighted that Milky Way-satellite, or n general galaxy-satellite interactions are a common phenomenon. Also recent cosmological simulations predict tidal streams from disrupted dwarf galaxies in the Milky Way's halo (Bullock and Johnston 2005). At least 11 substantial streams have been detected in the SDSS and 2MASS (Newberg et al. 2002; Majewski et al. 2003; Grillmair 2009).

The Galactic halo is nowadays known as a not homogeneous but structured part of the Milky Way, containing debris streams from both disrupted and existing (i.e. the progenitor is still visible) accreted satellites (i.e., globular clusters and dwarf galaxies). They were disrupted by tidal forces and stretched into stellar tidal streams and clouds. Well-known examples are e.g. the Sagittarius and Ophiuchus streams.

Stellar streams are of great interest as their orbits are sensitive tracers of the Galactic potential. They are a probe the potential's morphology and the total mass of the Milk Way, as disrupted galaxies and globular clusters follow, and therefore trace, the orbit of their progenitor which followed the gravitational potential (Koposov et al. 2010; Newberg et al. 2010; Sesar et al. 2014; Belokurov et al. 2014). Stellar streams are also helpful tracers for galaxy evolution. As dynamical times are very long in the outer regions of the halo, accreted material remains coherent for billions of years (Johnston et al. 1996).

A Map of the Halo

The fact that almost every Milky Way dwarf satellite galaxy has at least one RR Lyrae star (including the faintest one, Segue 1, Simon et al. 2011), enables them to be tracers to locate extremely low-luminous Milky Way dwarf satellites by detecting the faint RR Lyrae within them (Sesar et al. 2014).

The work at hand results into a panoramic map of the entire Milky Way north of $\delta{\sim}30^\circ$ ($\sim$30,000 deg^2), constructed by likely RR Lyrae candidates. Using single-epoch photometry reaching to $r_{\rm P1}{\sim}22$ mag, it is sensitive to stellar substructures with distances up to $\gtrsim 120$ kpc. Within this volume, the map recovers almost all previously reported streams and globular clusters.

The majority of stellar streams known nowadays has been discovered to SDSS, which observed about 14,555 deg^2 of the sky at a depth comparable to PS1 3π. As this area is completely contained within PS1 3π, the ability to recover these features provides a check on the accuracy of the methodology. As a reference, in the following Grillmair and Carlin (2016) is used. The map shown in Fig. 5.23 clearly reveals all prominent structures that have been reported previously,

and shows also sources in most fainter substructure. In particular, the map recovers many features listed by Grillmair and Carlin (2016) clearly, namely the Sagittarius Stream, the Virgo Overdensity, Boötes I dSph, Draco dSph, Sextans dSph, Ursa Minor dSph, as well as the globular clusters NGC1904, NGC4590, NGC5024, NGC5053, NGC5272, NGC5466, NGC5897, NGC5904, NGC6864, NGC6934, NGC6981, NGC7078, NGC7089. For others, scattered instances of some RR Lyrae are found, such as Segue 2, Pisces II, Palomar 13.

In the following, dwarf spheroidals and globular clusters that are recovered clearly – e.g. with an obvious number of stars in an overdensity – are given in alphabtical order. For each of them, a Figure is given in the Figure section, and they are labeled in the map of likely RR Lyrae candidates (Fig. 5.23). In the following, central coordinates are from the SIMBAD Astronomical Object Database[7](Wenger et al. 2000), and the apparent sizes from the paper given for each dwarf spheroidal or globular cluster. Distances D in parsec were derived from

$$D = 10^{((\langle r_{\mathrm{P1}}\rangle_{\mathrm{deredd}} - M_r + 5)/5)} \tag{5.4}$$

where $\langle r_{\mathrm{P1}}\rangle_{\mathrm{deredd}}$ is the dereddened r_{P1} mean magnitude. The absolute r band magnitude $M_r \sim M_V = 0.60$ mag is taken from Sesar et al. (2010).

Some of these objects show possibly tidal features, extending beyond their indicated sizes in a stream-like way. This is especially the case for NGC 5024, NGC 5053, NGC 5272, NGGC 7075. As they are not found in literature, they need further investigation. But for possible tidal features around globular clusters and dwarf spheroidals, this thesis does not contain a thorough examination of false positives.
The apparent sizes are indicated as in the papers in case, where it was often not indicated how this apparent size was derived.

Boötes I dSph (Boo I)
This dwarf spheroidal, located at $l = 358°.0361$, $b = 69°.6423$ at a heliocentric distance of $D \sim 60$ kpc is described in Belokurov et al. (2006).
Boötes was found by Belokurov et al. (2006) in a systematic search for stellar overdensities carried out in the north Galactic cap using SDSS DR5. It shows in the color-magnitude diagramm a well-defined turn-off, red giant branch as well as an extended horizontal branch. With an absolute magnitude of -5.8, it is one of the faintest known galaxies. From its isodensity contours, as shown in Belokurov et al. (2006), its progenitor is likely a dwarf spheroidal galaxy.
Boötes I dSph as appearing in the likely RR Lyrae sample is shown in the upper left panel of Fig. 5.24. Within the apparent radius of $27'$ indicated in the plot, 11 likely RR Lyrae candidates are found, all at $D \sim 60$ kpc.

Draco dSph (Dra)
This dwarf spheroidal, located at $l = 86°.3679$, $b = 34°.7217$ at a heliocentric distance of $D \sim 76$ kpc (75.8 ± 5.4 kpc by Bonanos et al. (2004), 82.4 ± 5.8 kpc by Kinemuchi et al. (2008),

[7]http://simbad.u-strasbg.fr/simbad/

own estimate 75.8 ± 3.9 kpc) is described in Kinemuchi et al. (2008).
Draco dSph as appearing in the likely RR Lyrae sample is shown in the upper right panel of Fig. 5.24. A detailed analysis on found sources and estimated distances is given below in a separate Subsection.

Sextans dSph (Sex)
This dwarf spheroidal, located at $l = 243°.5$, $b = 42°.3$ at a heliocentric distance of $D = \sim 85$ kpc is described in Irwin et al. (1990) Sextans dSph is reported by them as a newly found dwarf elliptical galaxy, discovered using APM measures of UK Schmidt atlas glass copy IIIaJ survey plates. The color-magnitude diagrams reveal a pronounced red horizontal branch and a well-defined asymptotic giant branch typical for dwarf spheroidal systems.
Sextans dSph as appearing in the likely RR Lyrae sample is shown in the lower left panel of Fig. 5.24. Within the apparent radius of $45'$ indicated in the plot, 140 likely RR Lyrae candidates are found, most of them at $D = \sim 85 - 90$ kpc.

Ursa Minor dSph (UMi)
This dwarf spheroidal, located at $l = 104°.9527$, $b = 44°.8028$ at a heliocentric distance of $D = 69 \pm 4$ kpc is described in Mighell and Burke (1999). Ursa Minor dSph was discovered by Wilson (1995) and Hubble independently. This faint dwarf spheroidal is the second closest satellite of the Milky way. Color-magnitude diagrams show a strong horiziontal branch. Ursa Minor may be the only dwarf galaxy within the Local Group that is composed only of stars older than 10 Gyr (Mateo et al. 1998).
Ursa Minor dSph as appearing in the likely RR Lyrae sample is shown in the lower right panel of Fig. 5.24. Within the apparent radius of $15'$ indicated in the plot, 74 likely RR Lyrae candidates are found, most of them at $D = \sim 70$ kpc.

NGC 1904
This globular cluster, located at $l = 227°.2291$, $b = -29°.3515$ at a heliocentric distance of $D \sim 13$ kpc is described in Shapley and Sawyer (1927) NGC 1904 is a globular cluster in the Lepus constellation, discovered by Pierre Méchain in 1780. It is one out of the two extragalactic globular clusters in the Messier catalog (the other is Messier 54). Both are thought to belong to the Canis Major Dwarf Galaxy.
NGC 1904 as appearing in the likely RR Lyrae sample is shown in the upper left panel of Fig. 5.25. Within the apparent radius of $4.8'$ indicated in the plot, 28 likely RR Lyrae candidates are found, with a distance of $D = 12 - 40$ kpc.

NGC 4590
This globular cluster, located at $l = 299°.6258$, $b = 36°.0508$ at a heliocentric distance of $D \sim 10.3$ kpc is described in Brocato et al. (1997) NGC 4590, also known as Messier 68 (M68) was discovered by Charles Messier in 1780, and described by William Herschel later on. NGC 4590 has a highly eccentric orbit ($\epsilon = 0.5$) reaching as far as 30 kpc from the Galactic center. Within the cluster, a total of 50 variable stars are identified up to now, most of them RR Lyrae.
NGC 4590 as appearing in the likely RR Lyrae sample is shown in the upper right panel of Fig.

5.25. Within the apparent radius of 5.5′ indicated in the plot, 41 likely RR Lyrae candidates are found, among them 30 at $D = \sim$10 kpc.

NGC 5024

This globular cluster, located at $l = 332°.9630$, $b = 79°.7642$ at a heliocentric distance of $D = \sim$18 kpc is described in Shapley and Sawyer (1927) and Hessels et al. (2007). It is also known as Messier 53 (M53) and was discovered by Johann Elert Bode in 1775.
NGC 4590 as appearing in the likely RR Lyrae sample is shown in the lower left panel of Fig. 5.25. Within the apparent radius of 6.3′ indicated in the plot, 35 likely RR Lyrae candidates are found, among them ∼20 at $D = \sim$20 kpc.

NGC 5053

This globular cluster, located at $l = 335°.6987$, $b = 78°.9461$ at a heliocentric distance of $D\sim$17.8 kpc is described in Clement et al. (2001) and Boberg et al. (2015). NGC 5053 was discovered by William Herschel in 1786. This globular cluster is located near the north Galactic cap, about 1° of M53, having old stars and being metal-poor. Its horizontal branch stars are about 16.65 mag, its brightest stars up to 13.8 mag.
NGC 5053 as appearing in the likely RR Lyrae sample is shown in the lower left panel of Fig. 5.25. Within the apparent radius of 5.25" indicated in the plot, 11 likely RR Lyrae candidates are found, among them 7 at 19 kpc.

NGC 5272

This globular cluster, located at $l = 42°.2170°$, $b = 78°.7069°$ at a heliocentric distance of $D\sim$10.4 kpc is described in Paust et al. (2010). NGC 5272, also known as Messier 3 (M3) was discovered by Charles Messier in 1764, and resolved into stars by William Herschel around 1784. This globular cluster, one of the largest and brightest, is estimated 8 Gyr old. Among its $\sim 5 \times 10^5$ stars, more than 270 are variables, among them 133 RR Lyrae.
NGC 5272 as appearing in the likely RR Lyrae sample is shown in the lower right panel of Fig. 5.25. Within the apparent radius of 9′ indicated in the plot, 201 likely RR Lyrae candidates are found, among them ∼55 at $D = \sim$10 kpc.

NGC 5466

This globular cluster, located at $l = 42°.1502$, $b = 73.5922$ at a heliocentric distance of $D\sim$15.9 kpc is described in Paust et al. (2010) and Buonanno et al. (1984). NGC 5466 was discovered by William Herschel in 1784. It contains a certain horizontal branch of stars and is metal poor, what makes it unusual. NGC 5466 might be the progenitor of the "45 Degree Tidal Stream" discovered in 2006 (Grillmair and Johnson 2006).
NGC 5466 as appearing in the likely RR Lyrae sample is shown in the upper left panel of Fig. 5.26. Within the apparent radius of 5.5′ indicated in the plot, 31 likely RR Lyrae candidates are found, among them ∼18 at $D = \sim$17 kpc.

NGC 5897

This globular cluster, located at $l = 342°.9460$, $b = 30°.2943$ at a heliocentric distance of $D\sim$12.5 kpc is described in Clement et al. (2001) and Koch and McWilliam (2014). This globular cluster

was discovered by William Herschel in 1784 and shows a low stellar density even in its center. NGC 5897 as appearing in the likely RR Lyrae sample is shown in the upper right panel of Fig. 5.26. Within the apparent radius of 5.5′ indicated in the plot, 15 likely RR Lyrae candidates are found, in the range $D = 10 - 50$ kpc.

NGC 5904

This globular cluster, located at $l = 3°.8587$, $b = 46°.7964$ at a heliocentric distance of D∼7.5 kpc is described in Paust et al. (2010). NGC 5904, also known as Messier 5 (M5) was discovered by Gottfried Kirch in 1702. It is one of the largest known globular clusters and assumed to be about 13 Gyr old, thus being one of the oldest globular clusters in the Milky Way. Among 105 stars in NGC 5904 being variable, 97 are RR Lyrae.
NGC 5904 as appearing in the likely RR Lyrae sample is shown in the lower left panel of Fig. 5.26. Within the apparent radius of 11.5′ indicated in the plot, 284 likely RR Lyrae candidates are found, among them ∼50 at $D = \sim$7 kpc, but also background stars distributed around 40 kpc.

NGC 6864

This globular cluster, located at $l = 0°.3031$, $b = -25°.7480$ at a heliocentric distance of $D = \sim$20.9 kpc is described in Skrutskie et al. (2006) and Harris (1996). NGC 6864, also known as Messier 75 (M75) was discovered by Pierre Méchain in 1780. It has an apparent magnitude of 9.18 and is one of the more densely concentrated globular clusters known.
NGC 6864 as appearing in the likely RR Lyrae sample is shown in the lower right panel of Fig. 5.26. Within the apparent radius of 3.4′ indicated in the plot, 9 likely RR Lyrae candidates are found, among them ∼5 at $D = \sim$21 kpc.

NGC 6934

This globular cluster, located at $l = 52°.1033$, $b = -18°.8929$ at a heliocentric distance of $D = \sim$16 kpc is described in Hessels et al. (2007). It was discovered by William Herschel in 1785.
NGC 6864 as appearing in the likely RR Lyrae sample is shown in the upper left panel of Fig. 5.27. Within the apparent radius of 3.6′ indicated in the plot, 20 likely RR Lyrae candidates are found, among them 23 at $D = \sim$18 kpc.

NGC 6981

This globular cluster, located at $l = 35°.1623$, $b = -32°.6831$ at a heliocentric distance of $D = 16.73 \pm 0.36$ kpc is described in Figuera (2011). It was was discovered by Pierre Méchain in 1780. NGC 6981, also known as Messier 72 (M72) belongs to the apparently smaller and fainter globular clusters in Messier's catalog, being one of its farthest, beyond the Galactic center. Its brightest stars are around 15.8 mag.
NGC 6981 as appearing in the likely RR Lyrae sample is shown in the upper right panel of Fig. 5.27. Within the apparent radius of 3.3′ indicated in the plot, 19 likely RR Lyrae candidates are found, among them ∼18 at $D = \sim$18 kpc.

NGC 7078

This globular cluster, located at $l = 65°.0126°$, $b = -27°.3126$ at a heliocentric distance of

$D = \sim$10 kpc is described in Hessels et al. (2007) and Clement et al. (2001). NGC 7078, also known as Messier 15 (M15) was discovered by Jean-Dominique Maraldi in 1746. Its brightest stars have an apparent magnitude of 12.6, and its horizontal branch giants are at $\sim 15.9 - 16.8$ mag in V (Behr et al. 2000). In NGC 7078, more than 150 variable stars have been found (Clement et al. 2001).
NGC 7078 as appearing in the likely RR Lyrae sample is shown in the lower left panel of Fig. 5.27. Within the apparent radius of $9'$ indicated in the plot, 81 likely RR Lyrae candidates are found, among them 29 at $D = \sim$12 kpc and a background at $D = \sim 40 - 60$ kpc.

NGC 7089
This globular cluster, located at $l = 53°.3709$, $b = -35°.7698$ at a heliocentric distance of $D\sim$10 kpc is described in Hessels et al. (2007) and Lee and Carney (1999). NGC 7089, known as Messier 2 (M2) is one of the largest known globular clusters and was discovered by Jean-Dominique Maraldi in 1746. The age of this compact and significant elliptical globular cluster is estimated as 13 Gyr, thus one of the oldest of Milky Way's globular clusters.
NGC 7089 as appearing in the likely RR Lyrae sample is shown in the lower right panel of Fig. 5.27. Within the apparent radius of $8'$ indicated in the plot, 51 likely RR Lyrae candidates are found, among them 8 at $D = 10 - 12$ kpc and a background at $D = \sim 40 - 60$ kpc.

An obvious and large substructure spanning large fractions of the sky is the Sagittarius stream. It will get an extra section here, and is discussed in detail in Chapter 6.

The Sagittarius Stream

The dominant substructure in the Galactic halo (aside from the Magellanic clouds) is the Sagittarius tidal stream, with its leading and trailing arms (Majewski et al. 2003). Already in Figure 5.23, the Sagittarius tidal stream can be seen as an overdensity crossing $l = 0°$ and $l = 180°$. It is useful to show the geometry of the Sagittarius stream by selecting stars near its presumed orbital plane, and then showing a projected view of this orbital plane. Specifically, Fig. 5.17 shows the angular and distance distribution for RR Lyrae candidates with $p_{\rm RRLyrae} \geq 0.27$ (formal purity=0.8, completeness=0.8) using the heliocentric Sagittarius (orbital plane) coordinates $(\tilde{\Lambda}_\odot, \tilde{B}_\odot)$ defined by Belokurov et al. (2014) and a distance modulus D from the mean magnitude $\langle r_{\rm P1}\rangle$. In this coordinate system, the equator is aligned with the plane of the Sagittarius trailing tail, and $\tilde{\Lambda}_\odot$ increases in the direction of Sagittarius motion. The latitude axis $\tilde{B}_\odot$ points to the Galactic North pole.
Distances D in parsec were taken from

$$D = 10^{\left(\left(\langle r\rangle_{\rm P1,deredd} - M_r + 5\right)/5\right)} \tag{5.5}$$

where $\langle r\rangle_{\rm P1,deredd}$ is the dereddened $r_{\rm P1}$ mean magnitude.
The absolute r band magnitude $M_r \sim M_V$=0.60 mag is taken from Sesar et al. (2010) who used the Chaboyer (1999) M_V − [Fe/H] relation under the assumption that the mean metallicity of

RRab stars is equal to the median metallicity of halo stars ([Fe/H] = −1.5, Ivezić et al. 2008). As this analysis doesn't distinguish between RRab and RRc stars from our analysis, and RRab stars are most common, making up 91% of the observed RR Lyrae, $M_r \sim M_V = 0.60$ mag is used for all RR Lyrae candidates.

Figure 5.17, showing the RR Lyrae candidates in the Sagittarius plane, provides a striking view of the stream, with its trailing and leading arm to distances of about 100 kpc. The structure in this Figure can be compared to to Fig. 6 in Belokurov et al. (2014) as well as to Fig. 6 and 17 in Law and Majewski (2010) that shows the best-fit N-body debris model in a triaxial halo and observational constraints from 2MASS + SDSS for the leading and trailing arm.

The results can be compared to Ruhland et al. (2011), who traced the Sagittarius stellar stream using BHB stars and compared it to Law et al. (2005). From the results of the work at hand, it can be confirmed that there is an extension of the trailing arm at distances of $60 - 80$ kpc from the Sun as given e.g. by Ruhland et al. (2011). Furthermore, a cloud-like overdensity is found at $\tilde{\Lambda}_\odot \sim 110°$, $5 \lesssim D \lesssim 25$ kpc, that can be identified with the Virgo overdensity. This overdensity can be seen in a number of works (Ruhland et al. 2011; Cole et al. 2008; Newberg et al. 2007), but the RR Lyrae candidates show the three-dimensional structure especially clearly. A more detailed analysis on structure and geometry of Sagittarius stream can be found in the subsequent Chapter 6.

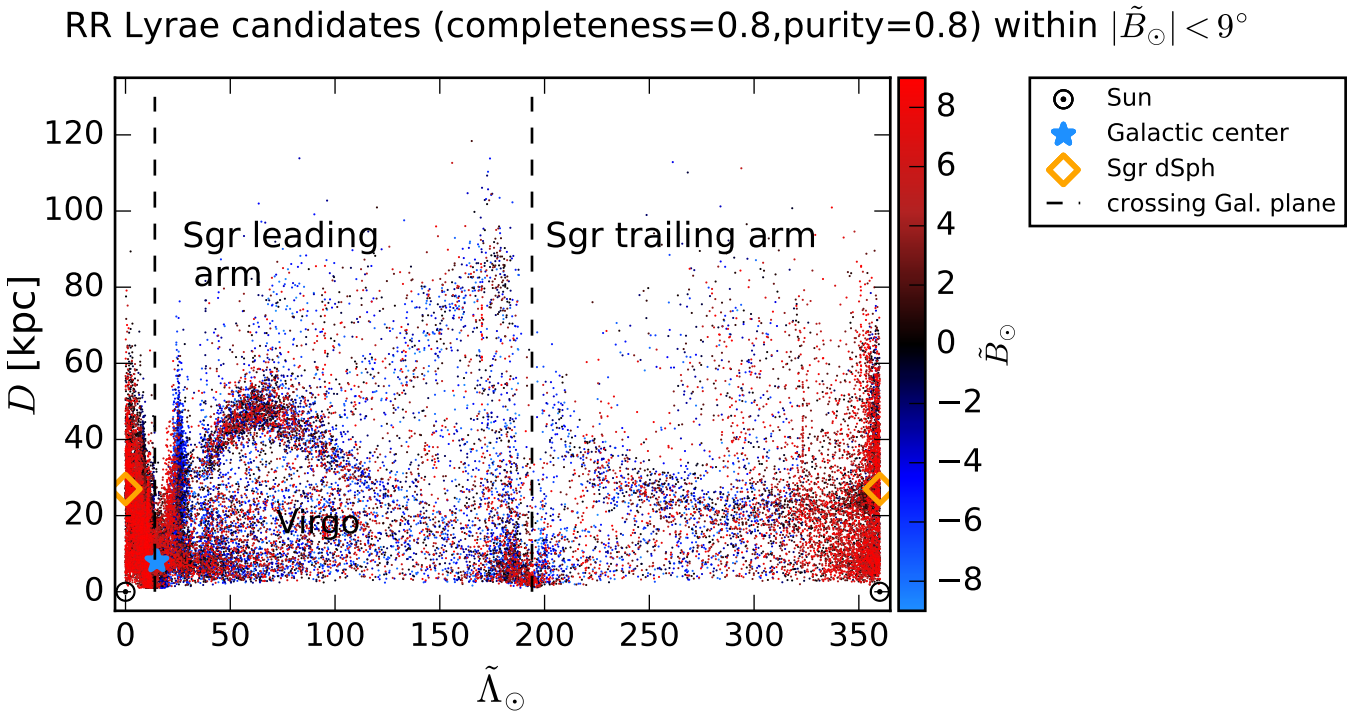


Figure 5.17 The extent of the Sagittarius tidal stream from the distribution of likely RR Lyrae candidates ($p_{\rm RRLyrae} \geq 0.27$, purity=0.8, completeness=0.8) within ±9 deg of the Sagittarius plane, shown in Sagittarius coordinates from Belokurov et al. (2014). The leading and trailing arm of Sagittarius stream can be identified, as well as several substructures up to more than 120 kpc. Distances are from distance modulus of dereddened $r_{\rm P1}$ band mean magnitude. The longitudes of the crossing Galactic plane at $l = 14°$ and $l = 190°$ are marked.

Distance Accuracy from the Draco dSph

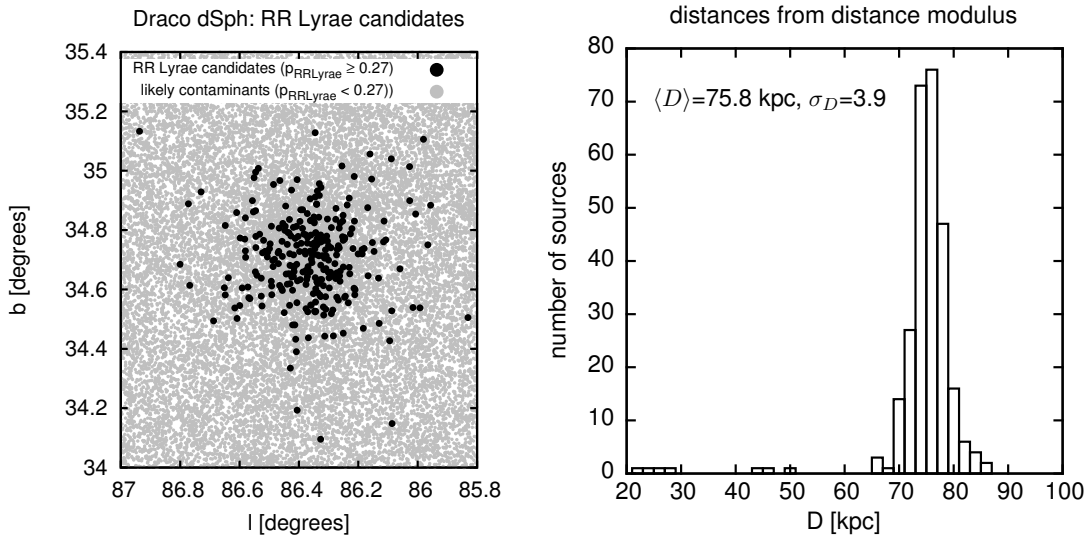
The Draco dwarf spheroidal galaxy (Draco dSph), at known distance and known to contain many RR Lyrae, provides an opportunity to estimate the distance precision of the RR Lyrae candidates, using their inferred mean magnitude in the $r_{\rm P1}$ band. Draco dSph is entirely dominated by old stars, and is affected by near-negligible reddening, which increases the likelihood of dealing with true RR Lyrae stars as compared to the candidates seen in the region of the Galactic disk. Out of the 272 RR Lyrae listed by Kinemuchi et al. (2008), in 269 cases a cross-matching source within 1" is found, all of them having $p_{\rm RRLyrae} \geq 0.27$. Also, among the likely RR Lyrae candidates within a 1.2×1.4 deg^2 patch around Draco dSph, there are only slightly more than in the Kinemuchi et al. (2008) set, namely 279. This results in a completeness of almost 100% w.r.t Kinemuchi et al. (2008), and a cross-identified fraction of 96% (i.e., they find 96% of the likely RR Lyrae candidate sample from PS1 3π within that region).

The first panel of Fig. 5.18 shows the angular distribution of the 279 sources within a 1.2×1.4 deg^2 patch around Draco dSph, having $p_{\rm RRLyrae} \geq 0.27$ (black points); the second panel shows their distribution in distance D. The obtained result of 75.8 ± 3.9 kpc is in very good agreement with Kinemuchi et al. (2008) who found a distance of 82.4 ± 5.8 kpc, and Bonanos et al. (2004) who

found a distance of 75.8 ± 5.4 kpc. Remarkably, the variance in the estimated distances from the likely RR Lyrae candidates is only $\sim$4 kpc, or 5% in distance. This provides an excellent empirical estimate of the distance precision of RR Lyrae candidates, before period-fitting (Sesar et al in prep). Note that many other satellites within $\sim$100 kpc also show clusters of RR Lyrae candidates (see Fig. 5.23).

As Draco dSph is in the training set, it was sensible to test how much of Draco dSph can be found without having it in the training set. In this case, among the likely RR Lyrae candidates within a $1.2 \times 1.4\mathrm{deg}^2$ patch around Draco dSph 83 likely candidates with $p_{\mathrm{RRLyrae}} \geq 0.27$ are found, among them 71 within the Kinemuchi et al. (2008) sample, leading to a completeness of 30, and a cross-identification fraction of 0.85.

This shows that the good match to the sample by Draco dSph is not only introduced by using Draco dSph as part of the training set (i.e.: it is not only reproducing the training set). On the other hand it shows how important it is to enhance the training set by Draco dSph RR Lyrae to identify distant RR Lyrae.



(a) (b)

Figure 5.18 Illustration of the distribution and distance precision for RR Lyrae candidates ($p_{\mathrm{RRLyrae}} \geq 0.27$) around Draco dSph. (a) Angular distribution of likely candidates, compared to that of likely contaminants, (b) distance estimates from distance modulus of dereddened r_{P1} band mean magnitude for the likely RR Lyrae candidates from panel (a). The distance estimates are in very good agreement with Kinemuchi et al. (2008) and Bonanos et al. (2004).

The Halo Profile

In Fig. 5.19, the heliocentric distance distribution of RR Lyrae candidates with $p_{\mathrm{RRLyrae}} \geq 0.27$ (purity=0.8, completeness=0.8) at Galactic latitudes $|b| \geq 20^\circ$ is shown. Half of them are within 20 kpc. The most distant candidates with $p_{\mathrm{RRLyrae}} \geq 0.27$ are $\sim$150 kpc away. An integrated

profile related to a galactocentric halo density profile $\rho\sim\mathrm{D}^{-2.62}$ is overplotted for illustrative purposes. Such a halo profile is in the ball-park of recent determinations (Xue et al. 2015; Deason et al. 2001; Sesar et al. 2013b). Comparing this profile to the distance distribution of our RR Lyrae candidates, this is found to fit well up to ∼80 kpc.

The underlying halo model for the profile, ρ_{halo} (Sesar et al. 2013b) is defined in Galactic coordinates (l, b):

$$\rho_{\mathrm{halo}}(X, Y, Z) = \rho_{\odot\mathrm{RRL}} \left(R_\odot / r\right)^n \tag{5.6}$$

with

$$\begin{aligned} X &= R_\odot - D\cos l\cos b \\ Y &= -D\sin l\cos b \\ Z &= D\sin b \\ r &= \sqrt{X^2 + Y^2 + (Z/q)^2} \\ n &= 2.62 \\ R_\odot &= 8.0\ kpc \\ q &= 0.71 \\ \rho_{\odot\mathrm{RRL}} &= 4.5\ \mathrm{kpc}^{-1}. \end{aligned}$$

$\rho_{\odot\mathrm{RRL}}$ is the number density of RR Lyrae at the position of the Sun.

5.5.4 The Catalog of Variable Sources in PS1 3π

While very useful for many Galactic studies, the existing catalogs of RR Lyrae stars (e.g. Vivas et al. 2001; Sesar et al. 2010, 2013b; Drake et al. 2013b) are not ideal: they are either deep with limited sky coverage (e.g., the SDSS Stripe 82 catalog covers 100 deg^2 and is complete up to 110 kpc, Sesar et al. (2010)), or have a wide coverage but are not very deep (e.g. the CRTS catalog overs 20,000 deg^2 and is complete up to 30 kpc, (Drake et al. 2013b). Also, none of the above catalogs cover the Galactic plane, and thus cannot support studies of the old (>10 Gyr) Galactic disk.

For PS1 3π PV2, 3.88×10^8 PS1 3π sources that fulfil the cuts described in Sec. 5.3.1 were processed. Supplementary to Hernitschek et al. (2016), the paper based on PV2, a catalog of all likely variable point sources in PS1 and of all likely QSOs is provided, having a total of 2.6×10^7 sources. The catalog includes all sources fulfilling the criterion of $\log_{10}\hat{\chi}^2 > 0.5$ (see Fig. 5.3) or $W12 > 0.5$. The latter criterion is intended to ensure that variability statistics are provided for almost all QSOs.

The Catalog of Variable Sources is available in its entirety in machine-readable format in the supplementary material to Hernitschek et al. (2016). A table structure is shown here for guidance regarding its form and content.

