

IDENTIFYING RR LYRAE VARIABLE STARS
IN THE NOIRLAB SOURCE CATALOG
WITH TEMPLATE FITTING

by

Kyle Louis Matt

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Physics

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2022

©COPYRIGHT

by

Kyle Louis Matt

2022

All Rights Reserved

ACKNOWLEDGEMENTS

This project was made possible through funding from the Montana State University Physics Department and a grant from the Montana Space Grant Consortium. I want to thank my advisor, Dr. David Nidever, for his guidance and support in this project and my degree generally.

My wife, D'Arcy White, has been an important pillar of support these past four years. They have been there for me during difficult times, comforted me during many stressful late nights, and helped proofread this thesis.

I also would like to thank my parents, Edwin and Sharon Matt, for the support they have provided me and for encouraging me to try difficult things. I also acknowledge my late grandfather, Dr. Douglas Henderson. A physicist and educator, he was a great inspiration for me and encouraged my love for physics.

VITA

Kyle Louis Matt was born on the 16th of September 1991 in Sandy, Utah, to Edwin and Sharon Matt. He graduated from Park City High School in 2010 in the top ten percent of his class. He graduated from Brigham Young University with a Bachelor of Science Degree in Physics-Astronomy in 2017. He attended Montana State University and received a Master of Science degree in Physics in 2022.

Kyle is a registered descendent of the Blackfeet Nation, headquartered in Browning, Montana.

His paternal grandparents are Lenore and Ward 'Skippy' Matt. Lenore was a secretary at Blackfeet Community College, and Skippy is a rancher. His maternal grandparents were Douglas and Rose-Marie Henderson, a physicist and a homemaker who raised three intelligent women with love for family.

TABLE OF CONTENTS

1. INTRODUCTION	1
2. METHODS	5
Templates and Light Curves	5
Period Search	7
Template Fitting	12
Distance Determination	18
3. RESULTS	20
4. CONCLUSIONS AND FUTURE DIRECTIONS.....	28
REFERENCES CITED.....	30
APPENDIX A: Template Fitter Code.....	33

LIST OF FIGURES

Figure	Page
1.1 \log_{10} Period versus Magnitude from Leavitt and Pickering [1912].	2
2.1 Sample template light curve of RR Lyrae type AB.	6
2.2 Example light curve using different periods.	6
2.3 An example of periodograms plotted.	10
2.4 Final Ψ periodogram, combining data from all available bands.	11
2.5 Demonstration of template stretching and offsets.	12
2.6 Light curve with the standard u , g , r , i , z and Y bands displayed.	13
2.7 Normalized light curve with u , g , r , i , z and Y bands superimposed.	14
2.8 Comparisons of amplitude ratios.	16
2.9 Final light curves for each band with best-fitting template.	17
2.10 Period-Luminosity plots from Cáceres and Catelan [2008].	19
3.1 Period-Amplitude diagram comparison.	21
3.2 Object with a detected period near 1/2 day.	22
3.3 Best-fit NSC periods versus Gaia periods.	22
3.4 Lomb-Scargle Periodogram with a beat period shown.	23
3.5 Light Curve example of an object with a Beat Period.	24
3.6 Periodogram of sparsely sampled object where beat period was detected.	25
3.7 Period-Amplitude plot showing lobe selection.	26
3.8 Spatial distribution of RRL sample. The Magellanic Clouds are highlighted in orange.	27

ABSTRACT

RR Lyrae are periodic variable stars generally with periods between 5 hours and 1 day. They can be used as standard candles for accurate distance measurements and thus are useful for studying the structure of the Milky Way and its stellar clusters. The second data release of the *NoirLab* Source Catalog is a large collection of 68 billion time-series measurements of 3.9 billion objects. To process this large volume of data, we designed a computer software package in Python called *Leavitt* to automate the detection process and measure their properties including period, magnitude, epoch of maximum brightness and amplitude of their pulsations by fitting their light curves to templates. In addition to identifying RR Lyrae, it is expected that *Leavitt* can be extended to identify similar variable stars such as Cepheids in the same dataset. Distances were calculated for the initial catalog of RR Lyrae candidates using parameters measured with this script.