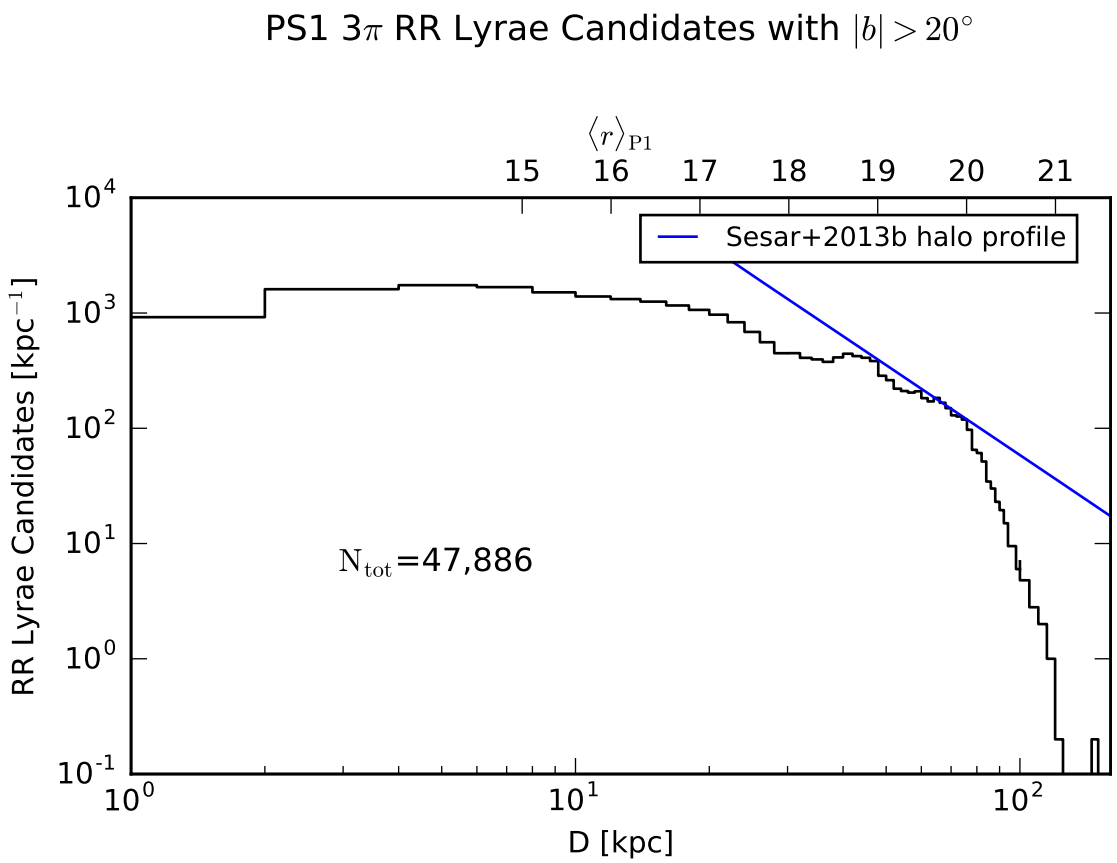


Figure 5.19 Distribution of the heliocentric distance estimates for halo RR Lyrae candidates ($p_{\rm RRLyrae} \geq 0.27$, $|b| > 20^\circ$). The corresponding apparent $r_{\rm P1}$ band magnitude, with no reddening assumed, is given. Distance estimates are done from distance modulus of dereddened $r_{\rm P1}$ band mean magnitude. The figure shows the distances for the 47,886 out of 48,199 halo RR Lyrae candidates within $|b| > 20^\circ$ having $r_{\rm P1}$ band mean magnitude available. An integrated number density profile from Equ. (5.6), $\sim$D$^{-1.62}$ is overplotted.

For PV3, a similar catalog can be built from the 1.1×10^9 sources that were processed. Under the same criterion as above, it would contain 7.7×10^7 sources (unpublished).

5.6 Period Finding for RR Lyrae Candidates

In a subsequent work by Branimir Sesar (Sesar, Hernitschek et al. 2016), features extracted by the author (see Table 5.7) are used together with template fitting in order to determine periods of RR Lyrae candidates and thus enhance the sample purity even more. The method is outlined here, as the cleaner sample together with precise distance estimates is used in the analysis presented in Chapter 6.

Table 5.4. The Catalog of Variable Sources in PS1 3π

Column	FITS Format Code	Description
1	E	right ascension in degrees
2	E	declination in degrees
3	E	scalar variability quantity $\hat{\chi}^2$, Equ. (5.1)
4	E	best fit structure function parameter ω_r (r band variability amplitude) on log-spaced grid
5	E	best fit structure function parameter τ (time scale) on log-spaced grid
6	E	error-weighted mean $g_{\rm P1}$ band magnitude $\langle g_{\rm P1}\rangle$
7	E	error-weighted mean $r_{\rm P1}$ band magnitude $\langle r_{\rm P1}\rangle$
8	E	error-weighted mean $i_{\rm P1}$ band magnitude $\langle i_{\rm P1}\rangle$
9	E	error-weighted mean $z_{\rm P1}$ band magnitude $\langle z_{\rm P1}\rangle$
10	E	error-weighted mean $y_{\rm P1}$ band magnitude $\langle y_{\rm P1}\rangle$
11	E	W1-W2 color from WISE
12	E	$p_{\rm QSO}$
13	E	$p_{\rm RRLyrae}$

Note. — Structure of the Catalog of Variable Sources in PS1 3π. The Catalog of Variable Sources is available for PV2 in its entirety in machine-readable format in the supplementary material to Hernitschek et al. (2016). A table structure is shown here for guidance regarding its form and content.

5.6.1 Template Fitting

For period finding, a template based method was applied. The paper by Sesar, Hernitschek et al. (2016) states that previous tests with the multi-band periodogram of VanderPlas and Ivezić (2015) (for a description see Sec. 4.1.3) had shown that it will fail with the data at hand for nearly half of the S82 RR Lyrae with PS1 3π photometry. The reason is that the model by VanderPlas and Ivezić (2015) is a mathematical multi-band light curve, thus for sparse data, the optimal estiamated light-curve shape (the best-fit model) will not necessarily be physical.

Sesar, Hernitschek et al. (2016) adopt a set of 482 of *griz* models, consisting of the lightcurve templates derived in the work by Sesar et al. (2010) from SDSS S82 RR Lyrae (with SDSS photometry). In total, the template set consists of 379 type *ab* multi-band templates and 104 type *c* multi-band templates for RRab and RRc stars, respectively.
In the subsequent analysis, Sesar, Hernitschek et al. (2016) use all $y_{\rm PS1}$ and $z_{\rm PS1}$ observations as if they would be from the same band ($z_{\rm PS1}$), as they had found from phased PS1 3π light curves of RR Lyrae that both are identical within photometric uncertainties.

To find the best-fit values of these parameters, the phase of each light curve given an assumed period $\mathcal{P}$ is

$$\phi(t|\mathcal{P},\phi_0) = \frac{(t-2400000)\,\mathrm{modulo}\mathcal{P}}{\mathcal{P}} + \phi_0. \tag{5.7}$$

In this equation, t is given in heliocentric Julian days, and the phase offset $-0.5 < \phi_0 < 0.5$ is used to enforce the maximum of the light curve occurring at $\phi = 0$.
Then, a χ^2-like statistic is minimized,

$$\chi^2_k = \sum_{m=g_{\rm P1},r_{\rm P1},i_{\rm P1},z_{\rm P1},y_{\rm P1}} \sum_{n=1}^{N_{\rm obs}} \left(\frac{m_n - m_k(\phi(t_n|\mathcal{P},\phi_0)|F,r')}{\sigma_{m_n}} \right)^2 . \tag{5.8}$$

In this equation, σ_{m_n} is the photometric uncertainty for the n-th observation in the $m = g_{\rm P1}, r_{\rm P1}, i_{\rm P1}, z{\rm P1}, y_{\rm P1}$ band. Sesar, Hernitschek et al. (2016) use the Differential Evolution algorithm of Storn and Price (1997) in order to minimize. this algorithm is very fast compared to others tested, and is already implemented in `scipy` (Millman and Aivazis 2011).

During the template fitting, to a given PS1 3π light curve first every template is fitted. While doing so, the minimization is constrained to periods from 0.4–0.9 days for type *ab* templates, and to 0.2–0.5 days for type *c* templates, as being typical for these classes. After that, depending on the type of the best-fit template, only type *ab* or *c* are fitted, now setting a more restrictive prior on the permissible periods, begin in the range of 2 min around the previous best-fit template. For further inference, only the best-fit outcome from this method is used.

As a test set, the S82 periods from SDSS were used by Sesar, Hernitschek et al. (2016). The method is capable of recovering the period for 85% out of 440 RR Lyrae. Even with a precision of 1 sec, the method can recover the period for 73% from PS1 3π photometry.

Thus, period estimation from template fitting is a powerful tool in feature extraction for subsequent classification.

5.6.2 A Cleaner Sample

The method is then applied by Sesar, Hernitschek et al. (2016) to full PS1 3π, where a more rigid cut than the one described in Section 5.3.2 is needed. For being processed, they require that the sources meet the following conditions after outlier cleaning took place:

(i) at least two epochs in each $g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$ bands, and at least a total of two epochs in $z_{\rm P1}$, $y_{\rm P1}$
(ii) a total of at least 23 epochs
(iii) an uncertainty-weighted mean magnitude of $15 < \langle m \rangle < 21.5$ in at least one of the $g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$ bands.

Sesar, Hernitschek et al. (2016) use the estimated period $\mathcal{P}$, a set of 10 features comparable to the ones used within this work (see Table 5.7), as well as ∼20 other features. The complete set of features is described in Sesar, Hernitschek et al. (2016) and are used to train a supervised classifier comparable to the approach described in Section 5.4.3.

The important differences are:

(i) the training set is similar to the S82-part of the training set used in this work (see Section 5.4.3), with the difference of sources must have at minimum 23 (instead of 10) epochs
(ii) the feature set of Table 5.7 is adopted, but replaced the feature `stellar_locus` (see Equ. (5.3)) used by Hernitschek et al. (2016) with $g_{\mathrm{P1}} - i_{\mathrm{P1}}$
(iii) instead of a RFC, a gradient boosting (as described in Section 4.2.2, implemented as `XGBoost` (Chen and Guestrin 2016) is used.

The approach uses three subsequently more detailed classifiers to overcome the huge computational effort of template fitting. As one template fitting takes ∼30 min CPU time per source, it will not be feasible to carry it out for the majority of the sources.
The three subsequent classifiers, as outlined in Sesar, Hernitschek et al. (2016) use

(i) optical/NIR colors and variability features
This analysis uses the feature set of Table 5.7, but replace the feature `stellar_locus` by Hernitschek et al. (2016) by $g_{\mathrm{P1}} - i_{\mathrm{P1}}$. The analysis is in purity and completeness comparable to the results presented by the author in Sec. 5.5.2. The classifier by Sesar, Hernitschek et al. (2016) also selects samples that are, for example, ∼80% pure and ∼80% complete. A cut is set on the outputted classification score to reduce the sample by more than three orders of magnitude, while losing only 2% of the true RR Lyrae.

(ii) multi-band periodogram
Multi-band periodograms are calculated for the sources passing step (i). The best 20 periods, as well as their power (i.e. height of the periodogram at the given period) are extracted. The classifier now uses the features of (i) plus the 20 best periods and their powers, resulting in 50 features.

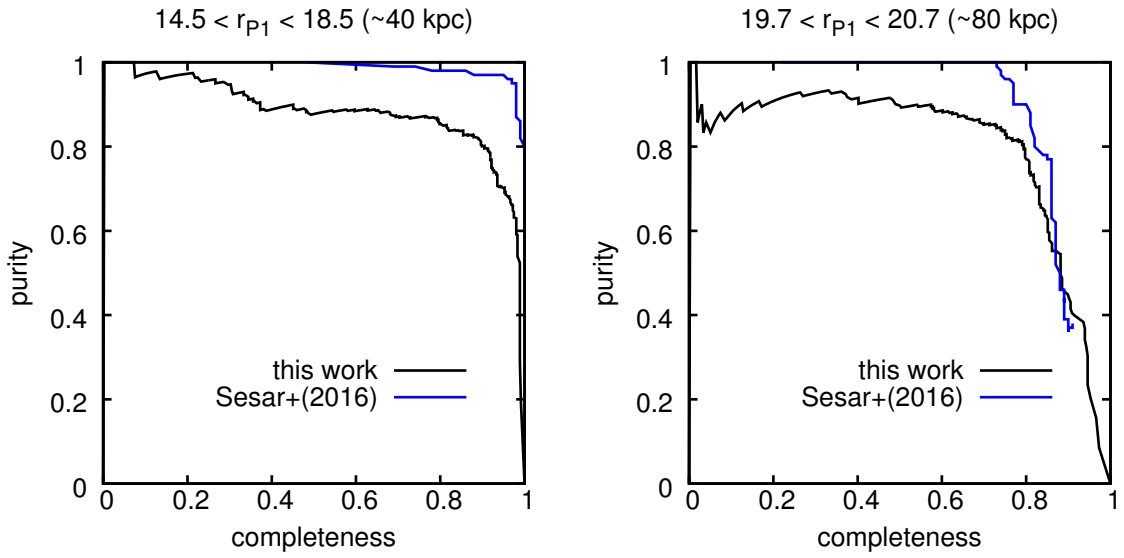
(iii) template fitting
To sources with a high classifier score from (ii), template fitting is applied.

This strategy avoids wasting CPU time for estimating periods or fitting templates to sources that are likely not RR Lyrae.

Comparing the purity and completeness of this classifier to the one by the author as described in Sec. 5.5.2, after the final step (iii) samples that are, for example, 90% pure and 90% complete can be selected.
Fig. 5.20 compares the purity-completeness curves from the classifier developed in this work to the purity-completeness curves after step (iii) from Sesar, Hernitschek et al. (2016) in order to show the remarkably high purity and completeness that can be reached using multi-band template fitting.

For exploring the Galactic halo, precise distances are important. To estimate distances, methods as Equ. 5.5 rely on knowledge on the metallicity, that is mostly not available.

Having the period at hand due to the analysis described before, it can be used for a more precise distance determination. In detail, this period can be used in combination with a period-absolute



(a) (b)

Figure 5.20 Trade-off between purity and completeness for the classifier within this work and the classifier of Sesar, Hernitschek et al. (2016).
Comparison of both classifiers for the (a) bright, (b) faint end of the sample. For both magnitudes ranges, the classifier by Sesar, Hernitschek et al. (2016) using period fitting provides a significant improvement, enabling the selection of samples being almost 100% pure at a completeness of ∼80% (faint end) or even > 90% (bright end).

magnitude-metallicity (PLZ) relation to measure the distances of RR Lyrae with a precision of 3% without knowledge on the metallicity (Sesar, Hernitschek et al. 2016).

The distance estimation by Sesar, Hernitschek et al. (2016) relies on empirical studies by Catelan et al. (2004), Sollima et al. (2006), Marconi et al. (2015) and other. They have shown that the absolute magnitude of a RR Lyrae can be modeled as

$$M = \alpha \log_{10}(\mathcal{P}/\mathcal{P}_{\rm ref}) + \beta([{\rm Fe/H}] - [{\rm Fe/H}]_{\rm ref}) + M_{\rm ref} + \epsilon \tag{5.9}$$

where $M_{\rm ref}$ is the absolute magnitude at a reference period $\mathcal{P}_{\rm ref}$ and metallicity $[{\rm Fe/H}]_{\rm ref}$. The variables α and β give the dependence of the absolute magnitude on period and metallicity, respectively. The intrinsic scatter in the absolute magnitude is modeled by ϵ, being a standard normal variable with mean 0 and standard deviation σ_M.
The details of how to constrain distances using this method are outlined in detail in Sesar, Hernitschek et al. (2016).

In the end, distance moduli of PS1 3π RR Lyrae candidates are computed from the flux-averaged mean $i_{\rm P1}$ magnitude as

$$M_{i_{P1}} = -1.77 \log_{10}(\mathcal{P}/0.6) + 0.49 \tag{5.10}$$

with an uncertainty in $M_{i_{P1}}$ of 0.06 mag, i.e. a ∼3% uncertainty in distance.

These distances of a sample selected by the method described in Section 5.6.2 (having a purity of 0.9, completeness of 0.8 at 80 kpc) will be used in Chapter 6 in order to precisely map the geometry of Sagittarius stream.

5.7 Discussion and Outlook

This part of the work entails how to identify, characterize and classify variable (point) sources in the PS1 survey, the most extensive, deep, multi-band, wide-area, multi-epoch imaging survey to date. Because photometry in different bands of PS1 is not observed simultaneously (as they were e.g. in SDSS), a new methodology for multi-band fitting of structure functions was developed, implemented and carefully tested, used to characterize non-simultaneous multi-band lightcurves. This allows to assign to each of 1.1×10^9 point sources in PS1 3π a basic, χ^2-based variability indicator, a variability amplitude (in the $r_{\rm P1}$-band) ω_r, and a variability time-scale τ.

The analysis then focused on the identification of two classes of variable sources among these objects, QSOs and RR Lyrae stars. Because it aids enormously in the identification of QSOs, additional to PS1 3π photometry, complementary WISE data was used. To classify objects on the basis of this mean photometry and lightcurves, the fact was utilized that SDSS Stripe 82 is covered by PS1 3π, as well as Draco dSph, providing together a full inventory of QSOs and RR Lyrae in these areas. Taking this as ground truth, a Random Forest Classifier was trained to classify all sources in PS1 3π that have at least 10 epochs after outlier cleaning.

To test the effects of missing information on classification, the classification was not only carried out with the full available feature set of variability parameters and colors from PS1 together with WISE colors, but also with more restricted pieces of information, using only color-related and only variability-related parameters. This had shown that the variability information is absolutely indispensable to define a sample with an interesting combination of purity and completeness.

One important limitation of the classification is that it relies on SDSS S82 for QSOs and RR Lyrae, supplemented for faint RR Lyrae within Draco dSph. While this area covers a wide range in Galactic latitude, $20° < b < 70°$ for S82 and ∼1 deg^2 for Draco dSph, there is no training in the Galactic plane. While the number of very likely RR Lyrae candidates, $p_{\rm QSO/RRLyrae} > 0.27$ drops near the Galactic plane, the number of possible RR Lyrae candidates, with $p_{\rm RRLyrae} \geq 0.06$ is much higher than around the Galactic north pole, by a factor of ∼6, which must reflect, foremost, increased contamination. In detail, the source densities are as follows: at $b > 60°$, the density of likely RR Lyrae candidates is 1.7/deg^2, and the density of possibly RR Lyrae candidates 2.9/deg^2. At low latitudes, $|b| < 20°$, the numbers are 3.1 deg^2 and 10.1 deg^2.
This fact implies, unsurprisingly, the likely presence of a considerably higher contamination, at least for samples with purity < 0.8, than tests in S82 would imply. The purity of low-latitude samples must be settled with follow-up observations and analysis. However, at higher Galactic

latitudes, $\gtrsim 20°$, PS1 3π appears to remain quite complete in its selection to $r_{\rm P1}\sim$21 mag, which enables candidate selection to more than ∼140 kpc.

As the treatment of reddening is limited right now, care must be taken applying any values of purity and completeness to regions of high reddenings.

Across the entire 3π, the analysis of PS1 3π PV3 identified 1.5×10^5 RR Lyrae candidates with $p_{\rm RRLyrae} \geq 0.27$. Based on the training, a purity and completeness of both 0.8 is expected among cross-matched sources. As mentioned above, these numbers on purity and completeness only apply away from the Galactic plane, and the bulge.

With this caveat on the low-latitude sample purity, the spatial distribution of RR Lyrae candidates is as follows: Within $|b| < 20°$, i.e. near the disk, there are 1.0×10^5 likely RR Lyrae candidates with $p_{\rm RRLyrae} \geq 0.27$ and 3.8×10^5 possibly RR Lyrae candidates with $p_{\rm RRLyrae} \geq 0.06$. Of them, 3.2×10^4 with $p_{\rm RRLyrae} \geq 0.27$ and 1.4×10^5 with $p_{\rm RRLyrae} \geq 0.06$ may be in the bulge as being in a radius of 20° around the Galactic center. Here the selection cuts on the parameter $p_{\rm RRLyrae}$ are mentioned rather than purity and completeness, because the mapping to purity and completeness in S82 may not apply at such low latitudes. In the Galactic halo, at Galactic latitudes of $|b| > 20°$ there are 4.8×10^4 candidates with $p_{\rm RRLyrae} \geq 0.27$, some extending to distances as large as ∼150 kpc.

This is the most extensive and faintest RR Lyrae candidate sample to date, extending to considerably fainter magnitudes than e.g. the CRTS sample of RR Lyrae stars.
Using the selected RR Lyrae in Draco, distances derived from $\langle r_{\rm P1} \rangle$ and $M_r = 0.6$ are precise to 6% at a distance of ∼80 kpc. A projection of the candidate sample into the orbital plane of the Sagittarius stream reveals the stream morphology clearly. Additionally, this sample shows a bunch of streams and satellites clearly. This indicates that this sample will be excellent for mapping stellar (sub-)structure in the Galactic halo.

Furthermore, there are 3.7×10^5 likely QSO candidates over the total PS1 3π area at a level of purity of 0.8, completeness of 0.8 ($p_{\rm QSO} \geq 0.56$), and 6.9×10^5 possible candidates at a level of purity of 0.75, completeness of 0.88 ($p_{\rm QSO} \geq 0.31$). At $|b| > 20°$, there are 3.3×10^5 candidates with $p_{\rm QSO} \geq 0.56$ and 6.1×10^5 candidates with $p_{\rm QSO} \geq 0.31$. The selection of candidates is homogeneous to a high degree away from the Galactic plane. Around the plane, the number density of QSO candidates with high $p_{\rm QSO}$ decreases because of dust.

Over all, this work has resulted in estimation of variability parameters and mean magnitudes for more than 1.1×10^9 sources, and a catalog of variable sources of almost 2.58×10^7 objects, being available as a 3π value-added catalog. These parameters of course allow the source classification based on different training sets than the one presented here.

These results of PS1 3π variability studies in the Milky Way context offer the possibility for all-sky detection of variable sources in general and to use RR Lyrae to precise distance estimates for finding streams and satellites. QSO candidates will be used as a reference frame for Milky Way

astrometry, to get absolute proper motions and study Milky Way disk kinematics. The general approach enables also the selection of Cepheid variables, as briefly summarized below.

Candidates of periodic variables can be processed further to increase their purity. As approaches for period finding and fitting are very computational expensive, it is necessary to apply it to pre-selected candidates (VanderPlas and Ivezić 2015). This is especially outlined by the approach of Sesar, Hernitschek et al. (2016).

Several approaches for detecting period lightcurve signals exist for well-sampled single-band data (e.g. Sesar et al. 2010; Graham et al. 2013), but must be adopted for the randomly sampled multi-band lightcurves as present from PS1 and LSST. Promising approaches for detecting periodicity in sparsely sampled multi-band time domain data are the multiband periodogram (VanderPlas and Ivezić 2015) as well as lightcurve template fitting (Sesar, Hernitschek et al. 2016).

Looking forward to catalogs of variable stars from Pan-STARRS, LSST and other multi-band all-sky time-domain surveys, the general approach of multi-band structure functions and mean magnitudes as features for a machine-learning classifier meets the constraints of being able to deal with noisy observational through different bands, accompanied by data from other surveys, and is fast enough to provide a sample pure and complete enough for further lightcurve analysis.

Cepheids

The approach applied so far in order to find highly pure and complete QSO and RR Lyrae candidate samples can be extended to find Cepheids in PS1 3π.

Cepheids in the Milky Way's disk are interesting as they are tracers of structure and evolution of the inner disk. High dust reddening and dust obscuration, as well as high source density poses difficulties to map the inner Galaxy down to $|b| < 5°$. Attempts were made by the VISTA Variables in the Via Lactea (VVV) ESO Public Survey (Catelan et al. 2011), using near-infrared time-series photometry, who revealed 35 classical Cepheids (Dékány et al. 2015).

In the following, a brief outlook is given on how the approach presented so far can be extended for Cepheids.

In order to build a training set, mock light curves are generated. Most of the known Milky Way Cepheids are too bright to build a training set of appropriate size. The following procedure is carried out to generate mock light curves (Laura Inno 2015, unpublished):

(i) Producing normalized light-curves templates in PS1 3π $g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$, $z_{\rm P1}$, $y_{\rm P1}$ bands based on 131 Cepheids from Monson and Pierce (2011) and the color transformation from Tonry et al. (2012).

(ii) Selecting PS1 3π sample light curves at various lines of sight close to the Galactic plane to obtain epoch sampling, together with reddening information from Marshall et al. (2006).

(iii) Make mock curves based on the obtained epoch sampling and reddening: each of the 131 Cepheids is "placed" at various lines of sight (thus, epoch sampling and reddening applied). Additionally, photometric errors are modeled.

The training set is composed of these mock light curves, plus PS1 3π light curves of "other" sources (being not known as Cepheids, QSO or RR Lyrae).
Training and classification is then done as previously or RR Lyrae: The feature set shown in Table is used to train a RFC, using variability and colors from PS1 3π, colors from WISE. Once likely Cepheid candidates are found, template fitting similar to that for RR Lyrae (see Sec. 5.6.2) is carried out.

For testing the classification, 46 Cepheids from the Kiso Observatory Survey (unpublished) are used, all of them within $|b| < 2°$. Within $0.1° \times 0.1°$ around each known Kiso Cepheid, PS1 3π light curves are obtained and classified.

True Cepheids from Kiso get typically a $p_{\mathrm{Cepheid}\geq 0.98}$ what is comparable to a completeness of 0.99, purity of 1. There are typically <10% sources getting a $0.4 < p_{\mathrm{Cepheid}<0.8}$, and the vast majority of sources has a $p_{\mathrm{Cepheid}<0.1}$. In total, among 40522 sources in the sample, 109 are identified as likely Cepheids with $p_{\mathrm{Cepheid}>0.98}$.

With respect to Kiso, for $p_{\mathrm{Cepheid}\geq 0.98}$ a completeness is obtained of 0.8 (i.e. finding 80% of their Cepheids, 37 out of 46), and a cross-identified fraction of 0.33 (i.e. they find 33% of the Cepheid candidate sample from PS1 3π, 37 out of 109), if adopting the above magnitude cuts and p_{RRLyrae} threshold. This points vaguely to a completeness of $\sim$0.8, purity of $\sim$0.33 for Cepheid candidates, but requires further investigation and possibly improvement of the feature set.

Technical Remarks

For the whole project, a total of 3.3×10^5 CPU hours of super-computing time was used by the author for PV2 and PV3 data each, among them 3.28×10^5 CPU hours for structure-function fitting and 680 for classification.
The computations are performed using C++ code for structure function estimation, and Python for classification.
An additional 4.0×10^5 CPU hours was used for the subsequent period finding by Branimir Sesar.

The PS1 3π catalog used in this work to obtain light curves is stored in the Large Survey Database (LSD) format (Jurić et al. 2011), which allows for a fast and efficient manipulation of very large catalogs ($>10^9$ objects).

5.8 FIGURES

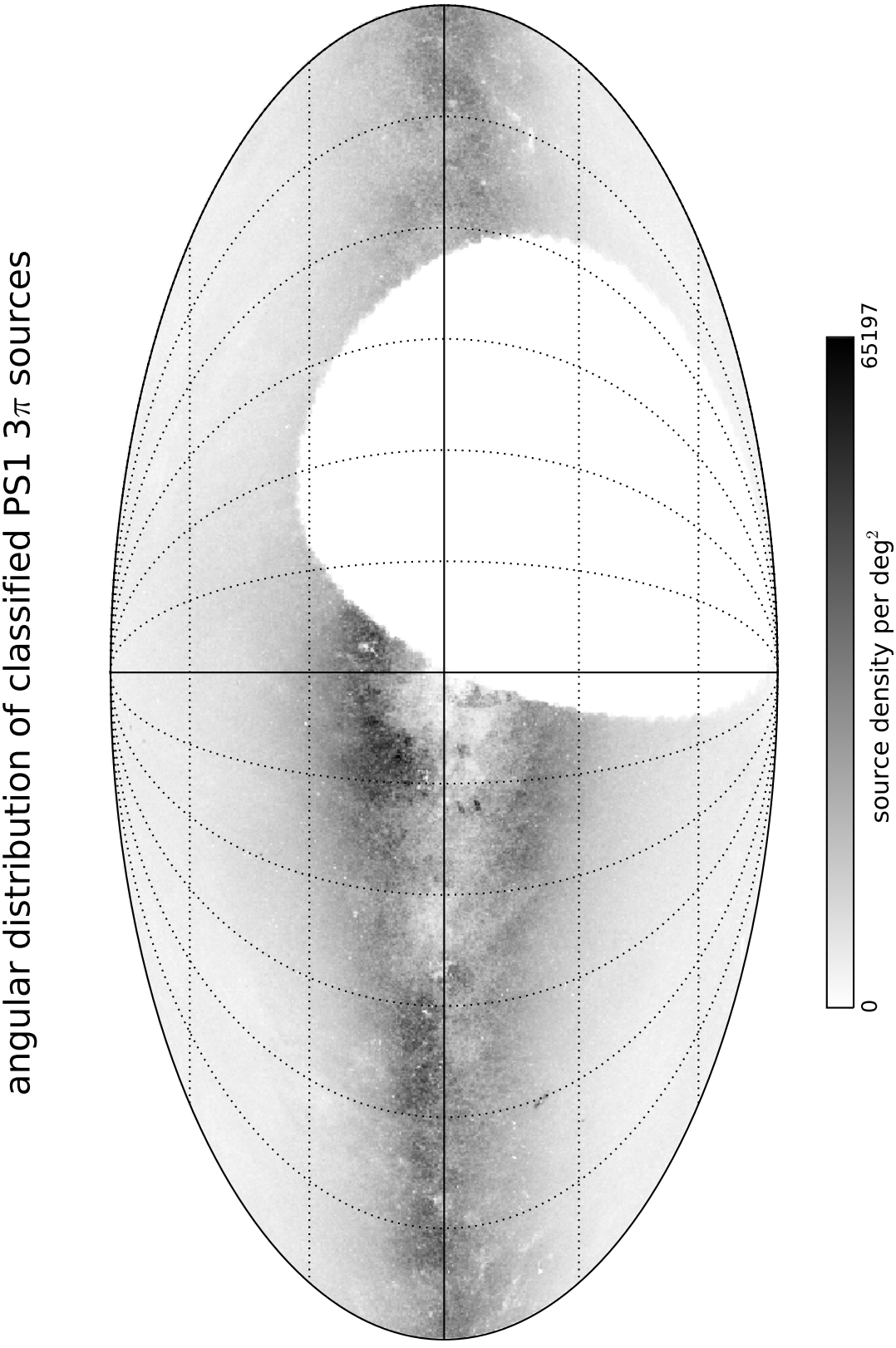


Figure 5.21 Density of processed 1.1×10^9 PS1 3π sources as Mollweide projection in Galactic coordinates using the healpy (https://healpy.readthedocs.org) pixelation.

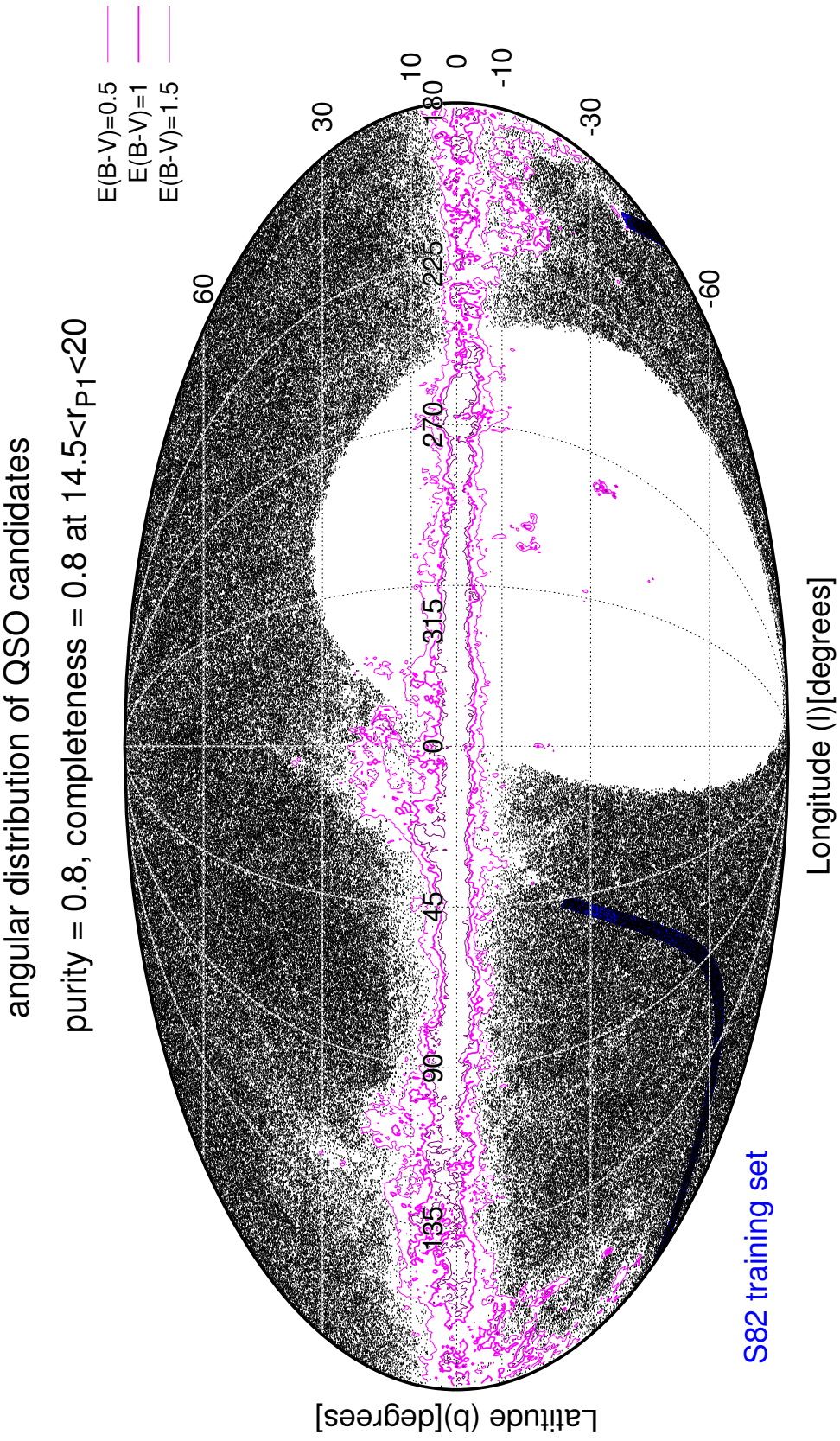


Figure 5.22 Angular distribution of the 3.7×10^5 likely QSO candidates ($0.56 \leq p_{\rm QSO}$, purity=0.8, completeness=0.8), shown in Mollweide projection in Galactic coordinates. A contour plot of the reddening-based $E(B-V)$ dust map (Schlafly et al. 2014) is overlayed.

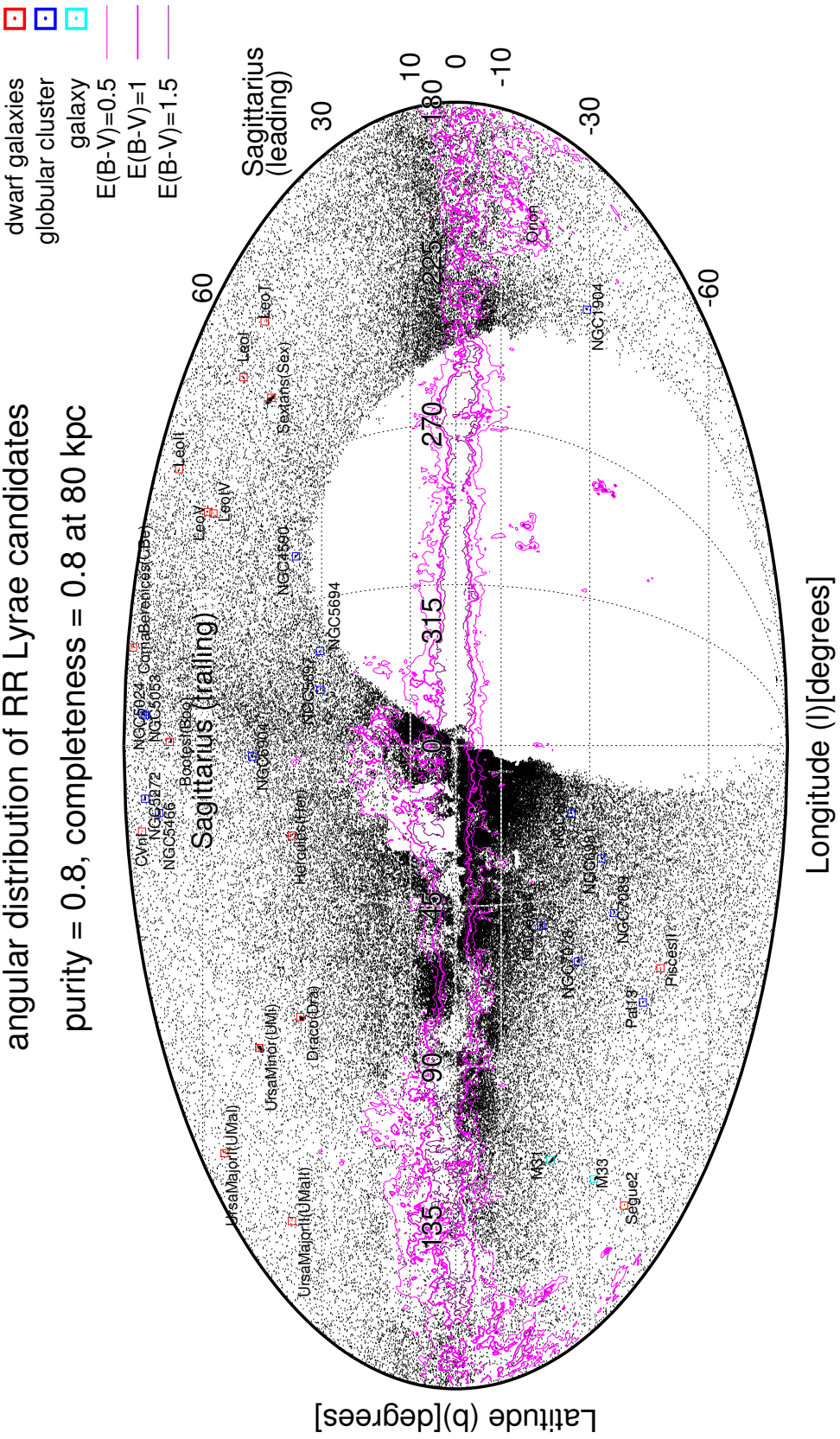


Figure 5.23 Angular distribution of the 1.5×10^5 likely QSO candidates ($0.27 \leq p_{\rm QSO}$, purity=0.8, completeness=0.8), shown in Mollweide projection of Galactic coordinates. A contour plot of the reddening-based $E(B-V)$ dust map (Schlafly et al. 2014) is overlayed, as well as identified known objects of the Milky Way spheroid substructure and its neighborhood.

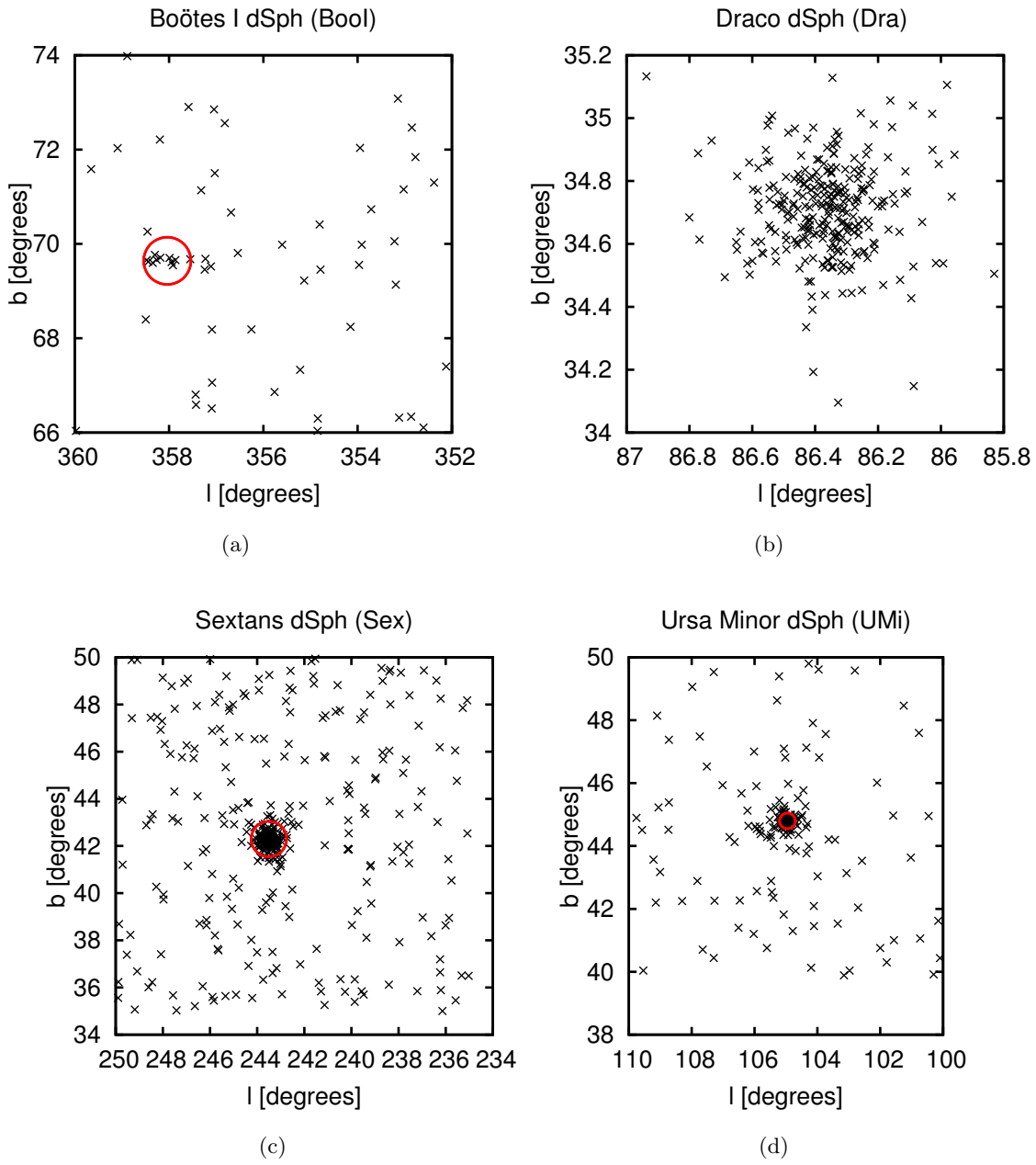


Figure 5.24 Dwarf spheroidals visible in the RR Lyrae candidate sample; central coordinates are from SIMBAD, and the apparent sizes from the paper given for each dwarf spheroidal. A description of these sources is given in Sec. 5.5.3.

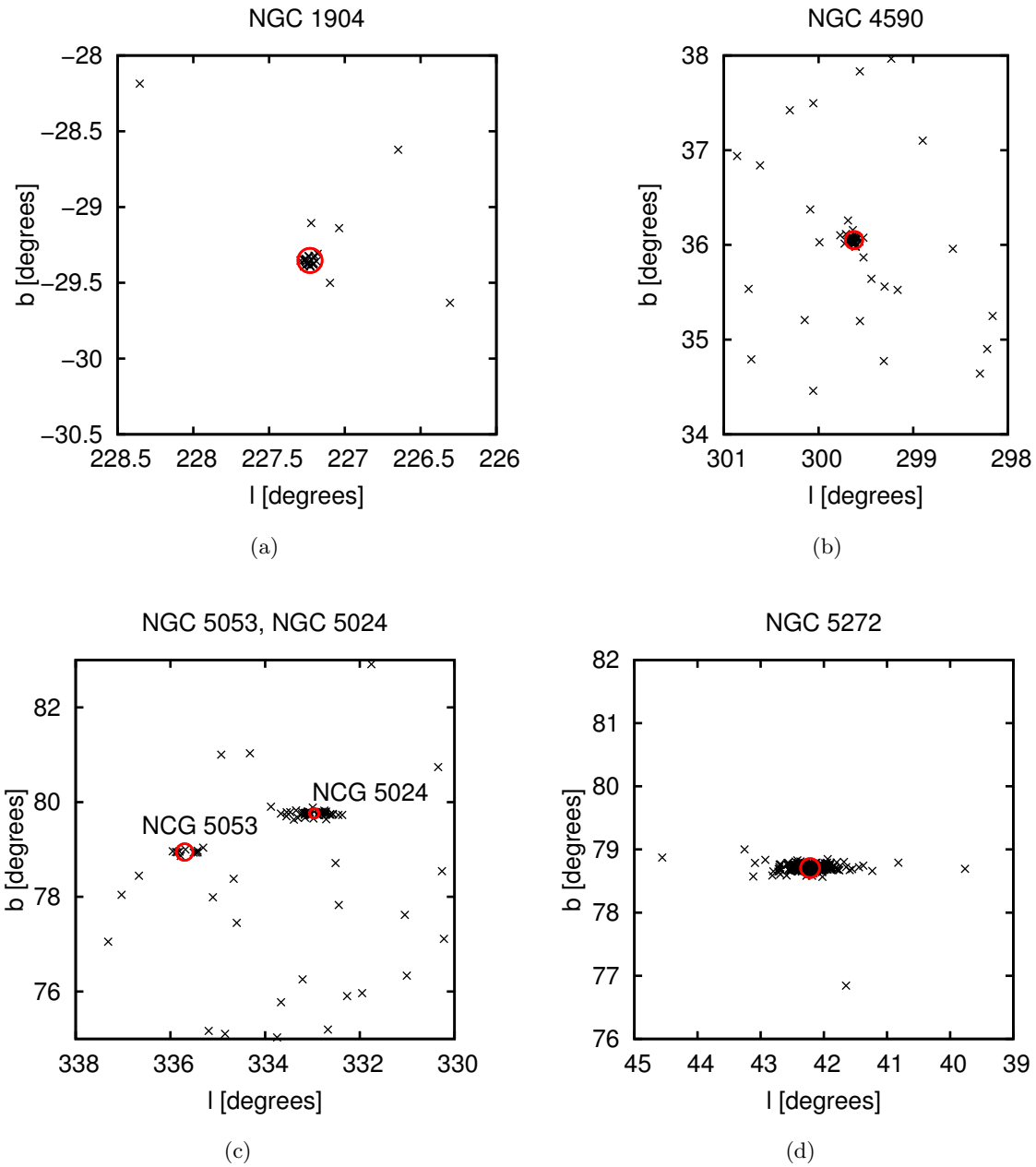


Figure 5.25 Globular clusters (I) visible in the RR Lyrae candidate sample; central coordinates are from SIMBAD, and the apparent sizes (indicated as red circles) from the paper given for each globular cluster. A description of these sources is given in Sec. 5.5.3. The seemingly strong features in NGC 5053, NGC 5024, NGC 5272 have not yet been verified.

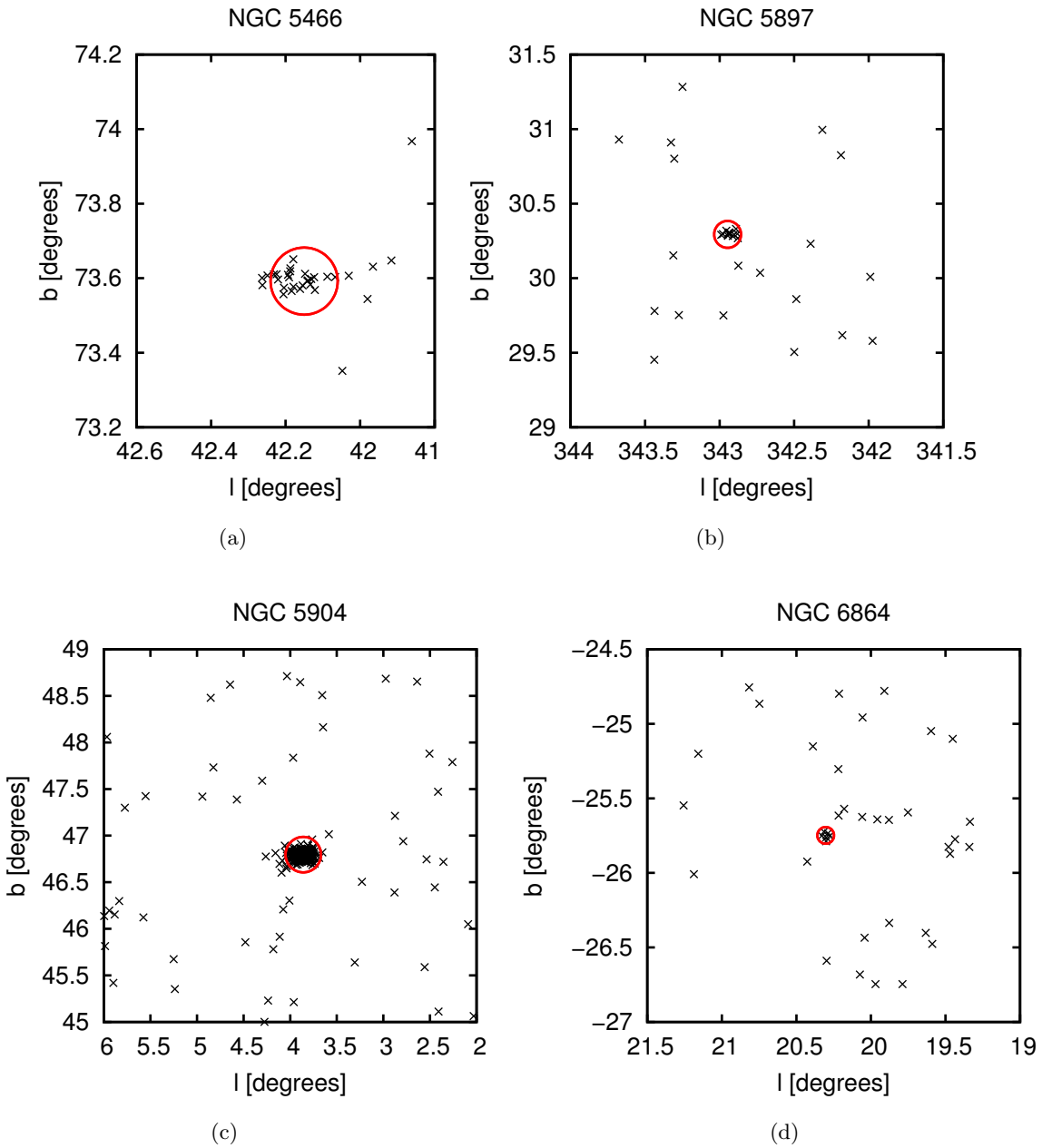


Figure 5.26 Globular clusters (II) visible in the RR Lyrae candidate sample; central coordinates are from SIMBAD, and the apparent sizes (indicated as red circles) from the paper given for each globular cluster. A description of these sources is given in Sec. 5.5.3.

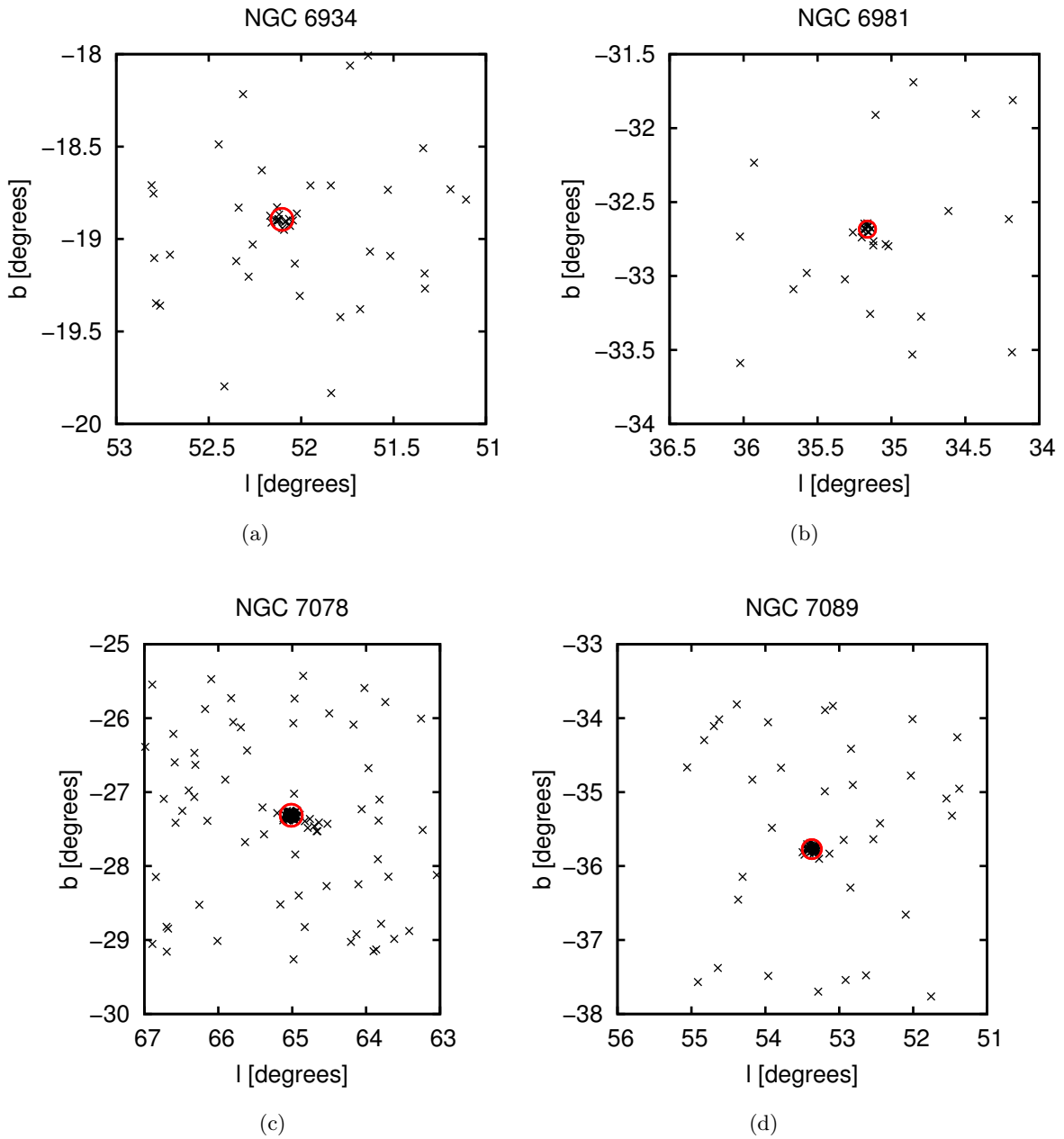


Figure 5.27 Globular clusters (III) visible in the RR Lyrae candidate sample; central coordinates are from SIMBAD, and the apparent sizes (indicated as red circles) from the paper given for each globular cluster. A description of these sources is given in Sec. 5.5.3.

Chapter 6

The Geometry of Sagittarius Stream

Stellar streams are of great interest as their orbits are sensitive tracers of the Milky Way's gravitational potential. As stellar streams as tracers of the gravitationally potential trace the mass, and thus the potential, enclosed in the orbit, it is crucial to have streams at large distances. Methods to constrain the mass by inferring the stream's progenitor orbit have mostly been carried out for closer systems, so for the GD-1 stream being only ∼15 kpc from the Galactic center (Koposov et al. 2010). However, there are numerous attempts on carrying this out for the Sagittarius tidal stream (e.g. Law et al. 2005; Peñarrubia et al. 2010; Gibbons et al. 2014).

The Sagittarius tidal stream is of special interest as of all known tidal streams belonging to the Milky Way, as it reaches out to more than 100 kpc and is the only stream that shows two nearly complete orbital loops, one called the "leading arm" and the other called the "trailing arm".

The main aim of this Chapter is to develop, test, apply and discuss a method to map the geometry of the Sagittarius (Sgr) stream. In the past, a few attempts have been carried out using N-body simulations constrained by observational data (e.g. Fellhauer et al. 2006; Law and Majewski 2010). Such methods suffer from the drawback that they are very computationally expensive. As having a reasonable high number of sources available by using the highly pure RR Lyrae candidate sample with precise distances as described in Sec. 5.6.2, the geometry of the Sgr stream can be directly fitted by a density model.

The structure of the Chapter is as follows: First, the data selected for the fitting are briefly described. In the methodology section, the underlying halo and stream model, as the fitting method are described. Results are given, analyzed and compared to previous publications. The Chapter concludes with a discussion of the results.

6.1 Data

In order to map the geometry of the Sgr stream, the sample of likely RR Lyrae candidates obtained after period fitting, as described in Sec. 5.6.2, is used. these sample has a purity of 0.9, completeness of 0.8 at 80 kpc. Heliocentric distance estimates come from the method described in Sec. 5.6.2.
For describing the angular and distance distribution of these sources, the heliocentric Sagittarius

coordinates $(\tilde{\Lambda}_\odot, \tilde{B}_\odot)$ as defined by Belokurov et al. (2014) are used. In this coordinate system, the equator $\tilde{B}_\odot = 0°$ is aligned with the plane of the stream. The source sample is then restricted to the sources within $|\tilde{B}_\odot| < 9°$ as also seen in the plots by Belokurov et al. (2014). This sample is plotted in Fig. 6.1, where the angular distance to the Sgr plane is indicated by color-coding.

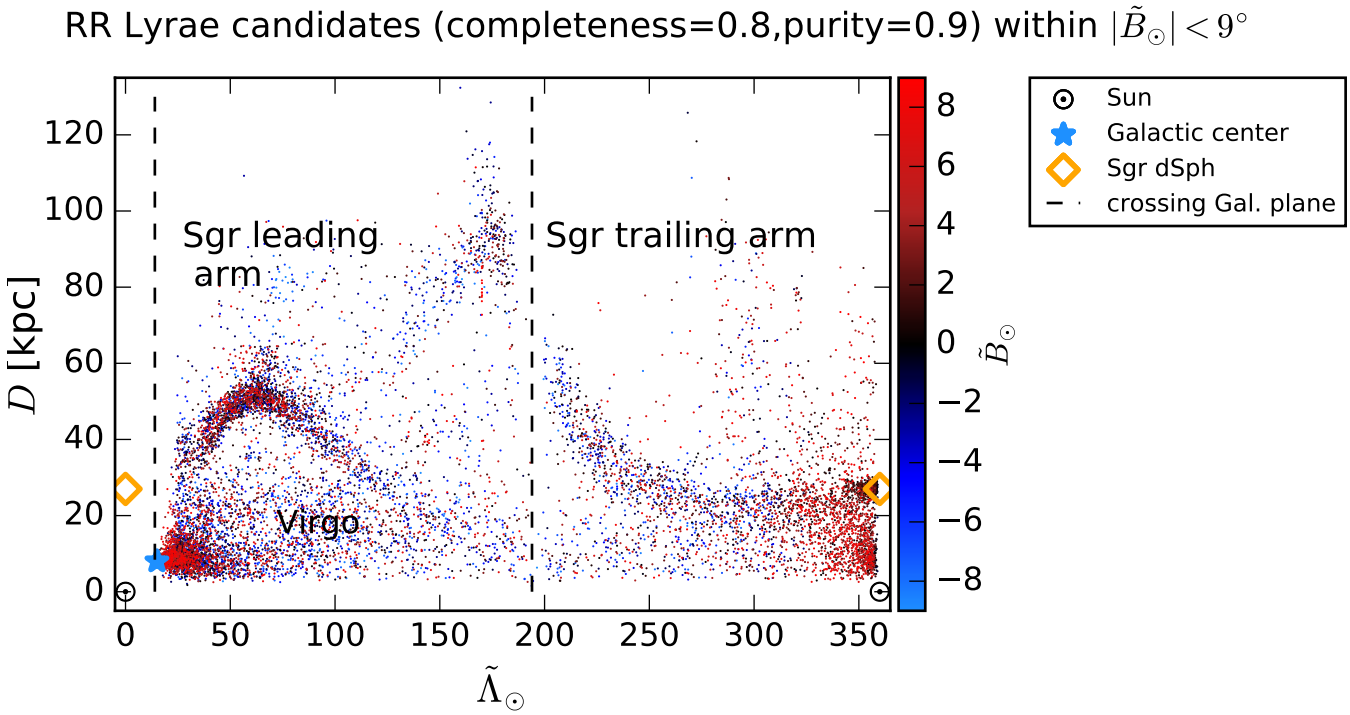


Figure 6.1 RR Lyrae candidates within $|\tilde{B}_\odot| < 9°$ as obtained after period fitting (see Sec. 5.6.2). The Sgr stream is clearly visible up to ∼130 kpc. The angular distance to the Sgr plane $\tilde{B}_\odot = 0°$ is indicated by color-coding.

Compared to the RR Lyrae candidate sample obtained without period fitting (see Fig. 5.17 in Section 5.5.3), the sample obtained after period fitting traces the Sgr stream with better contrast, thanks to higher purity of 0.9 compared to 0.8 before. Sources are found out to a heliocentric distance of more than 130 kpc.

Belokurov et al. (2014) have demonstrated that the trailing arm of the Sgr stream can be traced out to its apocenter at ∼100 kpc. Belokurov et al. (2014) give also a trace of the stream's leading arm to its apocenter at ∼50 kpc. The enormous extent of the Sgr stream has therefore only recently became apparent. It spans a huge range of distances, unparalleled when compared to other debris belonging to the Milky Way.

6.2 Methodology

In the following, the methodology for tracing the Sgr stream is described. Also, information on the test of the fitting method is given.

6.2.1 The Model

The distribution of RR Lyrae candidates is modeled as a composition of a power-law halo model $\rho_{\rm halo}$ in Galactic coordinates describing the background, and a Gaussian describing the heliocentric distance $D_{\rm sgr}$ and width $\sigma_{\rm sgr}$ of the stream. Data will be fitted in $\tilde{\Lambda}_\odot$ slices. This leads to the following model for the observed distances D:

$$\begin{aligned}\hat{p}(D|\vec{p}) =&(1-f_{\rm sgr})\frac{\rho_{\rm halo}(l,b,D,q,n)}{\int\limits_{D_{min}}^{D_{max}}\rho_{\rm halo}(l,b,D,q,n)\mathrm{d}D}\\ &+f_{\rm sgr}\frac{\rho_{\rm sgr}(l,b,D,D_{\rm sgr},\sigma_{\rm sgr})}{\int\limits_{D_{min}}^{D_{max}}\rho_{\rm sgr}(l,b,D,D_{\rm sgr},\sigma_{\rm sgr})\mathrm{d}D}\end{aligned} \tag{6.1}$$

with the parameter set $\vec{p}=(f_{\rm sgr},D_{\rm sgr},\sigma_{\rm sgr},n)$, composed of the fraction of the stars $f_{\rm sgr}$ being in the Sgr stream at the given $\tilde{\Lambda}_\odot$ slice, the heliocentric distance of the stream $D_{\rm sgr}$, its line-of-sight width $\sigma_{\rm sgr}$, and the power-law index n of the halo model.

The underlying halo model $\rho_{\rm halo}$ (Sesar et al. 2013b) is defined in Galactic coordinates (l,b):

$$\rho_{\rm halo}(X,Y,Z)=\rho_{\odot\rm RRL}\left(R_\odot/r\right)^n \tag{6.2}$$

with

$$\begin{aligned}X&=R_\odot-D\cos l\cos b\\ Y&=-D\sin l\cos b\\ Z&=D\sin b\\ r&=\sqrt{X^2+Y^2+(Z/q)^2}\\ n&=2.62\\ R_\odot&=8.0\ \mathrm{kpc}\\ q&=0.71\\ \rho_{\odot\rm RRL}&=4.5\ \mathrm{kpc}^{-1}.\end{aligned}$$

$\rho_{\odot\rm RRL}$ is the number density of RR Lyrae at the position of the Sun, q gives the halo flattening.

The underlying stream model is a Gaussian, defined in Galactic coordinates (l,b) and Galactocentric distance r, where r is given as function of the heliocentric distances D, $D_{\rm sgr}$,

$$\rho_{\rm sgr}(l,b,D,D_{\rm sgr},\sigma_{\rm sgr})=\frac{1}{\sqrt{2\pi}\sigma_{\rm sgr}}\exp\left(-\frac{(r(D)-r(D_{\rm sgr}))^2}{2\sigma_{\rm sgr}^2}\right)D^2. \tag{6.3}$$

6.2.2 The Fitting Method

For fitting, the data is splitted in slices of $\tilde{\Lambda}_\odot$, each 10° wide. Data is not binned in D.
The halo power law index n is fitted individually for each $\tilde{\Lambda}_\odot$ slice to account for incompleteness of the data, the flattening parameter q is kept fixed at 0.71.

The likelihood of the model given the data is then estimated as

$$\mathcal{L}(\vec{p}|D_{\mathrm{obs},i}) = \sum_i \ln \hat{p}(D_{\mathrm{obs},i}|\vec{p}) + \ln prior \tag{6.4}$$

with

$$\ln \mathrm{prior} = \begin{cases} 1, & \text{if} \quad 0 < f_{\mathrm{sgr}} < 1 \\ & \quad\ 1.70 < n < 10 \\ & \quad\ 1 < \sigma_{\mathrm{sgr}}[\mathrm{kpc}] < 6 \\ & \quad\ D_{\mathrm{minprior}} < D_{\mathrm{sgr}} < D_{\mathrm{maxprior}} \\ -\infty, & \text{else} \end{cases} \tag{6.5}$$

with D_{minprior}, D_{maxprior} indicated in Fig. 6.2.

D_{minprior}, D_{maxprior} are basically constrained by the minimum and maximum distance in the $\tilde{\Lambda}_\odot$ slice in case, but are also set in order to mask dense regions at low heliocentric distances as well as to separate the leading and trailing arm where both are present at the same line of sight. The prior given by D_{minprior}, D_{maxprior} is indicated in Fig. 6.2.

The likelihood of the model given the data is explored using the Affine Invariant Markov chain Monte Carlo (MCMC) ensemble sampler (Goodman and Weare 2010) as implemented in the `emcee` package (Foreman-Mackey et al. 2012). For a description of the algorithm, see Section A.2.3.

To gain confidence in any inferences obtained from the fitting method, it was tested on mock data using a mock halo sampled from the underlying halo model, superimposed by a mock stream inserted as a stellar density sheet, whose number density is uniform perpendicular to the line of sight, and Gaussian along the line of sight. The fraction of the stream stars w.r.t. the halo stars, described by f_{sgr}, was subsequently lowered. Also, the fit was carried out in the limit of many and few stars in each $\tilde{\Lambda}_\odot$ slice to make sure that reasonable fits can be obtained for densities like the ones present for the PS1 3π RR Lyrae candidates.

6.3 Results

The model given in Sec. 6.2.1 was then applied to the complete sample of candidates within $|\tilde{B}_\odot| < 9°$. In Fig. 6.2, the extent of the Sgr stream as well its fitted distance and width are

depicted. Here, the extent of the Sgr stream is traced in both its leading and trailing arm by $D_{\rm sgr}$, shown as black points centered on the $\tilde{\Lambda}_\odot$ slice in case. Its line-of-sight width $\sigma_{\rm sgr}$ is indicated by black bars. The grey areas mark the priors set on $D_{\rm sgr}$.