INTRODUCTION

In the field of astronomy, distance is one of the hardest things to measure, yet distances are critical to understanding three dimensional structures like our galaxy. For objects relatively close to the Earth, their distance can be determined by parallax, a geometric principle where a nearby object appears to shift relative to the background when viewed from different positions. For great distances, the parallax angle is very small and thus this method struggles to measure distances more than a few thousand parsecs.

Fortunately, distance can also be estimated if one knows the true brightness of a celestial object by comparing it to the observed brightness. Since light radiates outward in a sphere, the intensity of that light will drop with the square of the distance. Using a logarithmic scale such as stellar magnitudes, this distance relationship can be expressed as,

$$m - M = 5\text{Log}_{10}(d/10) \tag{1.1}$$

where m is the apparent magnitude, M the absolute magnitude and d the distance in parsecs. These objects, for which the true brightness can be determined, are known as standard candles.

In 1912, Henrietta Leavitt discovered that the period of variation for Cepheid variable stars corresponds to their absolute luminosity, making them the the first class of standard candle [Leavitt and Pickering, 1912]. Cepheids are bright, thus they are great for measuring distances to other galaxies, but they are rare. A plot from Leavitt's 1912 paper showing Period-Luminosity of Cepheids is included in Figure 1.1.

Pulsating variable stars, like Cepheids, are unstable stars that vary in brightness due to radial swelling and shrinking. There exist several classes of pulsating variable star including

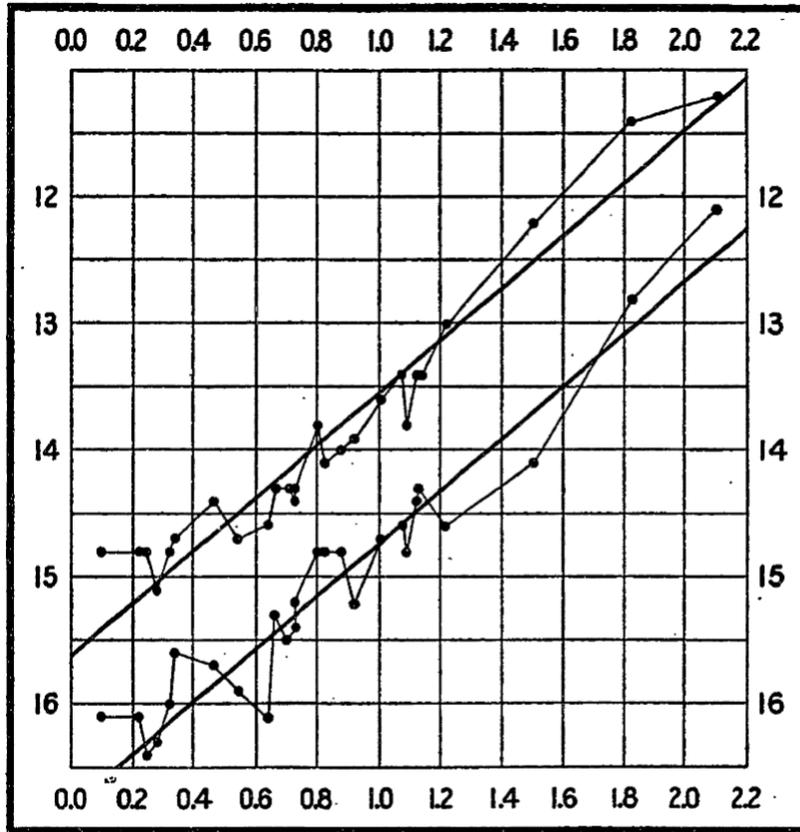


Figure 1.1: Log_{10} Period versus Magnitude from Leavitt and Pickering [1912].

the Classical Cepheids, δ Scuti and RR Lyrae. Each class differs in some characteristics such as mass, metallicity and age, but each has a period-luminosity relationship.

The existence of a relationship between luminosity and period suggests there is an underlying mechanism linking these properties together. Sir Arthur Eddington [1941] was an early astrophysicist to work on the problem of how Cepheid variables work. He realized that the period roughly corresponded to the time it would take for a sound wave to propagate from one side of a star to the other, so a pressure wave is likely involved.