The fitted parameters are given in Tab. B.6 and B.7 in the Table Appendix.

It is clearly visible that the distance and width estimates trace the stream well all the way out to more than 100 kpc, with individual sources tracing it even out to more than 130 kpc. From this detailed picture of the Sgr stream, many features can be seen in great detail, some of them reported previously.

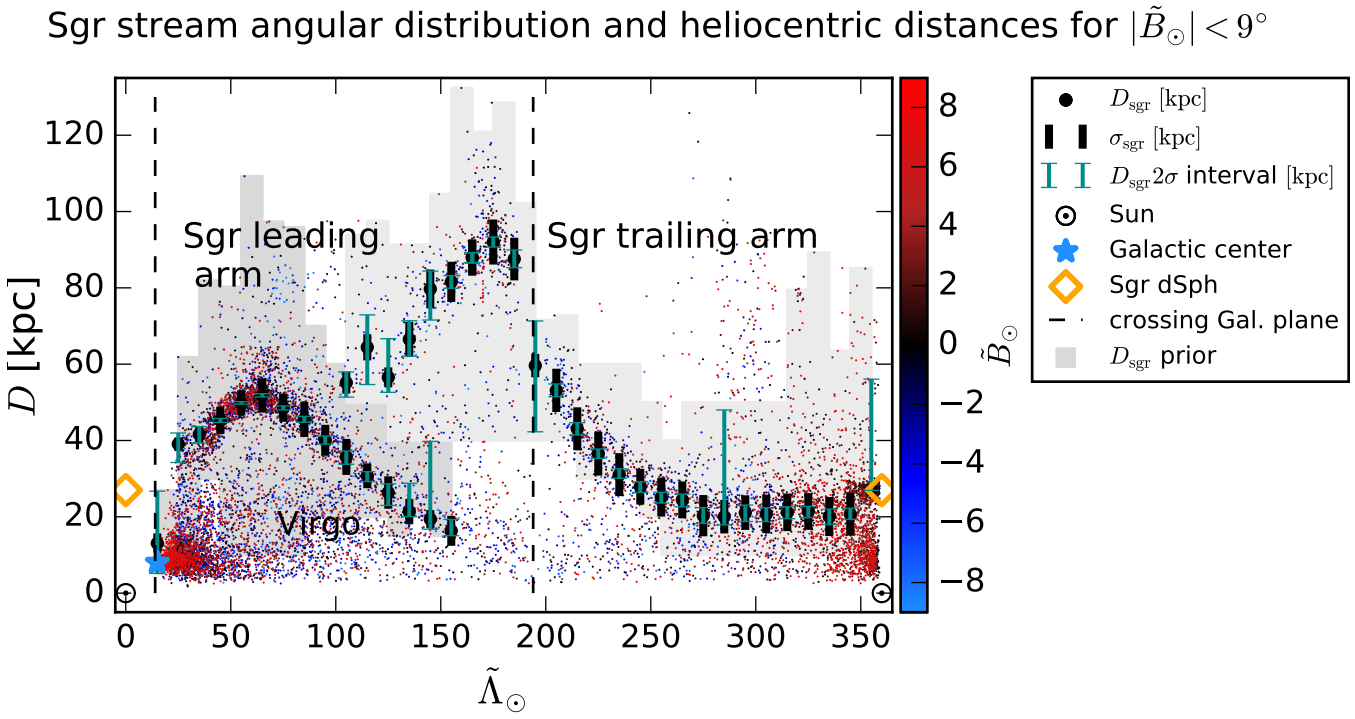
(i) The stream shows clearly distinct leading and trailing arms. The shape and extent looks similar to those found by Belokurov et al. (2014), see also Section 6.3.4.

(ii) The stream turns over at $\tilde{\Lambda}_\odot{\sim}70°$ and $\tilde{\Lambda}_\odot{\sim}170°$. These are the leading and trailing apocenters, according to Belokurov et al. (2014), Fig. 6 therein.

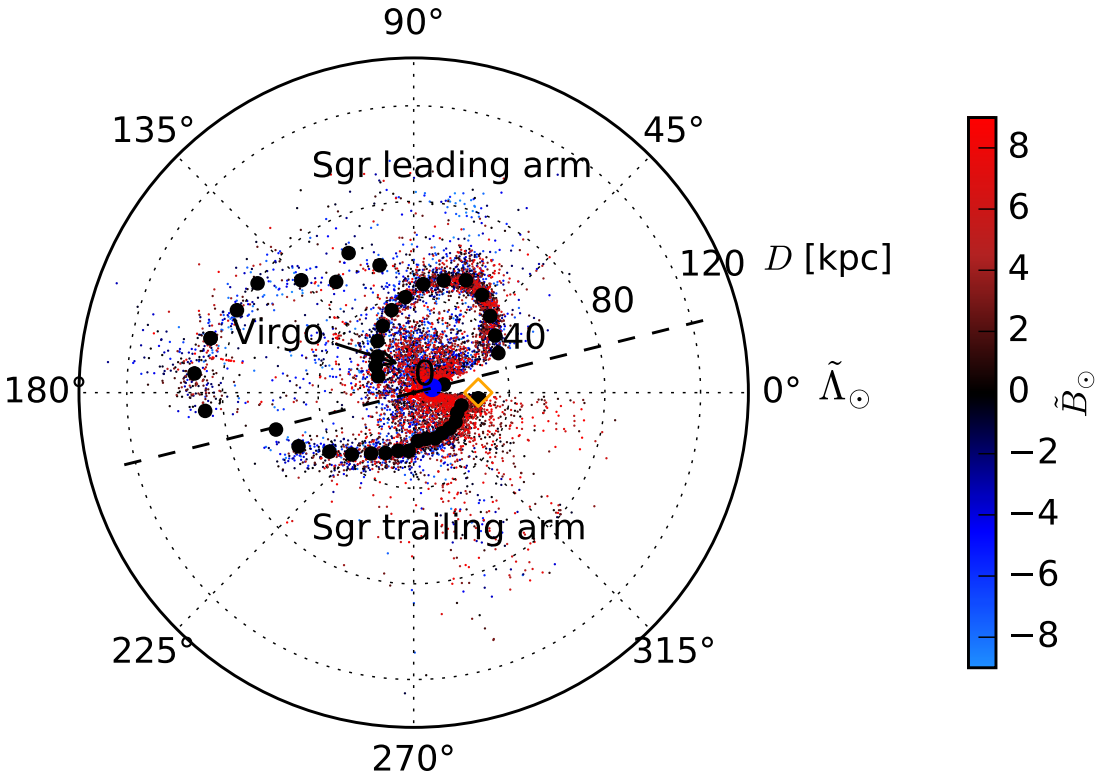
(iii) The overdensity found at $\tilde{\Lambda}_\odot \sim 95°$ and $\tilde{\Lambda}_\odot \sim 180°$, which was previously reported by Sesar (2012) as moving group, and by Drake et al. (2013a) as new stream, can now clearly be associated with the distant part of the trailing arm.

(iv) An apparent continuation of the stream at $\tilde{\Lambda}_\odot \sim 180\,\mathrm{deg}$, reaching up to 130 kpc, is clearly visible. This feature was previously predicted by Gibbons et al. (2014) from dynamical models of the stream. According to them, this feature of the stream is debris from the most recent orbital passage. The Gibbons et al. (2014) model also explains the two moving groups found by Sesar (2012) and Drake et al. (2013a) as debris from the oldest and from the most recent stripping epochs. This is discussed in Sesar, Hernitschek et al. (2016) in detail. Finding this feature in observational data is a new discovery, possible thanks to the wide and deep view of the Galactic halo possible through PS1 3π, and the precise RR Lyrae candidate selection possible through methods like shown in Section.

In the following, the results are laid out in greater detail and compared to previous distance estimates by Belokurov et al. (2014).



(a)



(b)

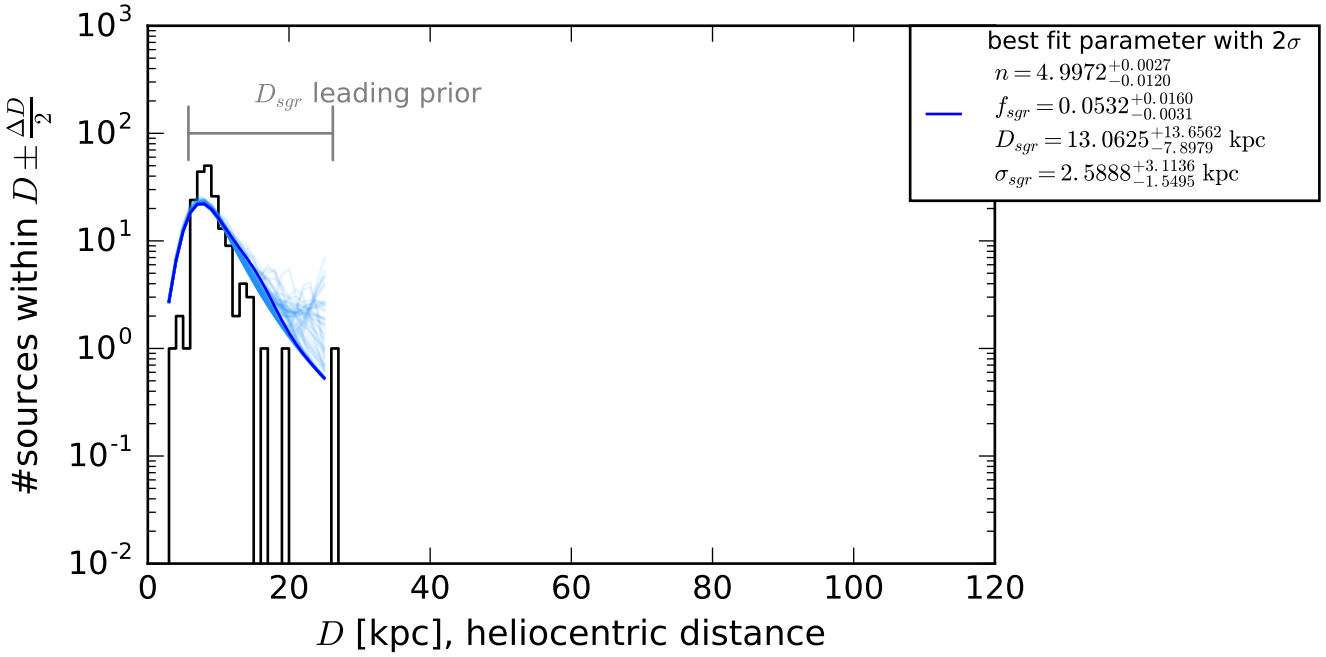
Figure 6.2 (a) The extent of the Sagittarius stream from the RR Lyrae candidates within ±9° of the Sagittarius plane, shown in Sagittarius coordinates from Belokurov et al. (2014). The best fit model, obtained for 10° slices in $\tilde{\Lambda}_\odot$, is overplotted. The black points indicate the center of the $\tilde{\Lambda}_\odot$ slices used to estimate the distance $D_{\rm sgr}$. (b) Alternative cylindric projection.

6.3.1 Fits to Individual $\tilde{\Lambda}_\odot$ Slices

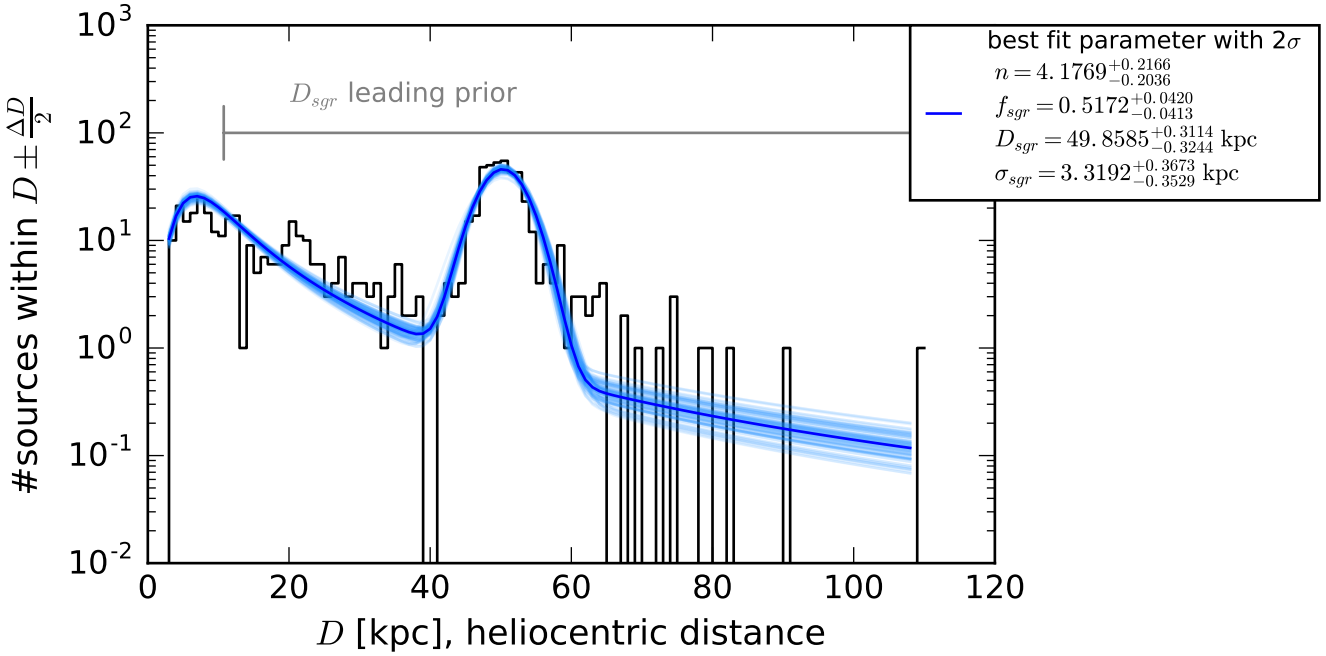
Each distance and width estimate $(D_{\rm sgr}, \sigma_{\rm sgr})$ in Fig 6.2 is obtained by optimizing Equ. (6.4) using a MCMC.
The following Figures 6.3 and 6.4 show these fits to individual slices in $\tilde{\Lambda}_\odot$. Here, also the previously mentioned issue with sample incompleteness becomes obvious.

Fig. 6.3 gives the fits for a 10° wide slice centered on $\tilde{\Lambda}_\odot = 10°$ and $\tilde{\Lambda}_\odot = 50°$, respectively. In these directions, only the leading arm is present. The plot indicates the prior on $D_{\rm sgr}$, in these cases, only set by the minimum and maximum distance available from sources in the $\tilde{\Lambda}_\odot$ slice in case. The distribution of the sources is shown, overplotted with the model from the best-fit parameters given as a solid blue line. The spread of transparent blue lines gives the spread of models within the 2σ range obtained by the MCMC.
In both cases, a halo profile much steeper than the $n = 2.62$ from the Sesar et al. (2013b) model is obvious. For $\tilde{\Lambda}_\odot = 10°$ (a), n reaches even the upper limit set by the prior. This is caused by sample incompleteness, leading the MCMC to choose a steeper profile even if the distribution of more distant sources indicates a flatter.
The estimate of $D_{\rm sgr}$ and $\sigma_{\rm sgr}$ is clearly seen as being sensible in Fig. 6.3 (b). Here, the 2σ range on the estimated parameters is very small, and the parameters fit well to what one would guess by visual inspection.
Even in Fig. 6.3 (a), where the 2σ range becomes significantly larger, a sensible estimate is found that fits well in the picture, see Fig. 6.2.

Fig. 6.4 gives the fits for $\tilde{\Lambda}_\odot = 150°$, where both leading and trailing arm are in the line of sight. Using distinct priors on $D_{\rm sgr}$, separates both debris and gives precise estimates on distance and width of both leading and trailing arm (see also Fig. 6.2 around $\tilde{\Lambda}_\odot = 150°$). This illustrates the importance of carefully set priors.



(a) $\tilde{\Lambda}_\odot = 10°$



(b) $\tilde{\Lambda}_\odot = 50°$

Figure 6.3 Combined halo and stream fit for a 10° wide slice centered on $\tilde{\Lambda}_\odot = 10°$ and $\tilde{\Lambda}_\odot = 50°$, respectively. At the $\tilde{\Lambda}$ shown here, only the leading arm of the Sgr stream is present.
The source distance distribution is shown, overplotted with the model from the best-fit parameters given as solid blue line. The spread of transparent blue lines gives the spread of models within the 2σ range obtained by the MCMC. The plots indicate the prior on D_{sgr}, in these cases, only set by the minimum and maximum distance available from sources in the $\tilde{\Lambda}_\odot$ slice in case.

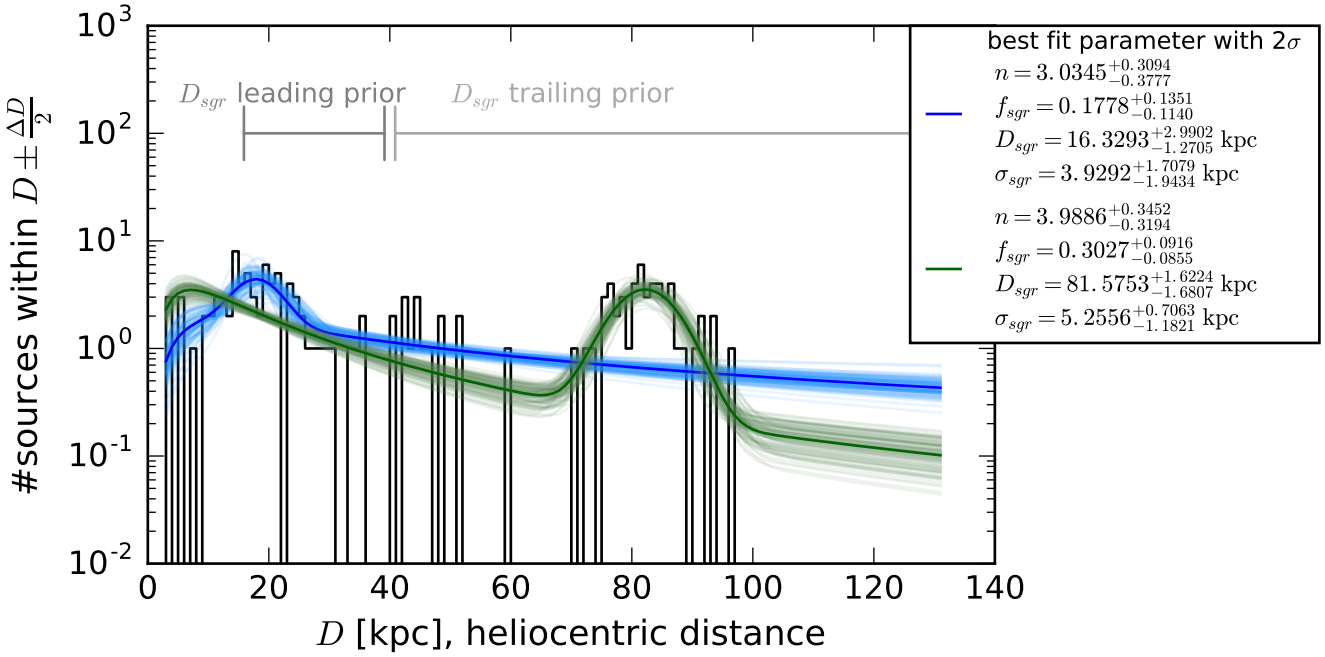


Figure 6.4 Combined halo and stream fit for a 10° wide slice centered on $\tilde{\Lambda}_\odot = 150°$ where both the leading and trailing arm of the Sgr stream are present. The figure is similar to Fig. 6.3, but showing the influence of a carefully chosen prior to separate both debris. Using distinct priors on $D_{\rm sgr}$, precise estimates on distance and width of both leading and trailing arm are possible.

6.3.2 The Width of Sagittarius Stream

Fig. 6.5 shows the estimated line-of-sight width $\sigma_{\rm sgr}$ of the stream, vs. $\tilde{\Lambda}_\odot$, given for the leading and trailing arm. The 2σ interval is indicated, as well as the prior on $\sigma_{\rm sgr}$, set to $1 < \sigma_{\rm sgr}$ [kpc] < 6. As also visible from Fig. 6.2(a), but now more obvious, the stream tends to broaden along its orbit from ∼1.75 kpc to 5 kpc for the leading arm, or 6 kpc (reaching the upper limit set by the prior) for the trailing arm, reaching its largest width close to the apocenters. However, it is imporant to note the large 2σ range.
Belokurov et al. (2014) give the leading tail's apocenter at $71°.3 \pm 3°.3$ and the trailing tail's apocenter at $170°.5 \pm 1°$ From the fit in 10° wide slices in $\tilde{\Lambda}_\odot$, the leading tail's apocenter can be estimated as being between $\tilde{\Lambda}_\odot = 60°$ and 70° where $D_{\rm sgr}$ reaches its largest extent of $48.5 - 49.6$ kpc, and the trailing tail's apocenter being at $\tilde{\Lambda}_\odot$∼170° reaching its largest extent of 92.0 kpc. Except towards the apocenters, $\sigma_{\rm sgr}$ raises also towards the "end" (the largest $\tilde{\Lambda}_\odot$) of the respective trailing or leading arm.

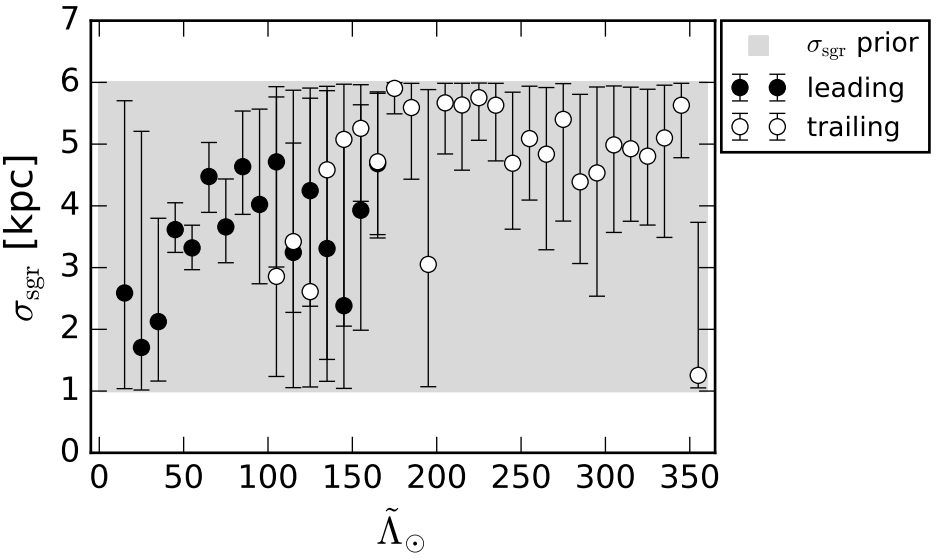


Figure 6.5 The width $\sigma_{\rm sgr}$ of the Sagittarius stream from the RR Lyrae candidates within ±9° of the Sagittarius plane. Error-bars indicate the 2σ range of the $\sigma_{\rm sgr}$ estimates. The grey area indicates the prior on $\sigma_{\rm sgr}$, set to $1 < \sigma_{\rm sgr}$ [kpc] < 6. A trend in the width can be seen, reaching maximum around the apocenters and towards the largest $\tilde{\Lambda}_\odot$ of each the leading and trailing arm, respectively.

6.3.3 Bifurcation

Belokurov et al. (2006) used a color cut to select the upper main-sequence and turnoff stars belonging to the Sgr stream. By doing so, they found a branching of the stream in the Galactic northern hemisphere, called the *bifurcation*. Starting at $\alpha \sim 190°$, the lower and upper declination branches of the stream, labeled A and B respectively, can be traced at least until $\alpha \sim 140°$. As stated by Fellhauer et al. (2006), the bifurcation likely arises from different stripping epochs, the

young leading arm providing branch A and the old trailing arm branch B of the bifurcation. Belokurov et al. (2006) states that branch B is significantly brighter and hence probably slightly closer than A. Their Fig. 4 shows a noticeable, but small difference in the distances estimated for branches A and B of 3 to 15 kpc. Also simulations by Fellhauer et al. (2006) find branch B being closer than branch A.

This brought up the question if this distance difference can also be found from the PS1 3π RR Lyrae candidate sample. For doing so, distance estimates were done for small patches on both branches, as shown by the polygons in Fig. 6.6. Belokurov et al. (2006) used a similar approach for their sources selected from SDSS Data Release 5, but with smaller and rectangle-shaped patches. Each patch was then fitted by the halo and stream model as described above in Section 6.2.1 in order to derive distance estimates.
The fitting led to the distance estimates as shown in Fig. 6.6 and in Tab. B.2 in the Table Appendix. Indeed a small distance difference between the two branches can be found, branch B being closer than branch A. However, because of the large distance uncertainty (see 2σ range indicated in Fig. 6.6 and in the table) this is not a significant result indicating such a distance difference between both branches.

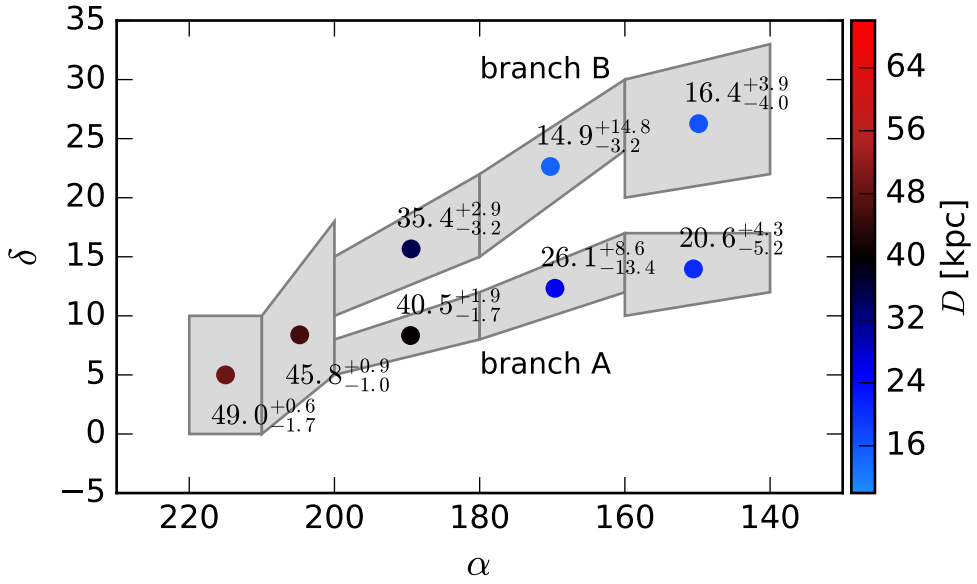


Figure 6.6 Heliocentric distance estimates for patches covering the branches A and B of the Sagittarius stream in equatorial coordinates (α, δ). For each patch, the fit using the halo and stream model as described above in Section 6.2.1 was carried out derive distance estimates. The points set at the centroid of each polygon indicate the heliocentric distance D in kpc as estimated from the sample within each polygon. The 2σ range is indicated.

6.3.4 Comparison to the Model by Belokurov et al. (2014)

The best estimate of the heliocentric distances for a large part of the Sgr stream obtained so far come from Belokurov et al. (2014). In Fig. 6.7, the obtained heliocentric distances from Belokurov

et al. (2014) (Figure 6 therein) are shown together with the $D_{\rm sgr}$ obtained within the work at hand.

They are in good agreement, however, Belokurov et al. (2014) does not trace the complete stream, and they don't give estimates on its width. Also, the distances from Belokurov et al. (2014) show a slight trend towards larger values.

As Belokurov et al. (2014) show, the apocenter of the leading trail is placed at ∼50 kpc and the trailing debris are revealed to reach out to ∼100 kpc from the Galactic center. The opening angle between the positions of the two apocenters, as viewed from Galactic center, is measured by Belokurov et al. (2014) to be $99°.3 \pm 3°.5$. From the tracing done in this work using the PS1 3π RR Lyrae candidates, the leading tail's apocenter can be estimated as being between $\tilde{\Lambda}_\odot = 60°$ and 70° where $D_{\rm sgr}$ reaches its largest extent of $48.5 - 49.6$ kpc, and the trailing tail's apocenter being at $\tilde{\Lambda}_\odot \sim 170°$ reaching its largest extent of 92.0 kpc. Keeping in mind that the analysis here was carried out using 10° wide slices in $\tilde{\Lambda}_\odot$, this is in good agreement.

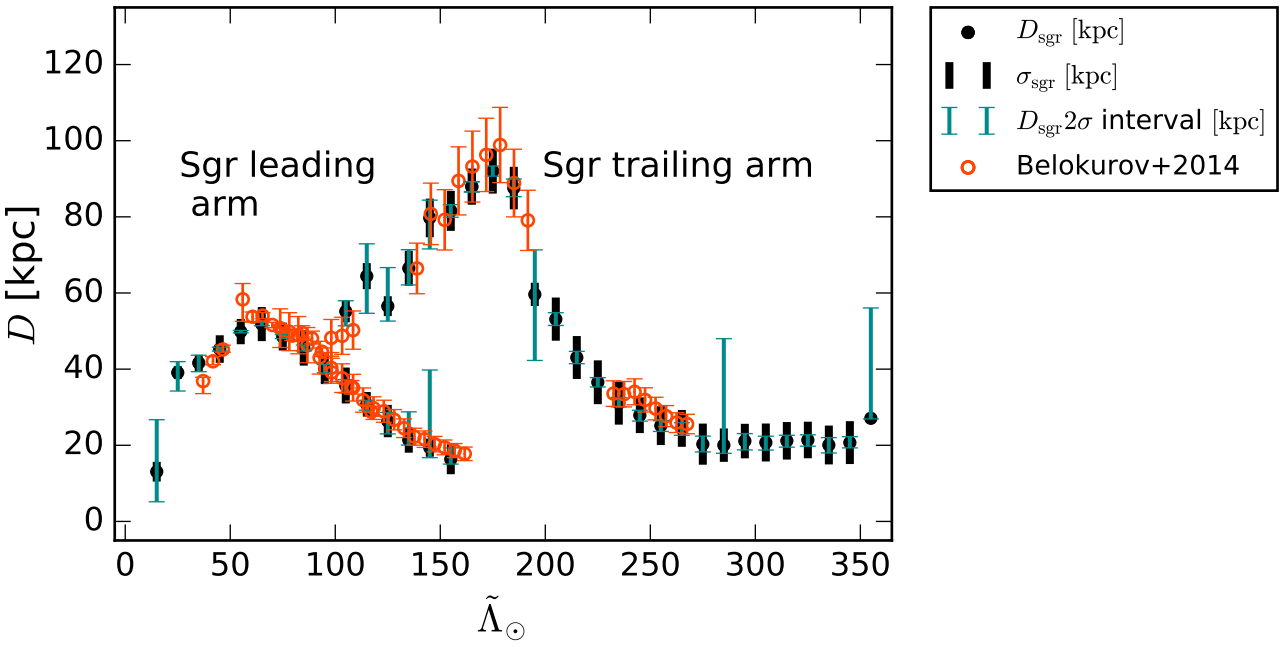


Figure 6.7 Comparison of the heliocentric distance estimates of the Sgr stream between this work and Belokurov et al. (2014). The $D_{\rm sgr}$, shown as black points together with their 2σ uncertainties (blue-green bars) and estimated stream width $\sigma_{\rm sgr}$, are compared to the estimates from Belokurov et al. (2014) (orange points) with their uncertainties. The distances from Belokurov et al. (2014) show a slight trend towards larger values. Over all, the distance estimates are in good agreement.

6.4 Discussion

In this Chapter, a method for fitting the geometry of the Sagittarius stream was developed, assuming a halo model as given by Sesar et al. (2013b) (shown here in Equ. (6.1)) and the line-of-sight density of the Sagittarius stream approximated by a Gaussian distribution centered on the distance $D_{\rm sgr}$, having the line-of-sight width $\sigma_{\rm sgr}$. This model was used to estimate distance and width of the Sgr stream as given by RR Lyrae candidates (completeness=0.8, purity=0.9, distance precision of 3%) resulting from the classification that incorporates period fitting.

The fitting resulted into the first complete (i.e., spanning $0^\circ < \tilde{\Lambda}_\odot < 360^\circ$) trace of Sgr stream's heliocentric distance, as well as line-of-sight width. Besides the distinct trace itself, another important finding was the discovery of the continuation of Sgr stream out to more than 120 kpc near its trailing apocenter. This confirms the simulations done by Gibbons et al. (2014).

Having now a model of the geometry of the Sgr stream at hand, it can be used to further constrain the Milky Way's potential.