One major problem though is an acoustic wave would dissipate away over a relatively short time frame, thus some mechanism must be driving the pulsations for them to continue. Eddington would go on to suggest the stars behave like a thermodynamic engine, converting heat into mechanical energy and proposed a method by which this engine might work. At the right temperature, helium ionizes and becomes opaque during the transition. Energy then becomes trapped within the star at this layer, allowing energy to build up until the pressure increases enough to push the opaque layer away and the energy gets released.

We would like to be able to identify all variable types in the future, but for now this project focuses on RR Lyrae (RRL). RRLs typically pulsate with periods between 0.2 and 1 days, they are old, metal-poor stars with lower mass than the Sun, in the horizontal branch phase. RRL are not as bright as Cepheids, but they are much more common. Therefore, they can be good for distances within the Milky Way, its halo, and satellite galaxies. In addition, because they are more common, we can create a fuller picture by mapping out whole structures in 3D space.

We are searching for variable stars in the second data release of the NOIRLab Source Catalog (NSC) which is an all sky survey containing large collection of 68 billion time-series measurements of 3.9 billion objects, in 7 bands; u , g , r , i , z , Y , VR . [Nidever et al., 2021] The NSC is especially great because, in addition to its large size, it contains fainter objects than most large surveys meaning it can see more distant objects. It also covers much of the

southern hemisphere, which has not been as well studied as the northern hemisphere.

Due to the very large amount of data in NSC DR2, automation is necessary to identify and classify variable stars. The primary method for identifying these variable stars we have settled on is the method of light curve template fitting. We can test a variety of templates, and select objects for which the RRL templates fit best. This is done as a computer script in the python language, we call our program `Leavitt` after Henrietta Leavitt for her instrumental contributions to the study of variable stars.

`Leavitt` can be found at <https://github.com/KyleLMatt/leavitt>

METHODS

Templates and Light Curves

A light curve is a plot of magnitudes over time, and can be used to show the variability of an object. We use empirical templates, made by Layden [1998] based on the shapes of light curves from known RRL stars. An example template is shown in Figure 2.1. These templates are normalized in magnitude and phase to the range of 0 to 1. This allows us to fit the template to data by multiplying the y axis by the peak-to-peak amplitude, converting it to magnitude by adding the mean magnitude, and multiplying the x axis (phase) by the period, converting it to time. We can then compare these templates to data to see how well they fit.

Figure 2.2a shows an example of a raw light curve. This plot covers over 2000 days of data. At this scale, it is difficult to see the variable pattern. We will want to “fold” this data into phase space, which can be done if we know the period by using Equation 2.1. We take the time (here measured in days), subtract by some reference epoch t_0 , then divide by the period (P) to get a phase value. The integer portion of this phase is the number of cycles since the reference epoch. We can subtract this cycle number off to leave a decimal value between 0 and 1, leaving the phase. Plotting the same data from Figure 2.2a, but with the days converted to phase using a period of 0.61833 days, we obtain Figure 2.2b, which appears to match the template shown in Figure 2.1.

$$\phi = \frac{t - t_0}{P} - \text{Floor} \left(\frac{t - t_0}{P} \right) \quad (2.1)$$

Note that the calculated phase value is extremely sensitive to the period. Rounding the period in the previous example to 0.61830 days gives us a messier plot in Figure 2.2c. Any error in the period compounds with each cycle, so after thousands of cycles this small error