Chapter 7

Summary and Discussion

In this thesis, techniques for identifying, characterizing and classifying astronomical sources from multi-band light curves were developed and applied. The methods were then applied to the specific case of light curves from Pan-STARRS1 3π, which are very sparse with about 65 observations (distributed among 5 bands) spread over a timeline of roughly 4 years.

In order to characterize variability, the approach fits multi-band, sparse light curves to extract features for subsequent machine learning. Structure function fitting was generalized to the fitting of structure functions for non-simultaneous, multi-band light curves; this new methodology was implemented, carefully tested and applied. This allows to assign to each of 1.1×10^9 point sources in PS1 3π a set of variability-based features, mainly a variability timescale and a variability amplitude, together with mean magnitudes.
The obtained *features* were then used to train and apply a supervised machine-learning classifier, analyzing PS1 3π data in the SDSS Stripe 82, where there is (presumably) complete identification of RR Lyrae stars and QSO. This made it possible to identify highly pure and complete samples of RR Lyrae and QSOs throughout PS1 3π; first tests show also the promising possibility of finding Cepheids in the Milky Way's disk.

In total, a sample of 1.5×10^5 likely RR Lyrae candidates in PS1 3π were identified, for which – based on SDSS S82 tests – a purity and completeness of each 0.8 for sources at a distance of 80 kpc (and a higher for closer sources) can be expected. Furthermore, a sample of 3.7×10^5 likely QSOs at the same level of purity and completeness expected for sources within $14.5 < r_{\mathrm{P1}} < 20$ is obtained.
The selection of candidates is homogeneous across the survey to a high degree away from the Galactic plane. Near the plane, the number density of highly likely RR Lyrae as well as QSO candidates decreases because of dust and source crowding. A projection of the RR Lyrae candidate sample into the orbital plane of the Sagittarius stream reveals the stream morphology clearly.

Optimal variable source classification was carried out with the just-described variability features, colors from PS1 3π and WISE colors. But also classification with more restricted pieces of information, only color related or only variability-related features was carried out. This reveals that the variability information is absolutely indispensable to define a sample of RR Lyrae or QSO with an interesting combination of purity and completeness. Furthermore it shows what one can

expect for purity and completeness if sources lack specific pieces of information.

Resulting RR Lyrae and QSO samples

Across the entire 3π, 1.5×10^5likely RR Lyrae candidates were identified. Based on the training in S82, a purity (under circumstances comparable to S82) of 0.8, and completeness of 0.8 is expected. As mentioned above, these numbers on purity and completeness only apply away from the Galactic plane, and the bulge. Among them, at $|b| > 20°$, 4.8×10^4 candidates are identified, and 9.0×10^4 with a completeness of 0.88, purity of 0.52. The sample within the Galactic halo extends to distances as large as ∼140 kpc. The selection of candidates is distributed according to expectations to a high degree away from the Galactic plane, showing the structured Milky Way halo in great detail. Around the plane, the number density of highly likely RR Lyrae candidates drops, caused by dust and source crowding.
Furthermore, 3.7×10^5 likely QSO candidates were identified over the total PS1 3π area at the same level of purity and completeness, 0.8. The QSO selection of candidates is isotropic to a high degree away from the Galactic plane.

One important limitation of the classification is that it relies on SDSS Stripe 82; while this area covers a wide range in Galactic latitude, $20° < b < 70°$, no training set exists in the Galactic plane. While the number of very likely RR Lyrae candidates drops near the Galactic plane, the number of possible candidates with less purity does not. This implies, unsurprisingly, having considerably higher contaminations towards the plane. The purity of low-latitude samples must be settled with follow-up observations and analysis. However, at high galactic latitudes, PS1 3π appears to remain quite complete in its selection to nearly $r_{\rm P1}$∼22 mag for QSO and RR Lyrae, which enables RR Lyrae candidate selection to nearly ∼140 kpc. This is the most extensive and faintest RR Lyrae candidate sample to date, extending to considerably fainter magnitudes than e.g. the CRTS sample of RR Lyrae stars. Using the RR Lyrae in Draco, it is shown that distances derived from $\langle r_{\rm P1} \rangle$ are precise to 6% at a distance of ∼80 kpc.

Candidates of periodic variables can be processed further to increase their purity. As approaches for period finding and fitting are very computational expensive, it needs to be applied to pre-selected candidates. Starting with the RR Lyrae candidates described here, Sesar et al. (2016), produced an even cleaner sample of RR Lyrae candidates by direct light curve fitting, with a completeness of 0.8 and purity of 0.9.

These results of PS1 3π variability studies in the Milky Way context offer for all-sky detection of variable sources. RR Lyrae can be used to precise distance estimates for finding streams and satellites, as carried out within this work for Draco dSph and the Sagittarius stream. QSO candidates will be used as a reference frame for Milky Way astrometry (what is beyond this thesis), to get absolute proper motions and study Milky Way disk kinematics.

Over all, this work has resulted in estimation of variability parameters and mean magnitudes for more than 1.1×10^9 sources, and a catalog of variable sources obtained from a previous PS1 3π

processing version, containing almost 2.58×10^7 objects, being available as a 3π value-added catalog. As all obtained variability features provided by the catalog are general, this catalog allows further source classification based on different training sets than the one presented here. This makes it possible to explore the catalog in order to find variables of other classes than the ones discussed here.

Fitting the Geometry of Sagittarius Stream

A projection of the RR Lyrae candidate sample into the orbital plane of the Sgr stream reveals its morphology clearly. The geometry of the Sgr stream was explored, quantified by its spatial extend and width as a function of the angle $\tilde{\Lambda}_\odot$ in its orbital plane. The geometry of the Sgr stream was fitted with a model that assumed a power-law halo for the background and the line-of-sight density of the Sagittarius stream approximated by a Gaussian distribution centered on the heliocentric distance and a line-of-sight width. This model was used to estimate distance and width of the Sgr stream as given by RR Lyrae candidates from Sesar et al. (2016) (completeness=0.8, purity=0.9, distance precision of 3%) resulting from the classification that incorporates period fitting.

The fitting resulted into the first complete (i.e., spanning the complete angular distribution) trace of Sgr stream's heliocentric distance, as well as the first comprehensive mapping of the line-of-sight depth. Besides the distinct trace itself, this dataset enabled the discovery of the continuation of Sgr stream out to more than 120 kpc near its trailing apocenter (Sesar et al. in prep.), in accord with the simulations by Gibbons et al. (2014). The precision of the model obtained from the data shows that this sample is excellent for mapping stellar (sub-)structure in the Galactic halo.

Having this model of the geometry of the Sgr stream at hand, it can be used in subsequent work to further constrain the Milky Way's potential.

Outlook and Conclusion

In this thesis, a thorough study of automated variable source classification from sparse, unevenly sampled multi-band light curves was developed and carried out, and it was shown that this can produce excellent samples of well-classified variables. The author attributes this success to all of the following advances: usage of non-periodic, general light-curve features; usage of a supervised machine-learning classifier; extension of the classification process by features specific for the assumed class of source (i.e. periodic features for likely RR Lyrae to constrain the sample even more).

In this thesis, three science cases were explored, namely time-series evaluation and fitting, source classification, and structure/overdensity fitting. Such applications are very common in a number of studies. For this reason, throughout this thesis, a high value was set on the fact that developed modeling techniques are as general as possible to enable both an extension to further classes of variable sources, and the application to upcoming all-sky time-domain surveys. The analysis had shown that machine-learning approaches are a powerful tool to inspect data sets being large,

sparse and showing multi-dimensional feature spaces that can not easily separated using hard pre-set cuts.

However, it is important to note that such approaches need to be carefully tested on comparable, reliable data, supported by either other surveys or mock data. Care must be taken in translating science cases and their questions into a form that is "understandable" by a supervised (or unsupervised) machine-learning approach - e.g. selecting a set of features that can act as a proxy for the source classes in case and let the classifier select the manifestation being relevant for each class in order to classify.
Additionally, it is important to notice that it is much harder to understand the results of a machine-learning method than of a "classical" one.

Among upcoming surveys, where the methods presented here can be applied to, LSST is of special interest. For this purpose, a detailed comparison between the technical aspects of both surveys and their possibilities regarding source classification was done. The methodology developed and applied here will be helpful in the early stages of LSST, when only a few observations on a short baseline are available. Also, LSST will carry out a sub-survey observing the Galactic plane with a cadence comparable to PS1 3π, with 12 observations in each of its *ugrizy* bands on a baseline of 4 years. Compared to PS1 3π, LSST will offer observations by 2 mag deeper, down to 24 mag in i band, making it possible to study RR Lyrae even close to the Galactic plane.

A

Appendix

A.1 Time Series Analysis

In this chapter, the mathematical background of time series models and analysis is presented. This is mostly based on Coad (2012) and Rybicki (1994).

A.1.1 Stationary Time Series Models

A time series is given as a sequence of random variables $\{\mathbf{X}_t\}_{t=1,2,\cdots}$. Since there may be an infinite number of random variables, we consider multivariate distributions of random vectors, that is, of finite subsets of the sequence $\{\mathbf{X}_t\}_{t=1,2,\cdots}$ in order to describe time series.

Definition 1. *A time series model for the observed data* $\{x_t\}$ *is defined to be a specification of all of the joint distributions of the random vectors* $\mathbf{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$, $n = 1, 2, \cdots$ *of which* $\{x_t\}$ *are possible realizations, that is, at all of these probabilities*

$$P(X_1 \leq x_1, \cdots, X_n \leq x_n),\ -\infty < x_1, \cdots, x_n < \infty,\ n = 1, 2, \cdots . \tag{A.1}$$

Such a specification is rather impractical. Instead, consider the first and second-order momenta of a joint distribution, that is the expectation values

$$\mathrm{E}(X_t) \text{ and } \mathrm{E}(X_{t+\Delta t} X_t) \text{ for } t = 1, 2, \cdots \text{ and } \Delta t = 0, 1, 2, \cdots , \tag{A.2}$$

and use these so-called *second-order properties* in order to describe the properties of the time series.

Definition 2. $\{\mathbf{X}_t\}$ *is a Gaussian time series if all of its joint distributions are multivariate normal, that is, if for any collection of integers* $i_1, \cdots, i_n$ *the random vector* $(X_{i_1}, \cdots, X_{i_n})^{\mathrm{T}}$ *has a multivariate normal distribution.*

Weak Stationarity and Autocorrelation

For a n-dimensional random vector $\mathbf{X}$, one can calculate the covariance matrix. As a time series usually involves a large (infinite in theory) number of random variables, this results into a very large number of pairs of variables. So the *autocovariance* γ is defined as an extension of the covariance matrix,

$$\gamma(x_{t+\Delta t}, x_t) = \mathrm{Cov}(X_{t+\Delta t}, X_t) \tag{A.3}$$

for all indices t and lags Δt.

Definition 3. *A time series $\{\mathbf{X}_t\}$ is called weakly stationary or just stationary if*

(i) $\mathrm{E}(\mathbf{X}_t) = \mu_{X_t} = \mu < \infty$*, that is, the expectation of* $\mathbf{X}_t$ *is finite and is not depending on* t *and*
(ii) $\gamma(x_{t+\Delta t}, x_t) = \gamma_\tau$*, that is, for each* Δt*, the autocovariance of* $(\mathbf{X}_{t+\Delta t}, \mathbf{X}_t)$ *is not depending on* t*.*

Remark 1. *If $\{\mathbf{X}_t\}$ is a weakly stationary time series, then the autocovariance $\gamma(x_{t+\Delta t}, x_t)$ may be viewed as a function of Δt. It is called the autocovariance function (ACVF). When it is clear which time series it refers to, it is often written as $\gamma(\Delta t)$.*
Note that

$$\gamma(0) = \mathrm{Var}(\mathbf{X}_t), \tag{A.4}$$

that is, the variance is constant for all t.

Definition 4. *Similarly, the autocorrelation function (AFC) is defined by*

$$\rho_X(\Delta t) = \frac{\gamma_X(\Delta t)}{\gamma_X(0)} = \mathrm{Corr}(\mathbf{X}_{t+\Delta t}, \mathbf{X}_t) \tag{A.5}$$

for all t and Δt.

Example 1. ***White noise***
A sequence $\{\mathbf{X}_t\}$ of uncorrelated random variables, each with zero mean and variance σ^2, is called white noise. It is denoted by

$$\{\mathbf{X}_t\} \sim \mathrm{WN}(0, \sigma^2). \tag{A.6}$$

The name "white" indicates that all possible periodic oscillations are present with equal strength, so it is an analogy with white light.

Example 2. ***MA(1) process***
The series defined by the combination of two neighboring white noise variables given by

$$X_t = Z_t + \theta Z_{t-1}, t = 0, \pm 1, \pm 2, \cdots \tag{A.7}$$

where

$$\{Z_t\} \sim \mathrm{WN}(0, \sigma^2) \tag{A.8}$$

and θ is a constant, called a first order moving average, what is denoted by MA(1). $\mathrm{WN}(0, \sigma^2)$ *refers to white noise.*

A.1.2 Sample Autocorrelation Function

The autocorrelation function is a helpful tool in time series analysis. It is used for assessing the degree of dependence and in recognizing what kind of model the time series follows.
When trying to fit a model to an observed time series, the so-called *sample autocorrelation function* based on the data is used. It is defined analogously to the ACF for a time series $\{\mathbf{X}_t\}$.

Definition 5. *Let $x_1, \cdots, x_N$ be observations of a time series. Then the sample autocovariance function is defined by*

$$\hat{\gamma}(\Delta t) = \frac{1}{n} \sum_{t=1}^{n-|\Delta t|} (x_{t+|\Delta t|} - \bar{x}),\ -n < \Delta t < n \tag{A.9}$$

where

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t. \tag{A.10}$$

The sample autocorrelation function is defined by

$$\hat{\rho}(\Delta t) = \frac{\hat{\gamma}(\Delta t)}{\hat{\gamma}(0)}\ -n < \Delta t < n. \tag{A.11}$$

Remark 2. *For a lag $\Delta t \geq 0$, the sample autocovariance function is approximately equal to the sample covariance of the $n - \Delta t$ pairs $(x_1, x_{1+\Delta t}), \cdots, (x_{n-\Delta t}, x_n)$. Note that, in (A.9), the sum is divided by n, not by $n - \Delta t$, and also the overall mean $\bar{x}$ is used for both x_t and $x_{t+\Delta t}$.*

The role of the ACF in prediction

Suppose that $\{\mathbf{X}_t\}$ is a stationary Gaussian time series and X_n is an observed value. Then we would like to predict $X_{n+\Delta t}$ with high precision. The mean square error,

$$\text{MSE} = \text{E}\left[\{X_{n+\Delta t} - f(X_{n+\Delta t}|X_n)\}^2\right] \tag{A.12}$$

is a good measure of precision of the prediction. It is minimized when the function f is the conditional expectation of $X_{n+\Delta t}$ given X_n, that is

$$f(X_{n+\Delta t}|X_n) = \text{E}(X_{n+\Delta t}|X_n). \tag{A.13}$$

For a stationary Gaussian time series, using the equation for conditional expectation and variance of a bivariate normal random variable:

$$\text{E}(X_{n+\tau}|X_n = x_n) = \mu_{n+\Delta t} + \rho(\tau)\sigma_{n+\Delta t}\sigma_n^{-1}(x_n - \mu_n) = \mu + \rho(\Delta t)(x_n - \mu) \tag{A.14}$$

and

$$\text{Var}(X_{n+\tau}|X_n = x_n) = \sigma_{n+\tau}^2 \left\{1 - \rho(\tau)^2\right\}. \tag{A.15}$$

From this it follows that as $\rho(\tau) \to 1$, the value of the precise measure MSE $\to 0$. This means that, the higher the correlation is at lag τ, the more precise is the prediction of $X_{n+\tau}$ based on the observed X_n. Similar conclusions can be drawn about the prediction of $X_{n+\tau}$ based on the observed $X_n, X_{n-1}, \cdots$.

Properties of the ACVF and ACF

In this section, some basic properties of the autocovariance function (ACVF) are outlined, as described in Coad (2012).

Proposition 1. *The ACVF of a stationary time series is a function $\gamma(\cdot)$ such that*

(i) $\gamma(0) \geq 0$
(ii) $|\gamma(\Delta t)| \leq \gamma(0)$ *for all* Δt
(iii) $\gamma(\cdot)$ *is even, that is,* $\gamma(\Delta t) = \gamma(-\Delta t)$ *for all* Δt.

Proof.

(i) That is obvious, as $\gamma(0) = var(X_t) \geq 0$.
(ii) From the definition of correlation and stationarity of the time series, we have

$$|\gamma(\Delta t)| = |\rho(\Delta t)|\sigma^2, \tag{A.16}$$

where $\sigma^2 = \text{Var}(X_t)$. Also, $|\rho(\Delta t)| \leq 1$. Hence,

$$|\gamma(\tau)| = |\rho(\Delta t)|\sigma^2 < \sigma^2 = \gamma(0). \tag{A.17}$$

(iii) Thus, $\gamma(\Delta t) = \text{Cov}(X_{t+\Delta t}, X_t) = \text{Cov}(X_t, X_{t+\Delta t}) = \gamma(-\Delta t)$.

□

Another important property of the ACVF is given by the following theorem.

Theorem 1. *A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and nonnegative definite.*

Proof. A real-valued function κ defined on the integers is nonnegative definite if

$$\sum_{i,j=1}^{n} a_i \kappa(i-j) a_j \geq 0 \tag{A.18}$$

for all positive integers n and real-valued vectors $\mathbf{a} = (a_1, \cdots, a_n)^{\mathrm{T}}$.
Consider a vector random variable $\mathbf{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$ whose covariance matrix C is given by

$$C = \begin{pmatrix} \gamma(0) & \gamma(1-2) & \cdots & \gamma(1-n) \\ \gamma(2-1) & \gamma(0) & \cdots & \gamma(2-n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{pmatrix}. \tag{A.19}$$

Then, letting $\mathbf{Z} = (X_1 - \mathrm{E}(X_1), \cdots, X_n - \mathrm{E}(X_n))^{\mathrm{T}}$, we can write

$$\begin{aligned} 0 \leq \mathrm{Var}(\mathbf{a}^{\mathrm{T}}\mathbf{Z}) =& \mathrm{E}\left\{(\mathbf{a}^{\mathrm{T}}\mathbf{Z})(\mathbf{a}^{\mathrm{T}}\mathbf{Z})^{\mathrm{T}}\right\} \\ =& \mathrm{E}\left\{\mathbf{a}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{a}\right\} \\ =& \mathbf{a}^{\mathrm{T}}C\mathbf{a} = \sum_{i,j=1}^{n} a_i \gamma(i-j) a_j. \end{aligned} \tag{A.20}$$

Hence, $\gamma(\tau)$ is a non-negative definite function.

□

A.1.3 Strict Stationary Time Series Models

A more restrictive definition of stationarity involves all the multivariate distributions of the subsets of time series random variables.

Definition 6. *A time series $\{\mathbf{X}_t\}$ is called strictly stationary if the random vectors $(\mathbf{X}_{t_1}, \cdots, \mathbf{X}_{t_n})^{\mathrm{T}}$ and $(\mathbf{X}_{t_1+\Delta t}, \cdots, \mathbf{X}_{t_n+\Delta t})^{\mathrm{T}}$ have the same joint distribution for all sets of indices $\{t_1, \cdots t_n\}$ and for all integers Δt and $n > 0$. It is written as*

$$(\mathbf{X}_{t_1}, \cdots, \mathbf{X}_{t_n})^{\mathrm{T}} \stackrel{\mathrm{d}}{=} (\mathbf{X}_{t_1+\Delta t}, \cdots, \mathbf{X}_{t_n+\Delta t})^{\mathrm{T}}, \tag{A.21}$$

where $\stackrel{\mathrm{d}}{=}$ means "equal in distribution".

Definition 7. *Properties of a Strictly Stationary Time Series*

(i) The random variables $\mathbf{X}_t$ are identically distributed for all t.
(ii) Pairs of random variables $(\mathbf{X}_t, \mathbf{X}_{t+\Delta t})^{\mathrm{T}}$ are identically distributed for all t and Δt, that is $(\mathbf{X}_t, \mathbf{X}_{t+\Delta t})^{\mathrm{T}} \stackrel{d}{=} (\mathbf{X}_1, \mathbf{X}_{1+\tau})^{\mathrm{T}}$
(iii) The series $\mathbf{X}_t$ is a weakly stationary time series if $\mathrm{E}(\mathbf{X}_t^2) < \infty$ for all t.
(iv) Weak stationarity does not imply strict stationarity.

Proofs of this properties can be found in Coad (2012).

A.1.4 Random Walk

A *Random Walk* is a time series where each point of time in the series moves randomly away from its current position. The model can then be written as

$$X_t = X_{t-1} + Z_t, \tag{A.22}$$

where Z_t is a white noise variable with zero mean and variance σ^2. This model is *not stationary.*

Repeatedly substituting for past variables results in

$$\begin{aligned} X_t &= X_{t-1} + Z_t \\ &= X_{t-2} + Z_{t-1} + Z_t \\ &= X_{t-3} + Z_{t-2} + Z_{t-1} + Z_t \\ &\vdots \\ &= X_0 + \sum_{j=0}^{t-1} Z_{t-j}. \end{aligned} \tag{A.23}$$

If the initial value X_0 is fixed, then the expectation value of X_t is fixed and equal to X_0, that is,

$$\mathrm{E}(X_t) = \mathrm{E}\left(X_0 + \sum_{j=0}^{t-1} Z_{t-j}\right) = X_0. \tag{A.24}$$

In contrast, the variance and covariance depend both on time and the lag.
Since the white noise variables Z_t are uncorrelated, the variance is

$$\begin{aligned} \mathrm{Var}(X_t) &= v\mathrm{Var}\left(X_0 + \sum_{j=0}^{t-1} Z_{t-j}\right) \\ &= \mathrm{Var}\left(\sum_{j=0}^{t-1} Z_{t-j}\right) \\ &= \sum_{j=0}^{t-1} \mathrm{Var}\left(Z_{t-j}\right) \\ &= t\sigma^2 \end{aligned} \tag{A.25}$$

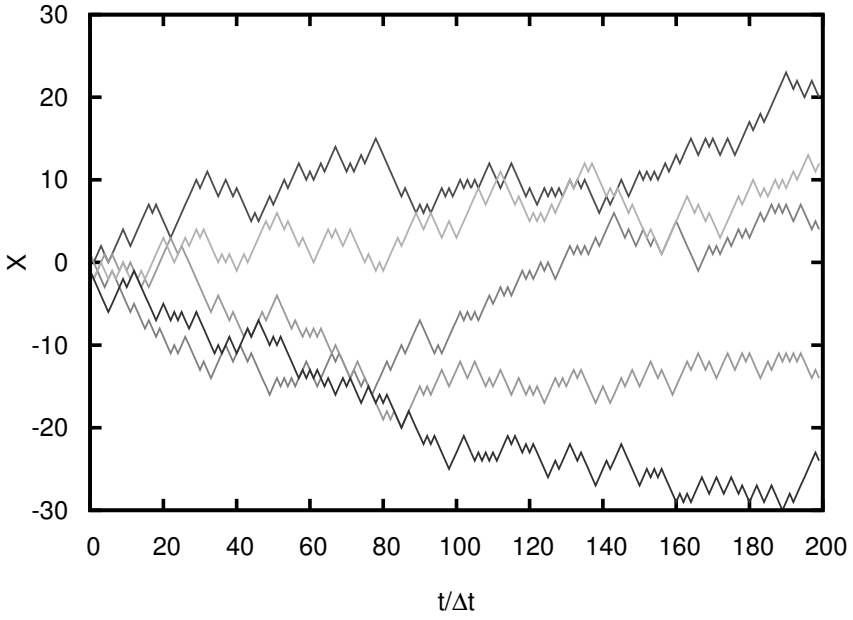


Figure A.1 Different realizations of a 1D Random Walk time series with 200 time steps.

and the covariance

$$\begin{aligned}\mathrm{Cov}(X_t, X_{t-\tau}) &= \mathrm{Cov}\left(\sum_{j=0}^{t-1} Z_{t-j}, \sum_{k=0}^{t-\tau-1} Z_{t-\tau-k}\right) \\ &= \mathrm{E}\left\{\left(\sum_{j=0}^{t-1} Z_{t-j}\right)\left(\sum_{k=0}^{t-\tau-1} Z_{t-\tau-k}\right)\right\} \\ &= \min(t, t-\tau)\sigma^2.\end{aligned} \tag{A.26}$$

It is evident from this that the random walk meanders away from its initial value in no particular direction without showing any clear trend, but, at the same time, is not stationary. An example showing different realizations of a 1D Random Walk time series is given in Fig. A.1.4.

A.1.5 Damped Random Walk - The Ornstein-Uhlenbeck Process

The Ornstein-Uhlenbeck process (named after Leonard Ornstein and Georg Eugene Uhlenbeck) is a stochastic process that original describes the velocity of a massive Brownian particle under the influence of friction. The process is stationary, Gaussian and Markovian, allows linear transformations of its space and time variables; it is the only nontrivial process satisfying all three conditions. Over time, the process tends to drift toward its long-term mean; such a process is called *mean-reverting*. The process can be considered as a modification of the damped random walk in continuous time, or Wiener process, in which the properties of the process have been changed in a way that it has a tendency to move back towards a central location (its mean), having a greater attraction when the process is further away from the center, which makes its

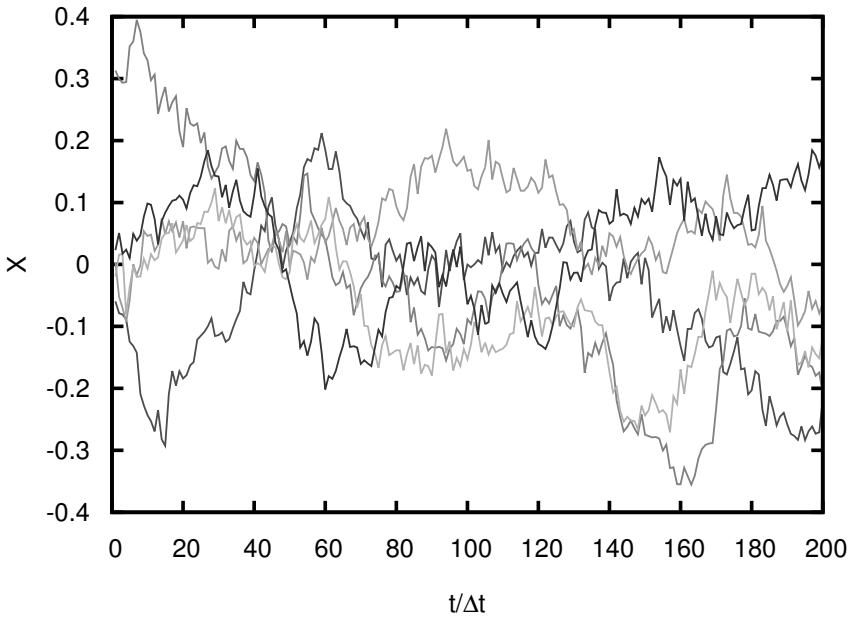


Figure A.2 Different realizations of a 1D Damped Random Walk time series with 200 time steps.

movement being "damped". Thus it is commonly named the *Damped Random Walk*. The following description of the Damped Random Walk is based on Rybicki (1994).

Representation via a Stochastic Differential Equation

A Damped Random Walk X_t satisfies the following stochastic differential equation:

$$\mathrm{d}X_t = \tau(\mu - X_t)\mathrm{d}t + \omega\,\mathrm{d}W_t \tag{A.27}$$

where $\omega > 0$, $\tau > 0$ and μ are parameters and W_t denotes the Random Walk.

An example showing different realizations of a 1D Damped Random Walk time series is given in Fig. A.1.5.

The Structure Function and Joint Probability Distribution

The Damped Random Walk has zero mean and the exponential correlation function

$$\Phi(\Delta t) = \langle x(t) + x(t+\Delta t)\rangle = \omega^2 \exp\left\langle -\frac{|\Delta t|}{\tau}\right\rangle . \tag{A.28}$$

The constants ω^2 and τ are, respectively, the variance and the decorrelation time of the process.

The structure function is then defined as

$$V(\Delta t) = \frac{1}{2}\langle [x(t) + x(t+\Delta t)]^2 \rangle = \Phi(0) - \Phi(\Delta t) = \omega^2 \left[1 - exp\left(-\frac{|\Delta t|}{\tau}\right)\right]. \tag{A.29}$$

Let now t_i, $1 \leq i \leq n$ be an ordered set of times, $t_1 < t_2 < ... < t_n$, and let $x_i = x(t_i)$ be the values of the process at those times. Then from the definition of a Gaussian process, the joint probability function of the set of values x_i is given by

$$P_n(x_1, ..., x_n) = [\det(2\pi \mathbf{C}_n)]^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{C}_n^{-1}\mathbf{x}\right) \tag{A.30}$$

where the n-dimensional vector $\mathbf{x}$ has components x_i and the $n \times n$ matrix $\mathbf{C}_n$ has components $C_{ij} = \Phi(t_i - t_j)$.

As shown by Rybicki and Press (1995), C_n has a simple inverse, the tridiagonal matrix

$$\mathbf{T}_n = \mathbf{C}_n^{-1} = \begin{pmatrix} d_1 & -e_1 & & & & \\ -e_1 & d_2 & -e_2 & & & \\ & -e_2 & d_3 & -e_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -e_{n-2} & d_{n-1} & -e_{n-1} \\ & & & & -e_{n-1} & d_n \end{pmatrix} \tag{A.31}$$

where the elements not indicated are 0. Defining

$$r_i = \begin{cases} 0, & i = 0 \\ \exp\left[-\frac{t_{i+1}-t_i}{\tau}\right], & 1 \leq i \leq n-1 \end{cases} \tag{A.32}$$

where the quantities e_i and d_i are given by

$$e_i = \begin{cases} 0, & i = 0 \\ r_i/(1-r_i^2), & 1 \leq i \leq n-1 \\ 0, & 1 = n \end{cases} \tag{A.33}$$

and

$$d_i = 1 + r_i e_i + r_{i-1} e_{i-1}, \quad 1 \leq i \leq n. \tag{A.34}$$

Using the tridiagonal matrix Equ. (A.31), the quadratic form in the exponent of Equ. (A.30) can be written as

$$Q_n \equiv \mathbf{x}^{\mathrm{T}}\mathbf{C}_n^{-1}\mathbf{x} = \omega^{-2} \sum_{i=1}^{n} (d_i x_i^2 + 2 e_{i-1} x_i x_{i-1}) \tag{A.35}$$

Substituting for d_i gives

$$\begin{aligned} Q_n &= \sum_{i=1}^{n} \left[(1 - r_{i-1}e_{i-1})x_i^2 + 2e_{i-1}x_i x_{i-1}\right] - \omega^2 \sum_{i=1}^{n} r_i e_i x_i^2 \\ &= \omega^{-2} \sum_{i=1}^{n} \left[1 - r_{i-1}e_{i-1})x_i^2 + 2e_{i-1}x_i x_{i-1} - r_{i-1}e_{i-1}x_{i-1}^2\right]. \end{aligned} \tag{A.36}$$

To end up with the second for, the index in the second sum was shifted. Using the definitions for e_i and d_i from Equ. (A.33) and (A.33),

$$Q_n = \omega^{-2} \sum_{i=1}^{n} \frac{(x_i - r_{i-1}x_{i-1})^2}{1 - r_{i-1}^2}. \tag{A.37}$$

The joint probability distribution function P_n can now be expressed as

$$P_n(x_1, ..., x_n) = [\det(2\pi \mathbf{C}_n)]^{-1/2} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - r_{i-1}x_{i-1})^2}{2\omega^2(1 - r_{i-1}^2)}\right). \tag{A.38}$$

Mathematical properties

The Damped Random Walk is an example of a Gaussian process having a bounded variance and admits a stationary probability distribution, in contrast to the Random Walk; the difference between the two is their "drift" term. For the Dandom Walk, the drift term is constant, whereas for the Damped Random Walk it depends on the current value of the process: if the current value of the process is less than the (long-term) mean, the drift will be positive; if the current value of the process is larger than the (long-term) mean, the drift will be negative, attracting towards the mean.
In other words, the mean acts as an equilibrium level for the process. This gives the process its informal name "mean-reverting". The stationary (long-term) variance is given by

$$\mathrm{Var}(x_t) = \frac{\omega^2}{2\tau}. \tag{A.39}$$

A.2 Markov Chain Monte Carlo Method

For estimating the maximum of a distribution that cannot be solved analytically, a rough approximation can be made on a parameter grid. However, for many reasons it is preferred to sample from this distribution instead. Such reasons are:

- having a high-dimensional distribution: A grid with N points in d-dimensional parameter space will demand N^d function evaluations ("curse of dimensionality"), while the convergence of a MCMC is $\mathcal{O}(N^{-1/2})$, which is dimensionally independent.
- being interested in not only the approximated maximum, but other statistical properties of the distribution
- having a distribution with steep maxima that might be missed if the parameter grid is too approximate, thus resulting in high computation time for a fine grid.

In this case, the Markov Chain Monte Carlo (MCMC) method is very useful to sample from the distribution.

This introduction in Markov Chains is mostly based on Richey (2010), Brooks et al. (2011) and Weinzierl (2000) who give an overview of MCMC research and its application.

A.2.1 The Beginning of Markov Chain Monte Carlo Methods

In 1947, von Neumann and others were working on methods to estimate neutron diffusion and multiplication rates in fission devices. Von Neumann proposed the plan to create a large number of simulated neutrons and use the computer to randomly simulate how they pass through the fissionable material. After doing so, the number of neutrons remaining is counted in order to estimate the desired rates. From this point forward, randomized simulations became an important technique in physics and engineering and have soon be called *Monte Carlo* methods.

Later, Metropolis (1953) simulated a liquid that is in equilibrium with its gas phase. To find out about the thermodynamic equilibrium, they simulated the dynamics of the system, and let it run until it reaches equilibrium. They realized that they did not need to carry out such a detailed an complicated simulation; it would be enough to simulate some Markov chain having the same equilibrium distribution. Simulations following the scheme of Metropolis (1953) are said to use the *Metropolis algorithm*.

The Metropolis algorithm was used by chemists and physicists for similar problems, but was not widely known among other fields until the 1990's. A generalization by Hastings (1970) led to the Metropolis-Hastings algorithm. A special case of the Metropolis-Hastings algorithm was introduced by Geman and Geman (1984), called the *Gibbs sampler*. After Gelfand and Smith (1990) made the wider Bayesian community aware of the Gibbs sampler, it was rapidly realized that most Bayesian inference could be done by Markov Chain Monte Carlo methods. Problems

that had previously been undoable, or extremely hard to solve, then suddenly became solveable in straightforward manner.

Nowadays, MCMC methods are a widely known tool among various scientific fields, from physics to life sciences, and a lot of special cases for dedicated applications exist.

A.2.2 Markov Chains

A Markov chain is a series of stochastic events whereby the *state* of the process at the next time step, $t+1$, depends on:

(i) the current state of the process (e.g., contained in a state matrix)
(ii) the probability of changing to another state in the next time step (e.g., defined in a transition matrix).

Given a finite state space $\mathbb{S} = \{1, 2, \cdots, N\}$, a *Markov chain* is a stochastic process that is defined by a sequence of random variables $X_i \in \mathbb{S}$, for $i = 1, 2, \cdots$, such that

$$p(X_{n+1} = x_{n+1} | X_1 = x_1, \cdots, X_n = x_n) = p(X_{n+1} = x_{n+1} | X_n = x_n). \tag{A.40}$$

In other words, the conditional distribution X_{n+1} given $X_1, ..., X_n$ depends only on X_n, and the set from which X_n take values in $\mathbb{S}$. Here, only Markov chains for which this dependence is independent of n are considered; such Markov Chains are said to have *stationary transition probabilities*. When the state space is countably infinite, one can think of an infinite transition matrix.

This gives a $N \times N$ *transition matrix* $\mathbf{P}_{ij} = (\mathbf{p}_{ij})$, defined by

$$\mathbf{P}_{ij} \equiv P(X_{n+1} = j | X_n = i). \tag{A.41}$$

Note that for $i = 1, 2, \cdots, N$,

$$\sum_{j=1}^{N} \mathbf{p}_{ij} = 1. \tag{A.42}$$

The (i, j)-entry of the kth power of $\mathbf{P}$ gives the probability of transitioning from state i to state j in k steps.

But most Markov chains at interest in MCMC have uncountable state space, so we must think of it as a conditional probability distribution.

Two desirable properties of a Markov chain are:

(i) it is irreducible: for all states i and j, there exists k such that $(\mathbf{P}^k)_{i,j} \neq 0$

(ii) it is aperiodic[8]: for all states i and j, $gcd\left\{k : (\mathbf{P}^k)_{i,j} > 0\right\}$ =1.

An irreducible, aperiodic Markov chain must have an unique distribution $\pi = (\pi_1, \pi_2, \cdots, \pi_N)$ on the state space $\mathbb{S}$ (π_i = the probability of state i) with the property that $\pi = \pi\mathbf{P}$.
A Markov chain is said to be *stable on the distribution* π, or that π is the *stable distribution* for the Markov chain. If π is the stable distribution for an irreducible, aperiodic Markov chain, then we can use the Markov chain to sample from π.

Drawing Samples

To obtain a sample, select $s_1 \in \mathbb{S}$ arbitrary. Then for any $k > 1$, if $s_{k-1} = i$, select $s_k = j$ with probability $\mathbf{P}_{ij}$. The resulting sequence $s_1, s_2, \cdots$ has the property that as $M \to \infty$,

$$\frac{|\{k \leq M \text{ and } s_k = j\}|}{M} \to \pi_j \tag{A.43}$$

with probability 1.
Any large (but finite) sub-sequence approximates a sample drawn from π. Often, the first m terms of the sequence are discarded, and remaining $s_{m+1}, s_{m+1}, \cdots, s_M$ are used.
When doing so, the process of removing the first m samples is referred to as "burn-in" (see also Sec. A.2.4).

No matter how they are obtained, samples from π provide a way to approximate the properties of π. For example, suppose f is any real-valued function on the state space $\mathbb{S}$ and the expectation value needs to be approximated,

$$\mathrm{E}[f] = \sum_{i=1}^{N} f(i)\pi_i. \tag{A.44}$$

[8] A state i has period k if any return to state i must occur in multiples of k time steps. Formally, the period of a state is defined as

$$k = \gcd\{n : \Pr(X_n = i | X_0 = i) > 0\}$$

(where "gcd" is the greatest common divisor).

To do so, select a sample $s_1, s_2, \cdots, s_M$ from π and the ergodic theorem[9] guarantees that

$$\frac{1}{M}\sum_{i=1}^{M} f(s_i) \to \mathrm{E}\left[f\right] \tag{A.47}$$

as $M \to \infty$ with the convergence $\mathcal{O}(M^{-1/2})$.

Stationarity

Some Markov chains possess a unique equilibrium distribution. Informally, this means that if starting the chain somewhere in the state space and run the chain long enough, the chain will settle into an equilibrium distribution independent of the initial condition. We say that the chain becomes *stationary*.

A sequence $X_1, X_2...$ of random elements of some set is called a stochastic process (Markov chains are a special case). A stochastic process is *stationary* if for every positive integer k the distribution of the k-tuple

$$(X_{n+1}, ..., X_{n+k}) \tag{A.48}$$

does not depend on n.

In a Markov chain, the conditional distribution of $(X_{n+2}, ..., X_{n+k})$ given X_{n+1} does not depend on n. It follows that a Markov chain is stationary if and only if the marginal distribution of X_n does not depend on n.

Stationarity implies stationary transition probabilities from X_n to X_{n+1}, but not vice versa. Consider an initial distribution concentrated at one point. The Markov chain can be stationary if and only if all elements are concentrated at the same point, that is, $X_1 = X_2 = ...$, so the chain goes nowhere and does nothing. Conversely, any transition probability distribution can be combined with any initial distribution, including those concentrated at one point. Such a chain is usually no stationary, even though the transition probabilities are stationary.

[9] Ergodic Theorems concern the limiting behavior of averages over time. The most famous ergodic theorem is the one for independent random variables.

Theorem 2. *Let P be irreducible and let π_0 be an arbitrary distribution. Suppose $X_n \sim Markov(\pi_0, P)$ and set the number of visits to state i before time n as*

$$N_i(n) = \sum_{k=0}^{n-1} 1(X_k = i). \tag{A.45}$$

So, $N_i(n)/n$ is the proportion of time spent in state i before n. Then,

$$P\left(\frac{N_i(n)}{n} \to \frac{1}{m_i} \text{ as } n \to \infty\right) = 1 \tag{A.46}$$

where $m_i = \mathrm{E}_i(T_i)$ is the expected return time to state i.

A.2.3 Markov Chain Monte Carlo Sampling

MCMC algorithms are a widely used tool for sampling from, and calculating integrals of, complicated and high dimensional distributions that occur in a range of contexts, from computational physics and engineering to Bayesian statistics. If one would carry out these integrations in a straight forward manner, by evaluating the distribution deterministically over the entire state space at a set resolution, the time necessary for the computation would quickly become prohibitive. MCMC algorithms take a different approach. Rather than providing a high dimensionality result only at the end of the computation, the MCMC algorithm takes a stochastic approach, and provides an approximation that gradually becomes more accurate over the time the program executes.

MCMC algorithms are based on the idea that, if we don't know how to analytically solve for a distribution, say for a posterior distribution, then we can at least learn about it by constructing a Markov chain whose stationary distribution is the one that we are interested in learning about. If a Markov chain is constructed in the right way, one can use the MCMC to learn about a distribution to arbitrary precision.

Starting from a initial sample (either randomly generated or given by some reasonable choice) over the state (parameter) space, the algorithm uses a stochastic transition function to produce new, though not necessarily different in value, sample using a proposal distribution. There is an acceptance probability for the newly generated sample, and it is what guarantees the chain will become an approximation of the target distribution that should be sampled. The transition function is then recursively applied to each newly sample, resulting in a chain of samples.

Update mechanisms of interest preserve a specified distribution, that is, if the state has the specified distribution before update, then it has the same after the update. This leads to the construction of Markov chains to sample the distribution. Update mechanism are called *elementary* if it is not made up of parts that are themselves update mechanisms preserving the specified distribution.

As long as the transition function can take the chain over the target distribution's entire state space, the chain will finally approximate the target distribution, and as the algorithm runs that approximation will become more accurate. Once the chain has covered the area of state space of statistical interest, the chain is said to have *mixed*.

It is one very desirable property to have a fast mixing chain, as that means that the distribution will have "forgotten" (it depends not longer on) where it has started and so has no bias toward being in the start location. However, it is also often the case that the chain will be initialized in a region of extremely low probability. The first stretch of the chain will then make the rest of the chain a biased approximation for all but very large numbers of samples. For this reason, it is often the case that a section of the chain at the beginning is removed once it has reached a region of non-negligible probability, or a few thousand samples. The removed section of the chain is called the *burn-in*.

Metropolis-Hastings Markow Chain Monte Carlo Algorithm

The main features of Monte Carlo Markov chains for sampling from a distribution with density $p(x)$ are Hastings (1970):

(i) The computations depend on $p(x)$ only through ratios of the form $p(x')/p(x)$, where x' and x are sample points. Thus, the normalizing constant of $p(x)$ need not to be known, no factorization of $p(x)$ is necessary, and the methods are very easily implemented on a computer. Additionally, conditional distributions do not require special treatment and therefore the methods provide a convenient means for obtaining correlated samples from conditional distributions.

(ii) A sequence of samples is obtained by simulating a Markov chain. The resulting samples are therefore correlated and estimation of the standard deviation of an estimate and assessment of the error of an estimate may require more care than with independent samples.

The basic MCMC algorithm is the Metropolis-Hastings Algorithm as described in the following.

Suppose that the specified distribution – the desired stationary distribution of the MCMC sampler in case – has a (general) unnormalized density p. Thus p is a nonnegative-valued function that integrates to a value that is finite and nonzero.
The Metropolis-Hastings Algorithm uses a stochastic transition function $q(x^*|x^{(i)}) = x^{(i)} + N(\mathcal{O}, I)$ which results in a step of random direction and length in the state space from the previous point. A commonly used proposal distribution here is a Gaussian distribution centered at $x^{(i)}$, which tends to move to points nearby $x^{(i)}$ and thus explores the probability space using a random walk. When the current state at iteration i is $x^{(i)}$, a move to a state x^* is proposed. Then the *Hastings ratio*

$$r(x^{(i)}, x^*) = \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})} \tag{A.49}$$

is calculated.
The proposed move to the newly generated sample x^* is accepted with a probability $\alpha(x^{(i)}, x^*) = \min\big(1, r(x^{(i)}, x^*)\big)$. The state $x^{(i+1)}$ after the update is x^* with probability $\alpha(x^{(i)}, x^*)$, and the new state $x^{(i+1)}$ is $x^{(i)}$ with probability $1 - \alpha(x^{(i)}, x^*)$. This is the *Hastings update*.
If one attempts to move to a point being more probable than the current one, the move will always be accepted. If a move to a less probable point is attempted, the move is sometimes rejected, and the more the relative drop in probability, the more likely the new point is rejected. This guarantees the chain will become an approximation of the target distribution $p(x)$.
The transition function is then recursively applied to each new sample, which produces a chain of samples. The random choice of a new parameter value is influenced by the current value. As long as the transition function can take the chain over the entire state space of the target distribution, the chain will eventually approximate this distribution. As the algorithm runs, the approximation

will increase in accuracy.

Algorithm 3: Metropolis-Hastings MCMC

Input: $p(x)$: probability distribution
$q(x)$: proposal distribution
N_{iter}: number of sample iterations
$x^{(i)}$: current state at iteration i
x^*: proposed state

Output: $\{x^i\}$: chain of samples

begin
  initialization $x^{(0)}$
  for $i = 1, ..., N_{\text{iter}}$ **do**
    sample $u \in U[0, 1]$
    sample $x^* \in q(x^*|x^{(i)})$
    $\alpha(x^{(i)}, x^*) = \min\left(1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right)$
    if $u < \alpha$ **then**
      $x^{(i+1)} = x^*$
    else
      $x^{(i+1)} = x^{(i)}$

The Hastings ratio Equ. (A.49) is undefined if $p(x^{(i)}) = 0$, thus one must always arrange that $p(x^{(i)}) > 0$ in the initial state. There is no problem if $p(x^*) = 0$, as in this case, all that happens is that $r(x^{(i)}, x^*) = 0$ and the proposal x^* will accepted with zero probability. For this reason, the Metropolis-Hastings update can never move to a proposed state having $p(x^*) = 0$. Note that the proposal x^* must satisfy $q(x^{(i)}|x^*) > 0$ with probability 1 because $q(x^{(i)}|\cdot)$ is the conditional density of x^* given $x^{(i)}$. Hence, still assuming $p(x^{(i)}) > 0$, as the Hastings ratio is well defined, the denominator of the Hastings ratio is nonzero with probability 1. Note that either term of the numerator of the Hastings ratio can be 0, so the proposal is almost surely rejected, as either $p(x^*) = 0$ or $q(x^*|x^{(i)}) = 0$. So, there is no need to arrange for proposals being always possible values of the desired equilibrium distribution. The only demand is to ensure that one's implementation of the (unnormalized) density function p works when given any possible proposal as an argument, including giving $p(x^*) = 0$ for invalid x^*.

There are some special cases of the general updating process, the above Metropolis-Hastings update. The *Metropolis update* describes the case of $q(x^{(i)}|x^*) = q(x^*|x^{(i)})$ for all $x^{(i)}$ and x^*. For

a symmetric proposal density like a Gaussian distribution being centered at $x^{(i)}$, it is $q(x^{(i)}|x^*) = q(x^*|x^{(i)})$, so it cancels out. Then the Hastings ratio Equ. (A.49) reduces to

$$r(x^{(i)}, x^*) = \frac{p(x^*)}{p(x^{(i)})} \tag{A.50}$$

and is called the *Metropolis ratio* or *odds ratio*.

In the special case of a *Gibbs update*, the proposal is from a conditional distribution of the desired equilibrium distribution. Thus, it is always accepted. Gibbs updates have one property not shared by other Metropolis-Hastings updates: they are idempotent, meaning the effect of multiple updates is the same as the effect just one.

Using the transition function $q(x^*|x^{(i)}) = x^{(i)} + N(\mathcal{O}, I)$ the Metropolis-Hastings MCMC algorithm will eventually approximate any distribution it is given.
In practice, of course, reliable results as soon as possible are required. In order to make sure the chain mixes quickly, the random step should be wide enough to mix quickly, meaning the average acceptance isn't too high, and to make sure that the average rate of new states that are accepted is not too low reducing deficiency of the algorithm.

This is where the idea of adaptive MCMC comes in, which aims to automatically tune the parameters of the transition function (i.e., the width of the random step distribution) towards good acceptance rates. There are several ones, e.g. the Parallel Tempered MCMC algorithm, where so-called "hot" chains are more eager to accept jumps to lower likelihood and hence sample a broad range of the parameter space, whereas "cooler" chains are more aversely to do so. We present here the one that is used in the program developed through this work, the *Affine Invariant Markov Chain Monte Carlo (MCMC) Ensemble sampler*.

Affine Invariant Markov Chain Monte Carlo (MCMC) Ensemble Sampler

The paper of Goodman and Weare (Goodman and Weare 2010) and a more implementation-related paper of Foreman-Mackey et al. (Foreman-Mackey et al. 2012) show an advanced usage of an ensemble of so-called *walkers* in the Affine Invariant Markov Chain Monte Carlo (MCMC) Ensemble sampler. Its implementation `emcee` is used throughout the work resented in this thesis.

MCMC sampling methods typically have parameters that must be justified for a specific problem. For example, a step size that is sensible for some probability density $\pi(x)$ with $x \in \mathbb{R}^n$, may work poorly for the scaled probability density

$$\pi_\lambda(x) = \lambda^n \pi(\lambda x) \tag{A.51}$$

where $\lambda \in \mathbb{R}$ is very large or very small. The performance in sampling the density π_λ that is independent of λ.

Affine invariant samplers can be implemented in several ways. The method described here simultaneously evolves an ensemble of K walkers $S = \{X_k\}$ where the walkers are almost like separate Metropolis-Hastings chains but the proposal distribution for a given walker k depends on the positions in the N-dimensional parameter space of the $K-1$ walkers in the complementary ensemble $S_{[k]} = \{X_j, \forall j \neq k\}$.
To update the position of a walker k at position X_k, a walker X_j is drawn randomly from the complementary ensemble $S_{[k]}$, and a new position in parameter space is proposed as

$$X_k(t) \to Y = X_j + Z\left[X_k(t) - X_j\right] \tag{A.52}$$

with Z being a random variable drawn from some distribution $g(Z = z)$.
The distribution g has to fulfil

$$g(z^{-1}) = zg(z)$$

to make the proposal of (A.52) symmetric in the sense that $P(X_k(t) \to Y) = P(Y \to X_k(t))$. On that condition, the sampling chain will satisfy detailed balance if the proposed state is accepted with probability

$$q = \min\left(1, Z^{N-1}\frac{p(Y)}{p(X_k t)}\right). \tag{A.53}$$

This is done for each walker in the ensemble following the algorithm shown in to complete one update step.
For the distribution g, Goodman and Weare (2010) and Foreman-Mackey et al. (2012) are using

$$g(z) = \begin{cases} \frac{1}{c}\frac{1}{\sqrt{z}} & \text{if } z \in \left[\frac{1}{a}, a\right] \\ 0 & \text{otherwise} \end{cases}$$

with a normalizing constant of

$$C = \frac{1}{2}\left(\sqrt{a} - \frac{1}{\sqrt{a}}\right)$$

The parameter $a > 1$ can be adjusted to improve performance and is set to 2 in most cases.

This algorithm outperforms standard MCMC methods like the Metropolis-Hastings algorithm in producing independent samples with a much shorter autocorrelation time Foreman-Mackey et al. (2012). Faster convergence is preferred due to the reduction of computational costs as the number of computations being necessary to obtain the equivalent level of accuracy can be reduced.

Algorithm 4: Affine Invariant Markov chain Monte Carlo (MCMC) Ensemble sampler

Input: $p(x)$: probability distribution
$q(x)$: proposal distribution
N_{iter}: number of sample iterations
K: number of walkers
$x_k^{(i)}$: current state at iteration i for walker k
x^*: proposed state for walker k

Output: $\{x_k^i\}$: chain of samples for walker k

begin
- initialization $x^{(0)}$
- **for** $i = 1, ..., N_{\text{iter}}$ **do**
  - **for** $k = 1, ..., K$ **do**
    - draw a walker x_j at random from the complementary ensemble $S_{[k]}^{(i)}$
    - sample $z \in g(z)$
    - sample $x^* = x_j + z[x_k^{(i)} - x_j] \in q(x^*|x^* x_k^{(i)})$
    - $\alpha(x_k^{(i)}, x^*) = z^{N-1} \frac{p(x^*) q(x_k^{(i)}|x^*)}{p(x_k^{(i)}) q(x^*|x_k^{(i)})}$
    - **if** $u < \alpha$ **then**
      - $x_k^{(i+1)} = x^*$
    - **else**
      - $x_k^{(i+1)} = x_k^{(i)}$

Parallel Affine Invariant Markov Chain Monte Carlo (MCMC) Ensemble Sampler

For most applications on nowadays computing facilities, it is needed or at last sensible to parallelize the algorithm. Due to the fact that the algorithm is based on an ensemble of walkers, it seems tempting to simple parallelize the algorithm by simultaneously advancing each walker based on the state of the ensemble instead of evolving the walkers in series. Unfortunately, this subtly violates detailed balance of the chain. Instead, the full ensemble has to be split up in two subsets $S^{(0)} = \{x_k, \forall k = 1, \cdots, K/2\}$ and $S^{(1)} = \{x_k, \forall k = K/2 + 1, \cdots, K\}$ and simultaneous update all walkers in $S^{(0)}$ based only on the positions of the walkers in the complementary set $S^{(1)}$. Then, based on the new positions in $S^{(0)}$, $S^{(1)}$ is updated. In this case, the outcome is a valid step for all of the walkers.
This algorithm is shown below. It is similar to Algorithm (4) but the computationally expensive inner loop can now run in parallel, so one can take now advantage of generic parallelization that makes this algorithm extremely powerful.

Algorithm 5: Parallel Affine Invariant Markov chain Monte Carlo (MCMC) Ensemble sampler

Input: $p(x)$: probability distribution
$q(x)$: proposal distribution
N_{iter}: number of sample iterations
K: number of walkers

Output: $\{x_k^i\}$: chain of samples for walker k

begin
- initialization $x^{(0)}$
- **for** $i = 1, ..., N_{\text{iter}}$ **do**
  - **for** $b = 0, 1$ **do**
    - **for** $k = 1, ..., K/2$ **do**
      - draw a walker x_j at random from the complementary ensemble $S_{[k]}^{(\sim b)(i)}$
      - $x_k^{(i)} = S_k^{(b)}$
      - sample $z \in g(z)$
      - sample $x^* = x_j + z[x_k^{(i)} - x_j] \in q(x^*|x^* x_k^{(i)})$
      - $\alpha(x_k^{(i)}, x^*) = z^{N-1} \frac{p(x^*) q(x_k^{(i)}|x^*)}{p(x_k^{(i)}) q(x^*|x_k^{(i)})}$
      - **if** $u < \alpha$ **then**
        - $x_k^{(i+1/2)} = x^*$
      - **else**
        - $x_k^{(i+1/2)} = x_k^{(i)}$
    - $i = i + 1/2$

A.2.4 Application of MCMC Methods

Despite there exists a lot of theory on the convergence of Markov chains, for carrying out a MCMC application, more practical approaches are needed. Without them, not much more than the output would be known about the Markov chain. This could lead to erroneous results on the one hand, and waste of a lot of computation time on the other hand due to chains running too long.

For this reason, it is important to have some methods at hand that give information on the internal states of MCMC and also give hints on the reliability of the results obtained by MCMC.

Pseudo-Convergence

A Markov chain can appear to have converged to its equilibrium distribution when it has not. This is often caused by a state space being poorly connected. In this case, it takes many iterations to get from one part of the state space to another in order to explore the state space fully, or maybe the chain never finds the other part as it "gets stuck" in one. When the number of steps needed for transition between these parts exceeds the length of the simulated Markov chain, the Markov chain can appear to have converged but the distribution it appears to have converged to is only the equilibrium distribution conditioned on the part in which the chain was started. This phenomenon is known as *pseudo-converge*. It has also been called "multimodality" since it may occur when the equilibrium distribution is multimodal. But multimodality does not necessarily cause pseudo-convergence when the troughs between the different parts of the state space are not severe. Also, pseudo-convergence does not only happen when in the presence of multimodality. Approaches to overcome the problem of pseudo-convergence are Affine Invariant MCMC sampler.

Autocorrelation Function

As a measure for the "quality" of an algorithm, the inverse convergence rate can be measured by the autocorrelation function. This is an estimate of the number of steps needed in the chain in order to draw independent samples from the target chain. A more efficient chain has a shorter autocorrelation time.

Something else to keep in mind is the error of the MCMC as all draws one gets after a finite number of iterations are only approximations to the true quantity one wants to compute. Determining how long we have to run the chain before we feel sufficiently confident that the MCMC algorithm has produced reasonably accurate draws from the distribution is therefore a very important problem. This standard error is usually given by the ratio between the sample standard deviation and square-root of the sample size n as shown below:

$$\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_n)^2}$$

where X_i are the individual draws and

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Length of Runs, Number of Runs

Without further investigation, one has no idea how many iterations are required to achieve a good mixing of the chain. A good indicator time-series plots (e.g. Fig. A.3) which give valuable information about mixing.

The phenomenon of pseudo-covariance brought up the idea of comparing multiple runs of the same sampler started at different points in state space. If the multiple runs appear to converge to the same distribution, then – according to the multistart heuristic – one can trust the result. However, this assumes that each part of the state space contains at least one starting point. If this cannot be guaranteed, the multistart heuristic is worse than useless, as it can give confidence that all is well while in fact your results are completely erroneous.

Burn-in

Burn-in is a term that describes the practice of neglecting some iterations at the beginning of a MCMC run, and also refers to the iterations being thrown away. The Markov chain is executed for n steps (the burn-in period) during which all data is thrown away. After the burn-in the chain is running as described above, using each iteration in the subsequent MCMC calculations. Fig. x illustrates the issue that burn-in addresses. In this figure, the starting position is chosen far out in the tail of the equilibrium distribution.
The special case of a Markov chain started close to the center of a (symmetric) equilibrium distribution would not require a burn-in.

Run Time

Metropolis (1953) has shown that the chain will reach a stationary state in finite time. However, this finite time does not necessarily mean a practically useful one. His proof also says nothing about how long to run the MCMC to achieve the desired amount of precision. These are two practical issues that need to be addressed in any MCMC analysis: How many iterations should the burn-in take and how long should the chain run after the burn-in. Both issues depend critically on how well the chain explores the state space (mixing).

Stationarity

Reaching stationarity means that the current state is independent of the chain's starting points in state space. This properly implies that if the chain is started with different initial conditions, the chain will eventually end up sampling in the same place, no matter where we started. MCMC algorithms are guaranteed to reach stationarity, but this guarantee says nothing about whether that will occur in any practically amount of time or not. There are a few diagnostic plots being

helpful to reduce the amount of time, as well as being reasonable certain that stationarity is reached.

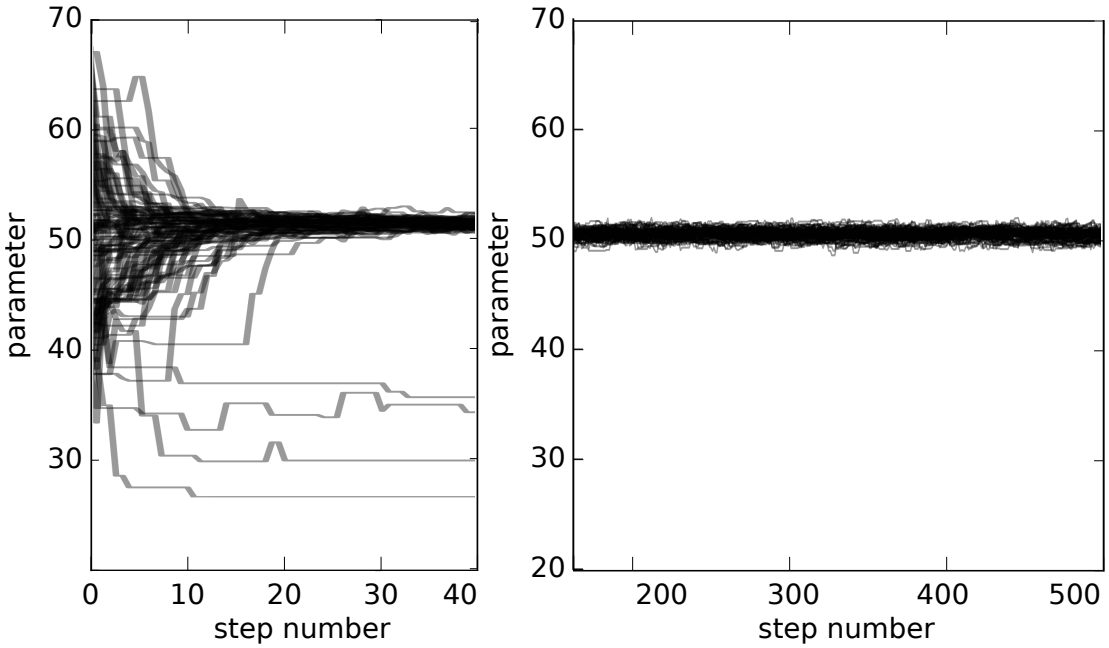
Diagnostic Plots

Stationarity is hard to proof. However, one of the simplest but very informative plots that is able to indicate that stationarity has not been achieved is to simply visualize the state of the Markov chain (the value it takes out of the state space) through iterations. Plotting each parameter as a function of iteration number to produce a time series plot of the chain.

In Fig. A.3, multiple walkers of a Parallel Affine Invariant MCMC Ensemble are shown.
It is quite evident from plot (a), that stationarity has not yet been reached. Each chain of the ensemble quickly diverges from it (b), the burn-in period has been removed, and the chain appears to be stationary.

It is quite evident from Fig. A.3, that in (a) stationary has not been achieved. The chain quickly diverges from its initial point, presumably towards stationary state. In Fig. (b), chains from the same process are shown that where the burn-in period is removed. No trend is evident, good mixing of the chains.

Additionally, visualizing the marginal posterior distributions helps to get a sense of where the parameters are in the state space.



(a) (b)

Figure A.3 Time series plot of a model parameter as the chain progresses. Multiple walkers of a Parallel Affine Invariant MCMC Ensemble are shown.
In plot (a), step zero was the initial point for the parameter. The chain hasn't reached stationarity yet as is evident by their trend. In plot (b), the burn-in period has been removed, and the chain appears to be stationary.

B

Tables

B.1 Expected Selection Completenesses and Purities

These tables give the obtained selection completeness and purity for QSO and RR Lyrae at different magnitude ranges, as obtained in Chapter 5.