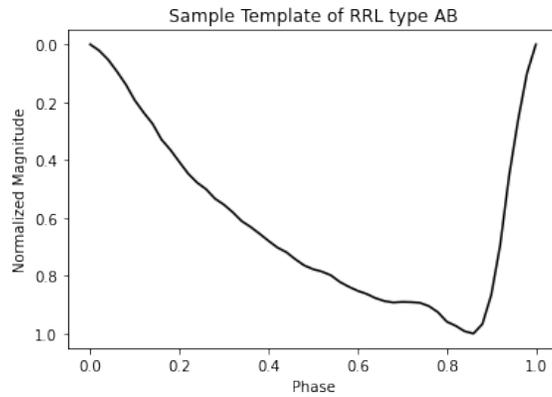
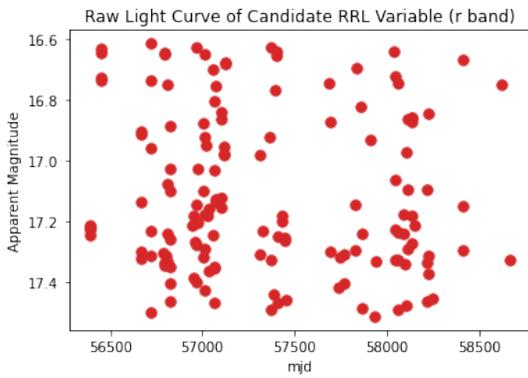
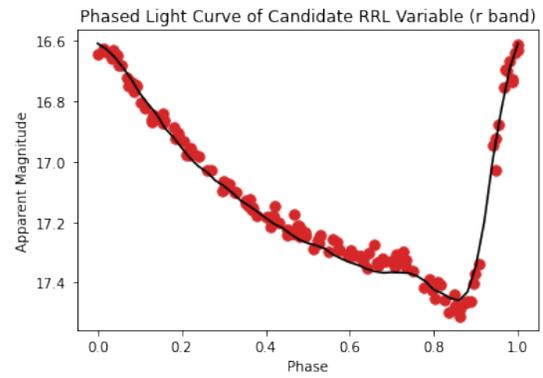


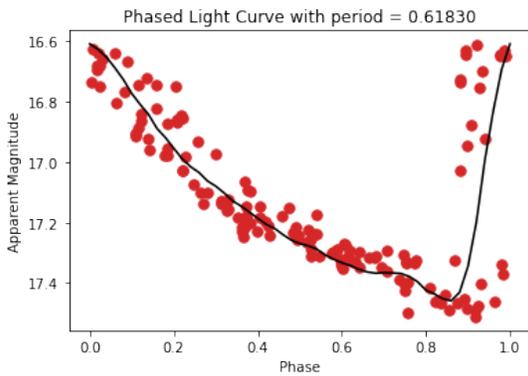
Figure 2.1: Sample template light curve of RR Lyrae type AB.



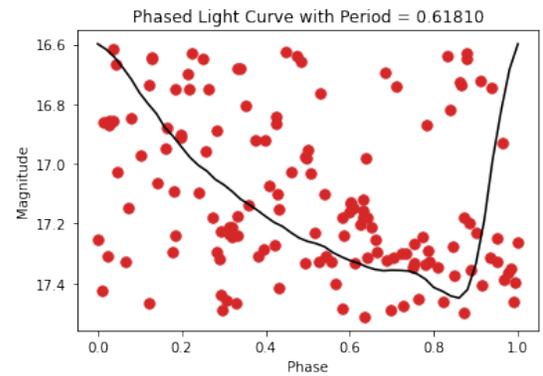
(a) Raw light curve, r band.



(b) Phased light curve.



(c) Phased light curve with small period error.



(d) Phased light curve with a slightly larger period error.

Figure 2.2: Example light curve using different periods.

adds up. When the error is small enough, it may be possible to find the correct period by trying small adjustments. But a slightly larger error may make the period unrecoverable like shown in Figure 2.2d.

Period Search

Rather than exhaustively fitting a template to tens of thousands of possible values of the period, we can narrow down the period search using a more computationally efficient algorithm. The traditional discrete Fourier transform, discussed below, is not very useful here because it requires evenly spaced data and most astronomical observations are taken at irregular intervals due to real world conditions like weather and daily or seasonal cycles. We instead use a hybrid algorithm from Saha and Vivas [2017] which combines the Lomb–Scargle periodogram and Laffer–Kinman’s Phase-Dispersion Minimization statistic.