Table B.1. Expected Selection Completeness and Purity for QSO

threshold on p_{QSO}	completeness	purity	threshold on p_{QSO}	completeness	purity
0.00	1.00	0.00	0.51	0.52	0.78
0.01	0.96	0.21	0.52	0.51	0.78
0.02	0.94	0.30	0.53	0.50	0.79
0.03	0.93	0.35	0.54	0.49	0.79
0.04	0.92	0.38	0.55	0.48	0.80
0.05	0.91	0.41	0.56	0.47	0.80
0.06	0.90	0.43	0.57	0.46	0.81
0.07	0.89	0.45	0.58	0.45	0.81
0.08	0.89	0.47	0.59	0.44	0.81
0.09	0.88	0.48	**0.60**	**0.43**	**0.82**
0.10	**0.87**	**0.50**	0.61	0.42	0.82
0.11	0.86	0.51	0.62	0.41	0.83
0.12	0.85	0.52	0.63	0.40	0.83
0.13	0.85	0.53	0.64	0.39	0.84
0.14	0.84	0.54	0.65	0.38	0.84
0.15	0.83	0.55	0.66	0.37	0.84
0.16	0.82	0.56	0.67	0.36	0.85
0.17	0.81	0.57	0.68	0.35	0.85
0.18	0.81	0.58	0.69	0.34	0.86
0.19	0.80	0.58	**0.70**	**0.32**	**0.86**
0.20	**0.79**	**0.59**	0.71	0.32	0.86
0.21	0.79	0.60	0.72	0.30	0.86
0.22	0.78	0.61	0.73	0.29	0.87
0.23	0.77	0.61	0.74	0.28	0.87
0.24	0.76	0.62	0.75	0.27	0.88
0.25	0.75	0.63	0.76	0.26	0.88
0.26	0.75	0.63	0.77	0.25	0.88
0.27	0.74	0.64	0.78	0.24	0.89
0.28	0.73	0.65	0.79	0.23	0.89
0.29	0.72	0.65	**0.80**	**0.22**	**0.90**
0.30	**0.71**	**0.66**	0.81	0.20	0.90
0.31	0.71	0.67	0.82	0.19	0.90
0.32	0.70	0.68	0.83	0.18	0.91
0.33	0.69	0.68	0.84	0.17	0.92
0.34	0.68	0.68	0.85	0.15	0.92
0.35	0.67	0.69	0.86	0.14	0.92
0.36	0.66	0.70	0.87	0.13	0.92
0.37	0.65	0.70	0.88	0.12	0.92
0.38	0.65	0.71	0.89	0.10	0.92
0.39	0.64	0.71	**0.90**	**0.09**	**0.93**
0.40	**0.63**	**0.72**	0.91	0.08	0.93
0.41	0.62	0.73	0.92	0.06	0.94
0.42	0.61	0.73	0.93	0.05	0.94
0.43	0.60	0.74	0.94	0.04	0.95
0.44	0.59	0.74	0.95	0.03	0.94
0.45	0.58	0.75	0.96	0.02	0.96
0.46	0.57	0.75	0.97	0.01	0.96
0.47	0.56	0.76	0.98	0.01	0.96
0.48	0.55	0.76	0.99	0.00	0.92
0.49	0.54	0.77	**1.00**	**0.00**	**0.00**
0.50	**0.53**	**0.77**			

Table B.2. Expected Selection Completeness and Purity for QSO within $14.5 < r_{\rm P1} < 20$

threshold on $p_{\rm QSO}$	completeness	purity	threshold on $p_{\rm QSO}$	completeness	purity
0.00	1.00	0.01	0.51	0.79	0.80
0.01	0.99	0.38	0.52	0.78	0.80
0.02	0.98	0.49	0.53	0.77	0.81
0.03	0.98	0.54	0.54	0.78	0.81
0.04	0.97	0.58	0.55	0.76	0.81
0.05	0.97	0.60	0.56	0.75	0.81
0.06	0.97	0.62	0.57	0.74	0.81
0.07	0.96	0.64	0.58	0.73	0.81
0.08	0.96	0.65	0.59	0.72	0.82
0.09	0.96	0.66	**0.60**	**0.71**	**0.82**
0.10	**0.95**	**0.67**	0.61	0.70	0.82
0.11	0.95	0.68	0.62	0.69	0.82
0.12	0.95	0.69	0.63	0.68	0.83
0.13	0.94	0.69	0.64	0.66	0.83
0.14	0.94	0.70	0.65	0.65	0.83
0.15	0.94	0.71	0.66	0.64	0.84
0.16	0.93	0.71	0.67	0.63	0.84
0.17	0.93	0.72	0.68	0.61	0.84
0.18	0.93	0.72	0.69	0.59	0.85
0.19	0.93	0.72	**0.70**	**0.58**	**0.85**
0.20	**0.92**	**0.73**	0.71	0.57	0.85
0.21	0.92	0.73	0.72	0.55	0.85
0.22	0.92	0.73	0.73	0.54	0.86
0.23	0.91	0.74	0.74	0.52	0.86
0.24	0.91	0.74	0.75	0.50	0.86
0.25	0.91	0.74	0.76	0.48	0.86
0.26	0.90	0.75	0.77	0.47	0.87
0.27	0.90	0.75	0.78	0.45	0.87
0.28	0.90	0.75	0.79	0.43	0.87
0.29	0.89	0.75	**0.80**	**0.41**	**0.88**
0.30	**0.89**	**0.75**	0.81	0.39	0.88
0.31	0.88	0.75	0.82	0.36	0.88
0.32	0.88	0.75	0.83	0.34	0.89
0.33	0.88	0.76	0.84	0.32	0.89
0.34	0.87	0.76	0.85	0.30	0.89
0.35	0.87	0.76	0.86	0.28	0.90
0.36	0.86	0.76	0.87	0.25	0.90
0.37	0.86	0.76	0.88	0.23	0.90
0.38	0.85	0.77	0.89	0.20	0.91
0.39	0.85	0.77	**0.90**	**0.18**	**0.92**
0.40	**0.85**	**0.78**	0.91	0.16	0.92
0.41	0.84	0.78	0.92	0.13	0.92
0.42	0.84	0.78	0.93	0.11	0.93
0.43	0.83	0.78	0.94	0.08	0.94
0.44	0.83	0.79	0.95	0.05	0.95
0.45	0.82	0.79	0.96	0.03	0.96
0.46	0.82	0.79	0.97	0.02	0.95
0.47	0.81	0.79	0.98	0.001	1.00
0.48	0.81	0.79	0.99	0.00	1.00
0.49	0.80	0.80	**1.00**	**0.00**	**0.00**
0.50	**0.79**	**0.80**			

Table B.3. Expected Selection Completeness and Purity for RR Lyrae within ∼130 kpc

threshold on p_{RRLyrae}	completeness	purity	threshold on p_{RRLyrae}	completeness	purity
0.00	1.0	0.00	0.51	0.65	0.83
0.01	0.93	0.28	0.52	0.64	0.84
0.02	0.92	0.37	0.53	0.63	0.84
0.03	0.91	0.43	0.54	0.63	0.84
0.04	0.90	0.47	0.55	0.63	0.84
0.05	0.89	0.51	0.56	0.61	0.85
0.06	0.89	0.53	0.57	0.60	0.85
0.07	0.88	0.55	0.58	0.58	0.85
0.08	0.88	0.57	0.59	0.58	0.85
0.09	0.87	0.59	**0.60**	**0.56**	**0.85**
0.10	**0.87**	**0.60**	0.61	0.55	0.85
0.11	0.86	0.62	0.62	0.54	0.85
0.12	0.86	0.63	0.63	0.53	0.85
0.13	0.85	0.64	0.64	0.52	0.85
0.14	0.85	0.65	0.65	0.50	0.85
0.15	0.84	0.66	0.66	0.50	0.86
0.16	0.84	0.67	0.67	0.49	0.86
0.17	0.84	0.68	0.68	0.47	0.86
0.18	0.84	0.69	0.69	0.46	0.86
0.19	0.84	0.70	**0.70**	**0.45**	**0.86**
0.20	**0.83**	**0.70**	0.71	0.43	0.86
0.21	0.82	0.71	0.72	0.42	0.86
0.22	0.82	0.71	0.73	0.40	0.86
0.23	0.82	0.71	0.74	0.38	0.87
0.24	0.81	0.72	0.75	0.36	0.88
0.25	0.81	0.73	0.76	0.35	0.88
0.26	0.81	0.73	0.77	0.33	0.88
0.27	0.80	0.74	0.78	0.32	0.88
0.28	0.80	0.75	0.79	0.31	0.88
0.29	0.79	0.76	**0.80**	**0.29**	**0.88**
0.30	**0.79**	**0.76**	0.81	0.27	0.88
0.31	0.78	0.76	0.82	0.24	0.90
0.32	0.78	0.77	0.83	0.22	0.90
0.33	0.77	0.78	0.84	0.21	0.90
0.34	0.76	0.78	0.85	0.19	0.90
0.35	0.76	0.79	0.86	0.17	0.89
0.36	0.76	0.79	0.87	0.15	0.88
0.37	0.75	0.80	0.88	0.13	0.89
0.38	0.74	0.80	0.89	0.11	0.89
0.39	0.74	0.80	**0.90**	**0.10**	**0.90**
0.40	**0.73**	**0.81**	0.91	0.08	0.90
0.41	0.72	0.81	0.92	0.06	0.94
0.42	0.72	0.81	0.93	0.05	0.92
0.43	0.71	0.81	0.94	0.03	0.89
0.44	0.71	0.81	0.95	0.02	0.83
0.45	0.71	0.82	0.96	0.01	0.82
0.46	0.69	0.82	0.97	0.01	0.80
0.47	0.69	0.83	0.98	0.00	1.00
0.48	0.68	0.83	0.99	0.00	0.00
0.49	0.67	0.83	**1.00**	**0.00**	**0.00**
0.50	**0.66**	**0.83**			

Table B.4. Expected Selection Completeness and Purity for RR Lyrae within ∼40 kpc ($14.5 < r_{\rm P1} < 18.5$)

threshold on $p_{\rm RRLyrae}$	completeness	purity	threshold on $p_{\rm RRLyrae}$	completeness	purity
0.00	1.00	0.00	0.51	0.80	0.86
0.00	0.99	0.42	0.52	0.80	0.86
0.00	0.99	0.50	0.53	0.79	0.87
0.00	0.98	0.55	0.54	0.79	0.87
0.04	0.98	0.59	0.55	0.78	0.87
0.05	0.98	0.60	0.56	0.78	0.87
0.06	0.98	0.62	0.57	0.77	0.87
0.07	0.97	0.64	0.58	0.76	0.87
0.08	0.97	0.65	0.59	0.74	0.87
0.09	0.97	0.67	**0.60**	**0.73**	**0.87**
0.10	**0.96**	**0.69**	0.61	0.71	0.87
0.11	0.95	0.69	0.62	0.70	0.87
0.12	0.95	0.70	0.63	0.69	0.87
0.13	0.95	0.70	0.64	0.68	0.88
0.14	0.94	0.70	0.65	0.67	0.88
0.15	0.94	0.71	0.66	0.66	0.88
0.16	0.94	0.72	0.67	0.65	0.89
0.17	0.94	0.73	0.68	0.63	0.89
0.18	0.94	0.73	0.69	0.61	0.89
0.19	0.93	0.74	**0.70**	**0.60**	**0.88**
0.20	**0.93**	**0.74**	0.71	0.59	0.89
0.21	0.92	0.75	0.72	0.58	0.89
0.22	0.92	0.76	0.73	0.56	0.88
0.23	0.92	0.76	0.74	0.54	0.88
0.24	0.92	0.76	0.75	0.52	0.88
0.25	0.91	0.77	0.76	0.50	0.88
0.26	0.91	0.77	0.77	0.48	0.88
0.27	0.91	0.78	0.78	0.46	0.89
0.28	0.91	0.78	0.79	0.45	0.90
0.29	0.91	0.79	**0.80**	**0.42**	**0.89**
0.30	**0.91**	**0.79**	0.81	0.39	0.89
0.31	0.91	0.79	0.82	0.37	0.89
0.32	0.91	0.80	0.83	0.36	0.91
0.33	0.90	0.80	0.84	0.34	0.91
0.34	0.90	0.80	0.85	0.33	0.93
0.35	0.90	0.81	0.86	0.31	0.93
0.36	0.89	0.82	0.87	0.28	0.95
0.37	0.88	0.82	0.88	0.27	0.95
0.38	0.88	0.83	0.89	0.25	0.96
0.39	0.87	0.83	**0.90**	**0.24**	**0.96**
0.40	**0.87**	**0.83**	0.91	0.21	0.97
0.41	0.86	0.83	0.92	0.18	0.97
0.42	0.85	0.83	0.93	0.14	0.96
0.43	0.85	0.84	0.94	0.11	0.97
0.44	0.84	0.84	0.95	0.08	0.97
0.45	0.84	0.84	0.96	0.06	1.00
0.46	0.83	0.84	0.97	0.04	1.00
0.47	0.82	0.84	0.98	0.02	1.00
0.48	0.82	0.85	0.99	0.00	0.00
0.49	0.81	0.85	**1.00**	**0.00**	**0.00**
0.50	**0.80**	**0.85**			

Table B.5. Expected Selection Completeness and Purity for RR Lyrae within ∼80 kpc ($19.7 < r_{\mathrm{P1}} < 20.7$)

threshold on p_{RRLyrae}	completeness	purity	threshold on p_{RRLyrae}	completeness	purity
0.00	1.00	0.00	0.51	0.71	0.85
0.01	0.95	0.26	0.52	0.70	0.86
0.02	0.94	0.36	0.53	0.69	0.86
0.03	0.91	0.40	0.54	0.67	0.87
0.04	0.90	0.45	0.55	0.66	0.87
0.05	0.89	0.49	0.56	0.66	0.87
0.06	0.88	0.52	0.57	0.64	0.87
0.07	0.88	0.54	0.58	0.64	0.87
0.08	0.86	0.57	0.59	0.64	0.88
0.09	0.85	0.59	**0.60**	**0.63**	**0.88**
0.10	**0.85**	**0.61**	0.61	0.62	0.88
0.11	0.84	0.64	0.62	0.62	0.88
0.12	0.84	0.65	0.63	0.60	0.88
0.13	0.83	0.66	0.64	0.60	0.89
0.14	0.83	0.68	0.65	0.58	0.88
0.15	0.83	0.69	0.66	0.58	0.90
0.16	0.83	0.70	0.67	0.57	0.90
0.17	0.82	0.71	0.68	0.57	0.90
0.18	0.82	0.72	0.69	0.55	0.90
0.19	0.82	0.73	**0.70**	**0.54**	**0.90**
0.20	**0.81**	**0.74**	0.71	0.54	0.90
0.21	0.81	0.75	0.72	0.53	0.90
0.22	0.81	0.76	0.73	0.52	0.90
0.23	0.80	0.77	0.74	0.50	0.90
0.24	0.80	0.78	0.75	0.50	0.90
0.25	0.80	0.79	0.76	0.48	0.91
0.26	0.80	0.79	0.77	0.46	0.91
0.27	0.80	0.80	0.78	0.45	0.91
0.28	0.80	0.80	0.79	0.43	0.91
0.29	0.80	0.80	**0.80**	**0.43**	**0.91**
0.30	**0.80**	**0.80**	0.81	0.42	0.90
0.31	0.80	0.81	0.82	0.40	0.90
0.32	0.80	0.81	0.83	0.40	0.91
0.33	0.80	0.81	0.84	0.39	0.91
0.34	0.77	0.81	0.85	0.38	0.92
0.35	0.76	0.81	0.86	0.33	0.93
0.36	0.76	0.82	0.87	0.31	0.93
0.37	0.76	0.82	0.88	0.30	0.93
0.38	0.76	0.82	0.89	0.27	0.92
0.39	0.75	0.83	**0.90**	**0.25**	**0.92**
0.40	**0.75**	**0.83**	0.91	0.23	0.92
0.41	0.75	0.83	0.92	0.19	0.90
0.42	0.75	0.84	0.93	0.17	0.91
0.43	0.74	0.84	0.94	0.13	0.89
0.44	0.74	0.85	0.95	0.10	0.89
0.45	0.73	0.84	0.96	0.07	0.88
0.46	0.73	0.85	0.97	0.04	0.86
0.47	0.73	0.85	0.98	0.02	0.88
0.48	0.72	0.85	0.99	0.01	1.00
0.49	0.73	0.85	**1.00**	**0.00**	**0.00**
0.50	**0.71**	**0.85**			

B.2 Sagittarius Stream

These tables give the geometry of the Sagittarius stream, represented by its extent and width as inferred from the analysis in Chapter 6.

Table B.6. Fitted Parameters for Sagittarius Stream, Leading Arm

$\tilde{\Lambda}_\odot$ [°]	$f_{\rm sgr}$ [a]	$D_{\rm sgr}$ [kpc][b]	$D_{\rm sgr}$ 2σ interval [kpc]	$\sigma_{\rm sgr}$ [kpc][c]	$\sigma_{\rm sgr}$ 2σ interval [kpc]
10	0.05315	13.062	[5.164, 26.718]	2.588	[1.039, 5.702]
20	0.05039	39.084	[34.246, 41.965]	1.706	[1.017, 5.206]
30	0.05119	41.628	[39.326, 43.656]	2.125	[1.162, 3.799]
40	0.37902	45.274	[44.742, 45.771]	3.615	[3.246, 4.050]
50	0.51720	49.858	[49.534, 50.169]	3.319	[2.966, 3.686]
60	0.61343	51.828	[51.406, 52.262]	4.475	[3.894, 5.026]
70	0.40678	48.549	[48.024, 49.061]	3.659	[3.078, 4.435]
80	0.35617	45.562	[44.775, 46.326]	4.634	[3.861, 5.535]
90	0.21225	40.088	[39.040, 41.173]	4.022	[2.738, 5.566]
100	0.22411	35.702	[33.760, 37.238]	4.712	[3.009, 5.930]
110	0.22507	30.744	[29.218, 31.830]	3.245	[2.274, 5.017]
120	0.21835	26.364	[22.988, 28.630]	4.246	[2.374, 5.907]
130	0.06867	21.364	[20.056, 28.786]	3.308	[1.157, 5.865]
140	0.13778	19.368	[16.743, 39.772]	2.382	[1.043, 5.029]
150	0.17783	16.329	[15.058, 19.319]	3.929	[1.985, 5.637]
160	0.36354	87.929	[86.482, 89.264]	4.683	[3.479, 5.846]

[a]fraction sources in Sgr stream

[b]mean heliocentric Sgr stream distance

[c]Sgr stream line-of-sight width

Table B.7. Fitted Parameters for Sagittarius Stream, Trailing Arm

$\tilde{\Lambda}_\odot$ [°]	$f_{\rm sgr}$ [a]	$D_{\rm sgr}$ [kpc][b]	$D_{\rm sgr}$ 2σ interval [kpc]	$\sigma_{\rm sgr}$ [kpc][c]	$\sigma_{\rm sgr}$ 2σ interval [kpc]
100	0.05428	55.155	[51.360, 57.960]	2.858	[1.236, 5.764]
110	0.05463	64.423	[54.662, 72.917]	3.421	[1.055, 5.873]
120	0.05886	56.539	[52.621, 66.671]	2.610	[1.066, 5.743]
130	0.07754	66.502	[62.112, 71.374]	4.583	[1.512, 5.935]
140	0.09194	79.711	[71.569, 84.388]	5.076	[2.051, 5.970]
150	0.30275	81.575	[79.894, 83.197]	5.255	[4.073, 5.961]
160	0.36581	87.959	[86.625, 89.198]	4.714	[3.534, 5.822]
170	0.52524	92.008	[90.682, 93.355]	5.902	[5.490, 5.996]
180	0.32859	87.550	[85.294, 89.928]	5.591	[4.431, 5.984]
190	0.08368	59.629	[42.281, 71.322]	3.051	[1.070, 5.883]
200	0.52845	53.135	[51.472, 54.817]	5.668	[4.840, 5.984]
210	0.56103	43.055	[41.427, 44.669]	5.634	[4.577, 5.984]
220	0.69129	36.552	[35.313, 37.770]	5.752	[5.061, 5.988]
230	0.53671	31.050	[30.087, 32.572]	5.630	[4.728, 5.984]
240	0.56661	27.889	[26.354, 29.187]	4.690	[3.621, 5.841]
250	0.64931	25.148	[23.634, 26.438]	5.089	[4.093, 5.937]
260	0.42848	24.436	[22.627, 25.939]	4.837	[3.288, 5.916]
270	0.49886	20.295	[18.264, 22.408]	5.400	[3.753, 5.979]
280	0.32116	20.032	[17.885, 48.008]	4.388	[3.065, 5.806]
290	0.26776	21.107	[18.780, 23.070]	4.533	[2.536, 5.925]
300	0.36671	20.730	[18.747, 22.387]	4.987	[3.569, 5.941]
310	0.45996	21.174	[19.509, 22.636]	4.924	[3.750, 5.922]
320	0.49185	21.391	[19.767, 22.747]	4.804	[3.688, 5.888]
330	0.37186	20.075	[17.979, 22.012]	5.100	[3.488, 5.954]
340	0.44994	20.701	[19.321, 22.260]	5.629	[4.779, 5.984]
350	0.46107	27.089	[26.926, 56.072]	1.254	[1.051, 3.732]

[a]fraction sources in Sgr stream

[b]mean heliocentric Sgr stream distance

[c]Sgr stream line-of-sight width

Table B.8. Possibly Sagittarius Stream Bifurcation

α [°] [a]	δ [°] [b]	$f_{\rm sgr}$ [c]	$D_{\rm sgr}$ [kpc][d]	$D_{\rm sgr}$ 2σ interval [kpc]	$\sigma_{\rm sgr}$ [kpc][e]	$\sigma_{\rm sgr}$ 2σ interval [kpc]
215	5	0.45309	49.02	[48.446, 49.620]	2.970	[2.379, 3.805]
204.783	8.391	0.37612	45.800	[44.847, 46.714]	4.500	[3.598, 5.474]
189.524	8.333	0.26316	40.513	[38.800, 42.410]	4.821	[2.054, 5.933]
189.444	15.667	0.16247	35.414	[32.233, 38.279]	4.877	[2.494, 5.957]
169.63	12.333	0.1954	26.075	[12.705, 34.648]	5.265	[1.962, 5.977]
170.256	22.641	0.20342	14.897	[11.666, 29.689]	4.161	[1.425, 5.960]
150.556	13.972	0.30127	20.563	[15.400, 24.890]	5.29	[2.369, 5.980]
149.841	26.27	0.19247	16.365	[12.347, 20.292]	5.00	[2.170, 5.956]

[a b]for each polygon, the centroid of its (α, δ) is given, as used in Fig. 6.6.

[c]fraction sources in Sgr stream

[d]mean heliocentric Sgr stream distance

[e]Sgr stream line-of-sight width

C

Bibliography

Abbas, M. A., Grebel, E. K., Martin, N. F., et al. (2014a). An Optimized Method to Identify RR Lyrae Stars in the SDSS × Pan-STARRS1 Overlapping Area Using a Bayesian Generative Technique. *AJ*, 148(1):8.

Abbas, M. A., Grebel, E. K., Martin, N. F., et al. (2014b). Newly discovered RR Lyrae stars in the SDSS-Pan-STARRS1-Catalina footprint. *MNRAS*, 41:1230.

Alcock, C., Allsman, R. A., Alvens, D. R., et al. (2001). The MACHO Project: Microlensing Detection Efficiency. *ApJS*, 136(2):439.

Aldering, G., Adam, G., Antilogus, P., et al. (2002). Overview of the Nearby Supernova Factory. In Tyson, A. and Wolff, S., editors, *Survey and Other Telescope Technologies and Discoveries*, volume 4836. Proceedings of the SPIE.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545.

Andersen, J. (1991). Accurate masses and radii of normal stars. *The Astronomy and Astrophysics Review*, 3:91.

Aretxaga, I., Fernandes, R. C., and Terlevich, R. J. (1997). QSO variability: probing the Starburst model. *MNRAS*, 286:271.

Arnett, D. (1996). *Supernovae and Nucleosynthesis: An Investigation of the History of Matter, from the Big Bang to the Present*. Princeton University Press.

Atkinson, R. (1931). Atomic synthesis and stellar energy. *ApJ*, 73:250.

Auvergne, M., Bodin, P., Boisnard, L., et al. (2008). The CoRoT satellite in flight: description and performance. *A&A*, 506(1):411.

Bailey, S., Aragon, C., Romano, R., et al. (2006). How to Find More Supernovae with Less Work: Object Classification Techniques for Difference Imaging. *ApJ*, 665(2):1246.

Bailey, S. I. (1902). A discussion of variable stars in the cluster ω Centauri. *Annals of Harvard College Observatory*, 38:1.

Bailey, S. I. and Pickering, E. C. (1913). Variable stars in the cluster Messier 3. *Annals of Harvard College Observatory*, 78:1.

Baker, N. and Kippenhahn, R. (1962). The pulsations of models of δ Cephei stars. *Zeitschrift für Astrophysik*, 54:11.

Ball, N. M., Brunner, R. J., Myers, A. D., et al. (2006). Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. *ApJ*, 650(1):497.

Behr, B. B., Cohen, J. G., and McCarthy, J. K. (2000). Rotations and Abundances of Blue Horizontal-Branch Stars in Globular Cluster M15. *ApJ*, 531:L37.

Beichman, C. A., Neugebauer, G., Habing, H., et al. (1986). IRAS Catalogs and Atlases: Explanatory Supplement. *NASA RP-1190*.

Bellm, E. (2014). The Zwicky Transient Facility. In Wozniak, P. et al., editors, *The Third Hot-wiring the Transient Universe Workshop (HTU-III)*, page 27.

Belokurov, V., Koposov, S. E., and Evans, N. W. (2014). Precession of the Sagittarius stream. *MNRAS*, 437(1):116.

Belokurov, V., Zucker, D. B., Evans, N. W., et al. (2006). A Faint New Milky Way Satellite in Bootes. *ApJ*, 647(2):L111.

Benedict, G. F., McArthur, B. E., Feast, M. W., et al. (2011). Distance scale zero points from the Galactic RR Lyrae star parallaxes. *ApJ*, 142(6):187.

Berk, D. E. V., Wilhite, B. C., Kron, R. G., et al. (2004). The Ensemble Photometric Variability of ∼25,000 Quasars in the Sloan Digital Sky Survey. *ApJ*, 601(2):692.

Bernard, S., Heutte, L., and Adam, S. (2009). Influence of Hyperparameters on Random Forest Accuracy. In Benediktsson, J. A., Kittler, J., and Roli, F., editors, *Lecture Notes in Computer Science*, volume 5519. Springer.

Bethe, H. A. (1939). Energy production in stars. *Physical Review*, 55:434.

Bethe, H. A. and Marshak, R. E. (1939). The physics of stellar interiors and stellar evolution. *Rep. Prog. Phys*, 6:1.

Blažko, S. (1907). Mittelungen über veränderliche Sterne. *Astronomische Nachrichten*, 175:325.

Boberg, O. W., E. D. Friel, E. V., et al. (2015). Chemical Abundances in NGC 5053: A Very Metal-poor and Dynamically Complex Globular Cluster. *ApJ*, 804(2):12.

Boccaletti, A., Lagage, P.-O., Baudoz, P., et al. (2015). The Mid-Infrared Instrument for the James Webb Space Telescope, V: Predicted Performance of the MIRI Coronagraphs. *PASP*, 127(953):633.

Bodenheimer, P. (2003). The basic physics of star formation. In Genorio-Tagle, G., Prieto, M., and Sánchez, F., editors, *Star Formation in Stellar Systems*, page 1. Cambridge University Press.

Bonanos, A. Z., Stanek, K. Z., Szentgyorgyi, A. H., et al. (2004). The RR Lyrae Distance to the Draco Dwarf Spheroidal Galaxy. *AJ*, 127(2):861.

Borucki, W. J., Koch, D., Basri, G., et al. (2010). Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327(5968):977.

Breiman, L. (1996). Bagging predictors. *Machine Learning archive*, 24(2):123.

Breiman, L. (1998). Randomizing Outputs To Increase Prediction Accuracy. *Technical Report Statistics Department, UCB*, 518.

Breiman, L. (1999). Using adaptive bagging to debias regressions. *Technical Report Statistics Department, UCB*, 547.

Breiman, L. (2001). Random Forests. *Machine Learning archive*, 45(1):5.

Breiman, L., Friedman, J. H., and Olshen, R. A. (1984). *Classification and regression trees*. Wadsworth, Belmont, CA.

Breiman, L., Olshen, R. A., and Stone, C. J. (1983). *CART: Classification and Regression Trees*. Wadsworth, Belmont, CA.

Brocato, E., Castellani, V., and Piersimoni, A. (1997). The Age of the Globular Cluster M68. *ApJ*, 491(2):789.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC.

Budavári, T. and Lubow, S. H. (2012). Catalog Matching with Astrometric Correction and its Application to the Hubble Legacy Archive. *ApJ*, 188(2):10.

Bullock, J. S. and Johnston, K. V. (2005). Tracing Galaxy Formation with Stellar Halos. I. Methods. *ApJ*, 635(2):931.

Buonanno, R., Buscema, G., Corsi, C. E., et al. (1984). Positions, magnitudes and colors for stars in the globular cluster NGC 5466. *A&A Supp. Ser.*, 5:79.

Cacciari, C. and Clementini, G. (2003). Globular cluster distances from rr lyrae stars. In Alloin, D. and Oieren, W., editors, *Lecture Notes in Physics, Stellar Candles for the Extragalactic Distance Scale*, volume 635, page 105. Springer.

Cáceres, C. and Catalan, M. (2008). The Period-Luminosity Relation of RR Lyrae Stars in the SDSS Photometric System. *ApJS*, 179:243.

Casetti-Dinescu, D. I., Nusde, D. A., Girard, T. M., et al. (2015). A Kinematically-Distinct RR-Lyrae Overdensity in the Inner Regions of the Milky Way. *ApJ*, 810(1):L4.

Catelan, M. (2004). The evolutionary status of M3 RR Lyrae variable stars: breakdown of the canonical framework? *ApJ*, 600(1):409.

Catelan, M., Minniti, D., Lucas, P. W., et al. (2011). The Vista Variables in the Vía Láctea (VVV) ESO Public Survey: Current Status and First Results. In McWilliam, A., editor, *RR Lyrae Stars, Metal-Poor Stars, and the Galaxy*, volume 5, page 145. Carnegie Observatories

Catelan, M., Pritzl, B. J., and Smith, H. A. (2004). The RR Lyrae Period-Luminosity Relation. I. Theoretical Calibration. *ApJS*, 154:633.

Catelan, M. and Smith, H. A. (2015). *Pulsating stars*. Wiley-VCH.

Chaboyer, B. (1999). Globular cluster distance determinations. In Heck, A. and Caputo, F., editors, *Post-Hipparcos Standard Candles, Astrophysics and Space Science Library*, volume 237, page 111. Dordrecht.

Chambers, K. (2011). The First year of the Pan-STARRS 1 System: Surveys, Cadences, Data Products, and Performance. *American Astronomical Society Meeting Abstracts*, 218:113.01.

Charbonneau, D., Allen, L. E., Megeath, S. T., et al. (2005). Detection of Thermal Emission from an Extrasolar Planet. *ApJ*, 626(1):523.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

Christy, R. F. (1966). A study of pulsation in RR Lyrae models. *ApJ*, 144:108.

Clement, C. M., Muzzin, A., Dufton, Q., et al. (2001). Variable Stars in Galactic Globular Clusters. *ApJ*, 112(5):2587.

Clerc, N., Merloni, A., Zhang, Y.-Y., et al. (2016). SPIDERS: the spectroscopic follow-up of X-ray selected clusters of galaxies in SDSS-IV. *MNRAS*, doi: 10.1093/mnras/stw2214:accepted.

Coad, S. (2012). Course Material for Time Series.

Cole, N., Newberg, H. J., Magdon-Ismail, M., et al. (2008). Maximum Likelihood Fitting of Tidal Streams with Application to the Sagittarius Dwarf Tidal Tails. *ApJ*, 683(2):750.

Collier, S. and Peterson, B. M. (2001). Characteristic Ultraviolet/Optical Timescales in Active Galactic Nuclei. *ApJ*, 555(2):775.

Cordes, J. M. and Downs, G. S. (1985). JPL pulsar timing observations. III - Pulsar rotation fluctuations. *ApJS*, 59:343.

Cox, A. N. (1998). Theoretical period changes in yellow giant pulsars. *ApJ*, 496(1):246.

Cox, A. N., Morgan, S. M., Rogers, F. J., et al. (1992). An opacity mechanism for the pulsations of OB stars. *ApJ*, 393(1):272.

Cox, J. P. (1960). A preliminary analysis of the effectiveness of second helium ionization in inducing Cepheid instability in stars. *ApJ*, 132:594.

Cox, J. P. (1963). On second helium ionization as a cause of pulsation instability in stars. *ApJ*, 138:487.

Cox, J. P. and Whitney, C. (1958). Stellar Pulsation. IV. A semitheoretical period-luminosity relation for classical Cepheids. *ApJ*, 127:561.

Cutler, A. and Zhao, G. (2001). PERT - perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *ICML '06 Proceedings of the 23rd international conference on Machine learning*. ACM New York.

Deason, A. J., Belokurov, V., and Evans, N. W. (2001). The Milky Way stellar halo out to 40 kpc: Squashed, broken but smooth. *MNRAS*, 416(4):2903.

Deb, S. and Singh, H. P. (2010). Physical parameters of the Small Magellanic Cloud RR Lyrae stars and the distance scale. *MNRAS*, 402(1):691.

Debosscher, J., Sarro, L. M., Lózep, M., et al. (2009). Automated supervised classification of variable stars in the CoRoT programme. *A&A*, 506:519.

Dékány, I., Minniti, D., Majaess, D., et al. (2015). The VVV Survey Reveals Classical Cepheids Tracing a Young and Thin Stellar Disk across the Galaxy's Bulge. *ApJL*, 812(2):L29.

Deming, D. and Seager, S. (2005). Detection of Infrared Radiation from an Extrasolar Planet. *Nature*, 434(7034):740.

Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization. *Machine Learning*, 40(2):139.

Downes, R. A., Margon, B., Homer, L., et al. (2004). Far-Ultraviolet Observations of RR Lyrae Stars in the Core of NGC 1851. *ApJ*, 128(5):2288.

Drake, A. J., Catelan, M., Djorgovski, S. G., et al. (2013a). Evidence for a Milky Way Tidal Stream Reaching Beyond 100 kpc. *ApJ*, 765(2):154.

Drake, A. J., Catelan, M., Djorgovski, S. G., et al. (2013b). Probing the Outer Galactic Halo with RR Lyrae from the Catalina Surveys. *ApJ*, 763(1):32.

Drake, A. J., Djorgovski, S. G., Mahabal, A., et al. (2009). First Results from the Catalina Real-time Transient Survey. *ApJ*, 696(1):870.

Eddington, A. S. (1926). *The Internal Constitution of the Stars*. Cambridge University Press.

Edelson, R. A., Alexander, T., Crenshaw, D. M., et al. (1996). Multiwavelength Observations of Short-Timescale Variability in NGC 4151. IV. Analysis of Multiwavelength Continuum Variability. *ApJ*, 470:364.

Ekström, S., Georgy, C., Eggenberger, P., et al. (2012). Grids of stellar models with rotation. I. Models from 0.8 to 120 $M_\odot$ at solar metallicity (Z=0.014). *A&A*, 537:A146.

Eyer, L. and Blake, C. (2004). Automated classification of variable stars for All-Sky Automated Survey 1-2 data. *MNRAS*, 358:30.

Eyer, L. and Mowlavi, N. (2008). Variable stars across the observational HR diagram. *J.Phys.Conf.Ser.*, 012010:118.

Falk, M. (2012). A First Course on Time Series Analysis.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861.

Fellhauer, M., Belokurov, V., Evans, N. W., et al. (2006). The Origin of the Bifurcation in the Sagittarius Stream. *ApJ*, 651:167.

Ferrarese, L., Pogge, R. W., and Peterson, B. M. (2001). Supermassive Black Holes in Active Galactic Nuclei. I. The Consistency of Black Hole Masses in Quiescent and Active Galaxies. *ApJ*, 555(2):L79.

Fey, A. L., Gordon, D., Jacobs, C. S., et al. (2015). The Second Realization of the International Celestial Reference Frame by Very Long Baseline Interferometry. *ApJ*, 150(2):58.

Figuera, R. J. (2011). Untitled. In Henney, J. W. and Torres-Peimbert, S., editors, *XIII Latin American Regional IAU Meeting*, volume 40, page 235. Revista Mexicana de Astronomía y Astrofísica (Serie de Conferencias).

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2012). emcee: The MCMC Hammer.

Foreman-Mackey, D., Price-Whelan, A., Ryan, G., et al. (2013). triangle.py v0.0.1,. Zenodo, doi:10.5281/zenodo.11020.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29:1189.

Gamov, G. (1939). Physical possibilities of stellar evolution. *Physical Review*, 55:718.

Gautschy, A. and Saio, H. (1996). Stellar pulsations across the HR diagram: Part 2. *Annual Review of Astronomy and Astrophysics*, 34:551.

Gelfand, E. A. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 685:398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gezari, S., Chornock, R., Rest, A., et al. (2012). An ultraviolet-optical flare from the tidal disruption of a helium-rich stellar core. *Nature*, 485:217.

Gibbons, S. L. J., Belokurov, V., and Evans, N. W. (2014). 'Skinny Milky Way please', says Sagittarius. *MNRAS*, 445:3788.

Gini, C. (1913). Variabilita e Mutabilita. *Journal of the Royal Statistical Society*, 76(3):326–327.

Goodman, J. and Weare, J. (2010). Ensemble Samplers with Affine Invariance. *Communications in Applied Mathematics and Computational Science*, pages 5–65.

Gould, A., Flynn, C., and Bahcall, J. N. (1998). Spheroid Luminosity and Mass Functions from Hubble Space Telescope Star Counts. *ApJ*, 503(2):798.

Graham, M. J., Drake, A. J., Djorgovski, S. G., et al. (2013). A comparison of period finding algorithms. *MNRAS*, 434(4):3423.

Grillmair, C. J. (2009). Four New Stellar Debris Streams in the Galactic Halo. *ApJ*, 693(2):1118.

Grillmair, C. J. and Carlin, J. L. (2016). Stellar Streams and Clouds in the Galactic Halo. In Newberg, H. J. and Carlin, J. L., editors, *Tidal Streams in the Local Group and Beyond, Astro-*

physics and Space Science Library, volume 420. Springer International Publishing Switzerland.

Grillmair, C. J. and Johnson, R. (2006). The Detection of a 45° Tidal Stream Associated with the Globular Cluster NGC 5466. *ApJ*, 639(1):L17.

Harris, W. E. (1996). A Catalog of Parameters for Globular Clusters in the Milky Way (2010 edition). *AJ*, 112:1487.

Hartwick, F. D. A. and Schade, D. (1990). The Space Distribution of Quasars. *ARA&A*, 28:437.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications.

Hawkins, M. R. S. (1996). Dark-Matter from Quasar Microlensing. *MNRAS*, 278:787.

Hernitschek, N., Rix, H.-W., Bovy, J., et al. (2016). Finding, characterizing and classifying variable sources in multi-epoch sky surveys: QSOs and RR Lyrae in PS1 3π data. *ApJ*, 817(1):73.

Hernitschek, N., Rix, H.-W., and Morganson, E. (2015). Estimating Black Hole Masses in Hundreds of Quasars. *ApJ*, 801(1):45.

Hertzsprung, E. (1911). Über die Verwendung Photographischer Effektiver Wellenlängen zur Bestimmung von Farbenäquivalenten. *Publ. Astrophys. Observ. Potsdam*, 22:1.

Hessels, J. W. T., Ransom, S. M., Stairs, I. H., et al. (2007). A 1.4 GHz Arecibo Survey for Pulsars in Globular Clusters. *ApJ*, 670(1):363.

Hillebrandt, W. and Niemeyer, J. C. (2000). Type Ia Supernova Explosion Models. *Annual Review of Astronomy and Astrophysics*, 38(1):191.

Hjellming, R. M. and Narayan, R. (1986). Refractive interstellar scintillation in 1741-038. *ApJ*, 310(1):768.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(8):832.

Hockey, T. (2009). *The Biographical Encyclopedia of Astronomers*. Springer Publishing.

Hodapp, K. W., Kaiser, N., Aussel, H., et al. (2004). Design of the Pan-STARRS telescopes. *Astronomical Notes*, 325(6):636.

Hoffmeister, C., Richter, G., and Wenzel, W. (1985). *Variable Stars*. Springer Publishing.

Hopkins, A. M. and Beacom, J. F. (2006). On the Normalization of the Cosmic Star Formation History. *ApJ*, 651(1):142.

Hopkins, A. M., Rao, S. M., and Turnshek, D. A. (2005). The Star Formation History of Damped Lyman Alpha Absorbers. *ApJ*, 630(1):108.

Howell, S. B., Sobeck, C., Haas, M., et al. (2014). The K2 Mission: Characterization and Early Results. *PASP*, 126(938):398.

Hughes, P. A., Aller, H. D., and Aller, M. F. (1992). The University of Michigan radio astronomy data base. I - Structure function analysis and the relation between BL Lacertae objects and

quasi-stellar objects. *ApJ*, 396(2):469.

Ibata, R. A., Gilmore, G., and Irwin, M. J. (1994). A dwarf satellite galaxy in Sagittarius. *Nature*, 370(6486):194.

Irwin, M. J., Bunclark, P. S., Bridgeland, M. T., et al. (1990). A new satellite galaxy of the Milky Way in the constellation of Sextans. *MNEAS*, 244:16.

Ivezić, Ž., Sesar, B., Jurić, M., et al. (2008). The Milky Way Tomography with SDSS. II. Stellar Metallicity. *AJ*, 684(1):287.

Ivezić, Ž. and "the LSST Science Council" (2011). The LSST System Science Requirements Document, v5.2.3.

Ivezić, Ž., Vivas, A. K., Lupton, R. H., et al. (2005). The Selection of RR Lyrae Stars Using Single-Epoch Data. *AJ*, 129(2):1096.

Jeans, J. H. (1902). The Stability of a Spherical Nebula. *Philosophical Transactions of the Royal Society*, 199:1.

Johnston, K. V., Hernquist, L., and Bolte, M. (1996). Fossil Signatures of Ancient Accretion Events in the Halo. *ApJ*, 465:278.

Jurić, M., Ivezić, Ž., Brooks, A., et al. (2008). The Milky Way Tomography with SDSS. I. Stellar Number Density Distribution. *AJ*, 673(2):864.

Jurić, M., Ivezić, Ž., Brooks, A., et al. (2011). Large Survey Database: A Distributed Framework for Storage and Analysis of Large Datasets. *AAS Meeting #217, Bulletin of the American Astronomical Society*, 43:433.19.

Jurić, M., Kantor, J., Lim, K.-T., et al. (2015). The LSST Data Management System.

Kaiser, N., W, W. B., Chambers, K., et al. (2010). *The Pan-STARRS wide-field optical/NIR imaging survey*. Proc. SPIE.

Kelly, B. C., Bechtold, J., and Siemiginowska, A. (2009). Are the variations in quasar optical flux driven by thermal fluctuations? *ApJ*, 698(1):895.

Kinemuchi, K., Smith, H. C., Harris, H. A., et al. (2008). The Variable Stars of the Draco Dwarf Spheroidal Galaxy: Revised. *AJ*, 136(5):1921.

Kippenhahn, R. and Weigert, A. (1990). *Stellar Structure and Evolution*. Springer Publishing.

Koch, A. and McWilliam, A. (2014). The chemical composition of a regular halo globular cluster: NGC 5897. *A&A*, 565:13.

Kolmogorov, A. N. (1941a). On degeneration of isotropic turbulence in an incompressible viscous liquid. *Dokl. Akad. Nauk SSSR*, 31:538.

Kolmogorov, A. N. (1941b). The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Dokl. Akad. Nauk SSSR*, 30:301.

Koopmann, R. A., Lee, Y.-W., Demarque, P., et al. (1994). Mass loss during the RR Lyrae phase of the horizontal branch: mass dispersion on the horizontal branch and RR Lyrae period changes. *ApJ*, 423:380.

Koposov, S. E., Rix, H.-W., and Hogg, D. W. (2010). Constraining the Milky Way Potential with a Six-Dimensional Phase-Space Map of the GD-1 Stellar Stream. *ApJ*, 712(1):260.

Korhonen, H., Berdyugina, S., Hackman, T., et al. (1999). Study of FK Comae Berenices, I. Surface images for 1994 and 1995. *A&A*, 346:101.

Kormendy, J., Bender, R., Ajhar, E. A., et al. (1996a). Hubble Space Telescope Spectroscopic Evidence for a 1×10^9 $M_\odot$ Black Hole in NGC 4594. *ApJL*, 473:L91.

Kormendy, J., Bender, R., Richstone, D., et al. (1996b). Hubble Space Telescope Spectroscopic Evidence for a 2×10^9 $M_\odot$ Black Hole in NGC 3115. *ApJL*, 459:L57.

Kozłowski, S., Kochanek, C. S., Udalski, A., et al. (2010). Quantifying Quasar Variability As Part of a General Approach To Classifying Continuously Varying Sources. *ApJ*, 708(2):927.

Krolik, J. H. (1999). *Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment*. Princeton University Press.

Küppers, M., Mottola, S., Lowry, S. C., et al. (2007). Determination of the light curve of the Rosetta target asteroid (2867) Steins by the OSIRIS cameras onboard Rosetta. *A&A*, 462(1):L13.

L. C. Ho and A. V. Filippenko and W. L. Sargent (1995). A search for 'dwarf' seyfert nuclei. 2: an optical spectral atlas of the nuclei of nearby galaxies. *ApJS*, 98(2):477.

Law, D. R., Johnston, K. V., and Majewski, S. R. (2005). A 2MASS All-Sky View of the Sagittarius Dwarf Galaxy: IV. Modeling the Sagittarius Tidal Tails. *ApJ*, 619(2):807.

Law, D. R. and Majewski, S. R. (2010). The Sagittarius Dwarf Galaxy: a Model for Evolution in a Triaxial Milky Way Halo. *ApJ*, 714(1):229.

Law, N. M., Kularni, S. R., Richard, G. D., et al. (2009). The Palomar Transient Factory: System Overview, Performance, and First Results. *Publications of the Astronomical Society of the Pacific*, 121(886):1395.

Lawrence, A., Warren, S. J., Almaini, O., et al. (2007). The UKIRT Infrared Deep Sky Survey (UKIDSS). *MNRAS*, 379(4):1599.

Laycock, S., Tang, S., Grindlay, J., et al. (2010). Digital Access to a Sky Century at Harvard: Initial Photometry and Astrometry. *AAS*, 140:4.

Layden, A. C., Hanson, R. B., Hawley, S. L., et al. (1996). The Absolute Magnitude and Kinematics of RR Lyrae Stars via Statistical Parallax. *AJ*, 112(5):2110.

Leavitt, H. S. and Pickering, E. C. (1912). Periods of 25 variable stars in the Small Magellanic Cloud. *Harvard College Observatory Circular*, 173:1.

Ledoux, P. (1947). Stellar models with convection and with discontinuity of the mean molecular weight. *ApJ*, 105:305.

Lee, J.-W. and Carney, B. W. (1999). BV Photometry of RR Lyrae Variables in the Globular Cluster M2 (NGC 7089). *AJ*, 117(6):2868.

Leighton, R. B. (1960). Untitled. In Thomas, R. N., editor, *Aerodynamic Pheonomena in Stellar Atmospheres, IAU Symposium*, number 12, page 321.

Leighton, R. B., Noyes, R. W., and Simon, G. W. (1962). Velocity fields in the solar atmosphere. I. Preliminary report. *ApJ*, 135:474.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447.

Longmore, A. J., Fernley, J. A., and Jameson, R. F. (1986). RR Lyrae stars in globular clusters: better distances from infrared measurements? *MNRAS*, 220:270.

LSST Science Collaborations and LSST Project (2009). LSST Science Book, Version 2.0. *arXiv:0912.0201*.

Maartens, R., Abdalla, F. B., Jarvis, M., et al. (2015). Overview of Cosmology with the SKA. In *Proceedings, Advancing Astrophysics with the Square Kilometre Array (AASKA14): Giardini Naxos, Italy, June 9-13, 2014*, page 16.

Madore, B. F. and Freedman, W. L. (1991). The Cepheid distance scale. Publications of the Astronomical Society of the Pacific. *Harvard College Observatory Circular*, 103:933.

Magnier, E. (2006). The Pan-STARRS PS1 image processing pipeline. In Ryan, S., editor, *Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conf*, page E5.

Magnier, E. (2007). Calibration of the Pan-STARRS 3π Survey. In Sterken, C., editor, *The Future of Photometric, Spectrophotometr ic and Polarimetric Standard- ization*, volume 364, page 153.

Magnier, E. A., Monet, D. G., and Chambers, K. C. (2008). Globular Cluster Distance Determinations. In Jin, W. J., Platais, I., and Perryman, M. A. C., editors, *IAU Symp. 248, A Giant Step: From Milli- to Micro-arcsecond Astrometry*, volume 237, page 553. Cambridge Univ. Press.

Magnier, E. A., Schlafly, E., Finkbeiner, D., et al. (2012). The Pan-STARRS 1 Photometric Reference Ladder, Release 12.01. *ApJS*, 205(2):20.

Magorrian, J., Tremaine, S., Richstone, D., et al. (1998). The Demography of Massive Dark Objects in Galaxy Centers. *AJ*, 115(6):2285.

Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., et al. (2003). A Two Micron All Sky Survey View of the Sagittarius Dwarf Galaxy. I. Morphology of the Sagittarius Core and Tidal Arms. *ApJ*, 599(2):1082.

Marconi, M., Cignoni, M., Criscienzo, M. D., et al. (2015). Predicted properties of RR Lyrae stars in the Sloan Digital Sky Survey photometric system. *MNRAS*, 371:1503.

Marshall, D. J., Robin, A. C., Reylé, C., et al. (2006). Modelling the Galactic interstellar extinction distribution in three dimensions. *A&A*, 453(2):635.

Mason, L., Baxter, J., Bartlett, P. L., et al. (2009). Boosting Algorithms as Gradient Descent. In Solla, S. A., Leen, T. K., and Müller, K., editors, *Advances in Neural Information Processing Systems*, number 12.

Mateo, M., Slzewski, E. W., Vogt, S. S., et al. (1998). The Internal Kinematics of the Leo I Dwarf Spheroidal Galaxy: Dark Matter at the Fringe of the Milky Way. *AJ*, 116:2315.

Mateu, C., Vivas, A. K., Downes, J. J., et al. (2012). QUEST RR Lyrae Survey III. Low Galactic latitude. *MNRAS*, 427(2):4.

Matsunga, N., Feast, M. W., and Menies, J. W. (2009). Period-luminosity relations for type II Cepheids. In Guzik, J. A. and Bradley, P., editors, *Stellar Pulsation: Challenges for Theory and Observation, AIP Conference Proceedings*, number 1170, page 96.

Metcalfe, N., Farrow, D. J., Cole, S., et al. (2013). The Pan-STARRS1 Small Area Survey 2. *MNRAS*, 435(3):1825.

Metropolis, N. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092.

Mighell, K. J. and Burke, C. J. (1999). WFPC2 Observations of the Ursa Minor Dwarf Spheroidal Galaxy. *ApJ*, 118(1):366.

Millman, J. K. and Aivazis, M. (2011). Python for Scientists and Engineers. *Computing in Science Engineering*, 13:9.

Monson, A. J. and Pierce, M. J. (2011). Near-infrared (JHK) photometry of 131 Northern Galactic classical Cepheids. *ApJS*, 193(1):12.

Morganson, E., Burgett, W. S., Chambers, K. C., et al. (2014). Measuring Quasar Variability with Pan-STARRS1 and SDSS. *ApJ*, 784(2):2.

Moya, A. and Rogríduez-López, C. (2010). Has a star enough energy to excite the thousand of modes observed with CoRoT? *ApJL*, 710:L7.

Nemec, J. M. (1985). Double-mode RR Lyrae stars in M15: reanalysis, and experiments with simulated photometry. *ApJ*, 90:240.

Newberg, H. J., Willett, B. A., Yanny, B., et al. (2010). The Orbit of the Orphan Stream. *ApJ*, 711(1):32.

Newberg, H. J., Yanny, B., Cole, N., et al. (2007). The overdensity in Virgo, Sagittarius debris, and the asymmetric spheroid. *ApJ*, 668(1):221.

Newberg, H. J., Yanny, B., Rockosi, C., et al. (2002). The Ghost of Sagittarius and Lumps in the Halo of the Milky Way. *ApJ*, 569(1):245.

Nikutta, R., Hunt-Walker, N., Nenkova, M., et al. (2014). The meaning of WISE colours - I. The Galaxy and its satellites. *MNRAS*, 442(4):3361.

Nun, I., Protopapas, P., Sim, B., et al. (2015). FATS: Feature Analysis for Time Series.

Oke, J. B. and Gunn, J. E. (1983). Secondary standard stars for absolute spectrophotometry. *ApJ*, 266(1):713.

Oosterhoff, P. T. (1939). Some remarks on the variable stars in globular clusters. *The Observatory*, 62:104.

Paczyński, B. (1996). Gravitational microlensing, the distance scale, and the ages. In Kochanek, C. S. and Hewitt, J. N., editors, *Astrophysical Applications of Gravitational Lensing, IAU Symposium*, volume 173, page 199.

Paust, N. E., Reid, I. N., Piotto, G., et al. (2010). The ACS Survey of Galactic Globular Clusters. VIII. Effects of Environment on Globular Cluster Global Mass Functions. *AJ*, 139(2):476.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Peñarrubia, J., Belokurov, V., Evans, N. W., et al. (2010). Was the progenitor of the sagittarius stream a disc galaxy? *MNRAS*, 408:L26.

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

Pickering, E. C., Colson, H. R., Fleming, W. P., et al. (1901). Sixty-four new variable stars. *ApJ*, 13:226.

Plummer, H. Z. C. (1913). Note on Variable Stars of Cluster Type. *MNRAS*, 73:652.

Pojmanski, G. (1997). The All Sky Automated Survey. *Acta Astronomica*, 47:467.

Pollock, J. T., Pica, A. J., Smith, A. G., et al. (1979). Long-term optical variations of 20 violently variable extragalactic radio sources. *AJ*, 84:1658.

Popper, D. M. (1980). Stellar masses. *Annual Review of Astronomy and Astrophysics*, 18:115.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2:3.

Preston, G. W., Shectman, S. A., and Beers, T. C. (1991). Detection of a galactic color gradient for blue horizontal-branch stars of the halo field and implications for the halo age and density distributions. *ApJ*, 375(1):121.

Pritchet, C. J. (2005). SNLS – The Supernova Legacy Survey. In Wolff, S. and Lauer, T. R., editors, *Observing Dark Energy, ASP Conference Series,*, volume 339. Astronomical Society of the Pacific.

Prusti, T. (2014). Gaia: scientific in-orbit performance. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 9143. Proceedings of the SPIE.

Rau, A., Kulkarni, S. R., Law, N. M., et al. (2009). Exploring the Optical Transient Sky with the Palomar Transient Factory. *PASP*, 121:886.

Rest, A., Scolnic, D., Foley, R. J., et al. (2014). Cosmological Constraints from Measurements of Type Ia Supernovae Discovered during the First 1.5 yr of the Pan-STARRS1 Survey. *ApJ*, 795(1):44.

Richards, J. W., Starr, D. L., Butler, N. R., et al. (2011). On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *ApJ*, 733(1):10.

Richey, M. (2010). The Evolution of Markov Chain Monte Carlo Methods. *The American Mathematical Monthly*, 117(5):383–413.

Rickett, B. J., Coles, W. A., and Bourgois, G. (1984). Slow scintillation in the interstellar medium. *A&A*, 134(2):390.

Ritter, A. (1879). Untersuchungen über die Höhe der Atmosphäre und die Constituion gasförmiger Weltkörper. *Annalen der Physik und Chemie*, 244(9):157.

Rosseland, S. (1949). *The Pulsation Theory of Variable Stars*. Clearedon Press, Oxford.

Ruhland, C., Bell, E. F., Rix, H.-W., et al. (2011). The Structure of the Sagittarius Stellar Stream as Traced by Blue Horizontal Branch Stars. *ApJ*, 731(2):119.

Rybicki, G. B. (1994). Notes on Gaussian Random Functions with Exponential Correlation Functions (unpublished notes).

Rybicki, G. B. and Press, W. H. (1992). Interpolation, Realization, and Reconstruction of Noisy, Irregularly Sampled Data. *ApJ*, 398(1):169.

Rybicki, G. B. and Press, W. H. (1995). Class of fast methods for processing irregularly sampled or otherwise inhomogeneous one-dimensional data. *Physical Review Letters*, 74:1060.

S. C. Keller, B. P. Schmidt, M. S. B. et al. (2007). The Two Micron All Sky Survey (2MASS). *Publications of the Astronomical Society of Australia*, 24(1):1.

Sandage, A. (1981). The Oosterhoff period groups and the age of globular clusters. II. Properties of RR Lyrae stars in six clusters: the P-L-A relation. *ApJ*, 248:161.

Sandage, A. (1990a). The Oosterhoff period effect: luminosities of globular cluster zero-age horizontal branches and field RR Lyrae stars as a function of metallicity. *ApJ*, 350:631.

Sandage, A. (1990b). The vertical height of the horizontal branch: the range in the absolute magnitudes of RR Lyrae sars n a given globular cluster. *ApJ*, 350:603.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *ApJ*, 263(1):835.

Schlafly, E. F. and Finkbeiner, D. F. (2011). Measuring Reddening with Sloan Digital Sky Survey Stellar Spectra and Recalibrating SFD. *ApJ*, 737(2):103.

Schlafly, E. F., Finkbeiner, D. F., Jurić, M., et al. (2012). Photometric Calibration of the First 1.5 Years of the Pan-STARRS1 Survey. *ApJ*, 756(2):158.

Schlafly, E. F., Green, G., Finkbeiner, D. F., et al. (2014). A Map of Dust Reddening to 4.5 kpc from Pan-STARRS1. *ApJ*, 789(1):15.

Schmidt, K., Marshall, P. J., Rix, H.-W., et al. (2010). Selecting Quasars by their Intrinsic Variability. *ApJ*, 714(2):1194.

Schmidt, K. B., Rix, H.-W., Shields, J. C., et al. (2012). The Color Variability of Quasars. *ApJ*, 744:147.

Schneider, D. P., Hall, P. B., Richards, G. T., et al. (2007). The Sloan Digital Sky Survey Quasar Catalog. IV. Fifth Data Release. *AJ*, 134(1):102.

Schwarzschild, K. (1900). Beiträge zur photographischen Photometric der Gestirne. *Publ. der V. Kuffner'schen Sternwarte in Wien*, V:1.

Schwarzschild, K. (1906). Ueber das Gleichgewicht der Sonnenatmosphäre. *Nachrichten von der Königlichen Gesellschft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1:4.

Sesar, B. (2012). Template RR Lyrae Hα, Hβ, and Hγ Velocity Curves. *AJ*, 144(4):114.

Sesar, B., Banholzer, S. R., Cohen, J. G., et al. (2014). Stacking the Invisibles: A Guided Search for Low-Luminosity Milky Way Satellites. *ApJ*, 793(2):135.

Sesar, B., Grillmair, C. J., Cohen, J., et al. (2013a). Tracing the Orphan Stream to 55 kpc with RR Lyrae Stars. *ApJ*, 776(1):26.

Sesar, B., Hernitschek, N., Mitrović, S., et al. (2016). Machine-Learned Identification of RR Lyrae Stars from Sparse, Multiband Data: The PS1 Sample. *ApJ submitted*.

Sesar, B., Ivezić, Ž., and Grammer, S. H. (2010). Light Curve Templates and Galactic Distribution of RR Lyrae Stars from Sloan Digital Sky Survey Stripe 82. *ApJ*, 708(1):717.

Sesar, B., Ivezić, Ž., Lupton, R., et al. (2007). Exploring the Variable Sky with the Sloan Digital Sky Survey. *ApJ*, 134(6):2236.

Sesar, B., Ivezić, Ž., Scott, J. S., et al. (2013b). Exploring the Variable Sky with LINEAR. II. Halo Structure and Substructure Traced by RR Lyrae Stars to 30 kpc. *AJ*, 146(2):21.

Shapley, H. (1914). On the Nature and Cause of Cepheid Variation. *A&A*, 40:448.

Shapley, H. (1918). No. 153. Studies based on the colors and magnitudes in stellar clusters. Eighth paper: The luminosities and distances of 139 Cepheid variables. *Contributions from the Mount Wilson Observatory*, 153:1.

Shapley, H. and Sawyer, H. B. (1927). A Classification of Globular Clusters. *Harvard College Observatory Bulletin*, 849:11.

Silk, J. and Rees, M. J. (1998). Quasars and galaxy formation. *A&A*, 331:L1.

Simon, J. D., Geha, M., Minor, Q. E., et al. (2011). A Complete Spectroscopic Survey of the Milky Way Satellite Segue 1: The Darkest Galaxy. *ApJ*, 733(1):46.

Simon, N. R. (1982). A plea for reexamining heavy element opacities in stars. *ApJ*, 260:L87.

Simoneti, J. H., Cordes, J. M., and Heeschen, D. S. (1985). Flicker of extragalactic radio sources at two frequencies. *A&A*, 296:46.

Skrutskie, M., Cutri, R., Stiening, R., et al. (2006). The Two Micron All Sky Survey (2MASS). *AJ*, 131:1163.

Smith, H. A. (2004). *RR Lyrae Stars*. Cambridge University Press.

Sollima, A., Cacciari, C., and Valenti, E. (2006). The RR Lyrae period-K-luminosity relation for globular clusters: an observational approach . *MNRAS*, 372:1675.

Soubiran, C. (1993). Kinematics of the Galaxy's Stellar Populations from a Proper-Motion Survey. *A&A*, 274:181.

Spitzer, L. and Saslaw, W. C. (1966). On the Evolution of Galactic Nuclei. *ApJ*, 143:400.

Stokes, H. G., Shelley, F., Viggh, H. E. M., et al. (1998). The Lincoln Near-Earth Asteroid Research (LINEAR) Program. *Lincoln Laboratory Journal*, 11(1):27.

Storn, R. and Price, K. (1997). Differential Evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341.

Stubbs, C. W., Doherty, P., Cramer, C., et al. (2010). Precise Throughput Determination of the Pan-STARRRS Telescope and the Gigapixel Imager Uing a Calibrated Silicon Photodiode and a Tunable Laser: Initial Results. *ApJ Suppl. Ser.*, 191:376.

Sweigart, A. V. and Renzini, A. (1979). Semi-convection and period changes in RR Lyrae stars. *A&A*, 71:66.

Szabó, R., Kollàth, Z., and Buchler, J. R. (2004). Automated nonlinear stellar pulsation calculations: Applications to RR Lyrae stars. The slope of the fundamental blue edge and the first RRd model survey. *A&A*, 425:627.

Tonry, C. W., Stubbs, K. R., Lykke, K. R., et al. (2012). The Pan-STARRS1 Photometric System. *ApJ*, 750(2):99.

Torrealba, G., Catelan, M., Drake, A. J., et al. (2009). Discovery of ∼9,000 new RR Lyrae in the Southern Catalina Surveys. *MNRAS*, 446(3):225.

Tremaine, S., Gebhardt, K., Bender, R., et al. (2002). The Slope of the Black Hole Mass versus Velocity Dispersion Correlation. *ApJ*, 574(2):740.

Udalski, A. (2003). The Optical Gravitational Lensing Experiment. Real Time Data Analysis Systems in the OGLE-III Survey. *Acta Astron.*, 53:305.

van Albada, T. S. and Baker, N. S. (1973). On the Two Oosterhoff Groups of Globular Clusters. *ApJ*, 185:477.

van der Marel, R. P., de Zeew, P. T., Rix, H.-W., et al. (1997). A massive black hole at the centre of the quiescent galaxy M32. *Nature*, 385(6617):610.

VanderPlas, J. T. and Ivezić, Ž. (2015). Periodograms for Multiband Astronomical Time Series. *ApJ*, 812(1):18.

Ďurech, J., Sidorin, V., and Kaasalainen, M. (1999). Study of FK Comae Berenices, I. Surface images for 1994 and 1995. *A&A*, 513:A46.

Vickers, J. J., Grebel, E. K., and Huxor, A. P. (2012). Identifying Blue Horizontal Branch Stars Using the z Filter. *AJ*, 143(4):86.

Vivas, A. K., Zinn, R., Andrews, P., et al. (2001). The QUEST RR Lyrae Survey: Confirmation of the Clump at 50 Kiloparsecs and Other Overdensities in the Outer Halo. *AJ*, 554:L33.

Watson, C. L., Henden, A. A., and Price, A. (2006). The International Variable Star Index (VSX). *Society for Astronomical Sciences and Annual Symposium*, 25:47.

Weinzierl, S. (2000). Introduction to Monte Carlo methods. *eprint arXiv:hep-ph/0006269*.

Wenger, M., Ochsenbein, F., Egret, D., et al. (2000). The SIMBAD astronomical database. The CDS reference database for astronomical objects. *A&AS*, 143:9.

Werner, M., Roelling, T., Low, F., et al. (2004). Spitzer Summary. *ApJS*, 154:1.

Winkler, P. F., Gupta, G., and Long, K. S. (2003). The SN 1006 Remnant: Optical Proper Motions, Deep Imaging, Distance, and Brightness at Maximum. *ApJ*, 585(1):324.

Woźniak, P. R., Vestrand, W. T., Akerlof, C. W., et al. (2004). Norther sky variability survey: public data release. *AJ*, 127:2436.

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. (2010). The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *AJ*, 140(6):1868.

Wuchterl, G. and Klessen, R. S. (2001). The first million years of the Sun: a calculation of the formation and early evolution of a solar mass star. *ApJL*, 560:L185.

Xue, X.-X., Rix, H.-W., Ma, Z., et al. (2015). The Radial Profile and Flattening of the Milky Way's Stellar Halo to 80 kpc from the SEGUE K-Giant Survey. *ApJ*, 809(2):144.

York, D. G., Adelman, J., Anderson, J., et al. (2000). The Sloan Digital Sky Survey: Technical Summary. *ApJ*, 120:1579.

Zackrisson, E. and Bergvall, N. (2003). Can microlensing explain the long-term optical variability of quasars? *A&A*, 408:17.

Zhevakin, S. A. (1963). Physical basis of the pulsation theory of variable stars. *Annual Review of Astronomy and Astrophysics*, 1:367.

Zu, Y., Kochanek, C. S., and Peterson, B. M. (2011). An alternative approach to measuring reverberation lags in active galactic nuclei. *ApJ*, 735(2):80.

List of Figures

List of Tables

Acknowledgements

Now, this huge piece of work is finished. Now, that's the right place to thank all the people who made this happen.

First of all, I would like to thank my supervisor Hans-Walter Rix for the continuous support and all the very useful comments and remarks.
Furthermore I would like to thank Eddie Schlafly and Branimir Sesar for many discussions on machine-learning, surveys and astronomy in general. Their guidance helped me in all the time and improved my scientific methods a lot. Also, especially, I want to thank Eddie Schlafly for supporting me with Pan-STARRS knowledge.

During my thesis, I was very glad to talk to scientists from many different research fields. I would like to thank Željko Ivezić and David Hogg for construtive comments on machine-learning, as well as Vasily Belokurov for answering many questions regarding Sagittarius stream.

Finally, beyond physics, I'd like to thank my spouse Florian Grimps for supporting me throughout all my studies. Thanks for patience, thanks for discussion on various things and for the non-astronomer's view on my work. Having someone to rely on is that precious!

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP 7) ERC Grant Agreement n. [321035].

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

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

The CSS survey is funded by the National Aeronautics and Space Administration under Grant No. NNG05GF22G issued through the Science Mission Directorate Near-Earth Objects Observations Program. The CRTS survey is supported by the U.S. National Science Foundation under grants AST-0909182 and AST-1313422. The services at IUCAA are supported by the University Grants Commission and the Ministry of Information Technology, Govt. of India under the Virtual Observatory - India project.

Computations were made using the supercomputer facilities of Jülich Supercomputing Centre (JSC) through the SFB881 and Max Planck Computing and Data Facility (MPCDF).