The Lomb-Scargle periodogram [Lomb, 1976, Scargle, 1982] is closely related to the Fourier transform, a classic mathematical operation to decompose a function into its component frequencies. An integrable function, $f(t)$, can be transformed to a corresponding complex function of frequency, $\hat{f}(\omega)$, using the Fourier transform integral, Equation 2.2. The imaginary unit is denoted as $i \equiv \sqrt{-1}$

$$\hat{f}(\omega) \equiv \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \quad (2.2)$$

For a finite, evenly spaced sample of data, this can be approximated with the discrete Fourier transform, Equation 2.3. Here m_j represents a measurement and N is the total number of measurements.

$$\hat{f}_\omega = \sum_{j=1}^N m_j e^{-i\omega t_j} \quad (2.3)$$

If we compute the squared amplitude of this transformed function, as in Equation 2.4, we can

eliminate the complex component and remove the phase term from our choice of temporal epoch. This quantity is called the classical periodogram. By convention, we include a normalization factor of $1/N$.

$$\begin{aligned}
\mathcal{P}_c(\omega) &= \frac{1}{N} \left| \hat{f}_\omega \right|^2 \\
&= \frac{1}{N} \left| \sum_{j=1}^N m_j e^{-i\omega t_j} \right|^2 \\
&= \frac{1}{N} \left[\left(\sum_j m_j \cos \omega t_j \right)^2 + \left(\sum_j m_j \sin \omega t_j \right)^2 \right]
\end{aligned} \tag{2.4}$$

While the classical periodogram can be useful for estimating the dominant frequencies of simple periodic signals like sinusoids, it is often very noisy even when the data is not. This problem is only compounded when calculated on data unevenly sampled in time. Scargle [1982] addresses this by modifying the periodogram. First, he applied a time delay, τ , defined such that \mathcal{P} is invariant to time-translations,

$$\tan(2\omega\tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j} \tag{2.5}$$

Then, considering a generalized form of the periodogram, (Equation 2.6), he solved for the arbitrary functions, A and B, such that the periodogram reduced to the classical form when observations are evenly spaced and the power spectrum remains as close as possible to the evenly spaced case.

$$\mathcal{P}_{gen}(\omega) = \frac{A^2}{2} \left[\sum_j m_j \cos \omega(t_j - \tau) \right]^2 + \frac{B^2}{2} \left[\sum_j m_j \sin \omega(t_j - \tau) \right]^2 \tag{2.6}$$

With $A = (\sum_j \cos^2 \omega t_j)^{-1/2}$ and $B = (\sum_j \sin^2 \omega t_j)^{-1/2}$ these criteria are met. Thus the final

modified periodogram is defined as,

$$\mathcal{P}_{ls}(\omega) = \frac{1}{2} \left[\frac{\left(\sum_j m_j \cos \omega(t_j - \tau) \right)^2}{\sum_j \cos^2 \omega(t_j - \tau)} + \frac{\left(\sum_j m_j \sin \omega(t_j - \tau) \right)^2}{\sum_j \sin^2 \omega(t_j - \tau)} \right] \quad (2.7)$$

This periodogram is, remarkably, equivalent to fitting the data to a simple sinusoid and recording the χ^2 goodness-of-fit at each frequency (Equation 2.12), a method developed by Lomb [1976]. Because of the contributions from these two men, equation 2.7 is called the “Lomb-Scargle periodogram”.

The Lomb-Scargle periodogram, being a form of Fourier analysis, does not just contain information on the fundamental period, but on the entire waveform shape. This means there will be additional signals in the periodogram, apart from the fundamental frequency. With a very well-sampled light curve, this may be desirable, but when data are sparse the power can leak into these sub-dominant peaks or into aliases, possibly boosting an incorrect period. In particular, for ground based observations, a spike at a period of 1 day may appear since data can only be collected during nighttime hours. Additional spikes can appear at periods of $1/n$ days for integer n for the same reason. Other prominent peaks can occur due to beat interference between the actual signal’s period and this nightly observing cadence. The beat frequencies are discussed further in the Results section.

To strengthen our period search we introduce a second technique, not closely related to Fourier analysis, “Lafleur–Kinman’s Phase-Dispersion Minimization”. When folding time-series data by the correct period, as done in Figure 2.2b, we expect the magnitudes to closely follow a pattern, such as the template light curve, with minimal scatter. However, if the wrong period is used, the data will be randomly distributed in magnitude, like in Figure 2.2d. This concept is the basis behind Phase-Dispersion Minimization. Therefore, if we divide our phased data into bins, measure the dispersion of magnitude within each bin and then sum this dispersion across all bins, we would expect the trial period for which this sum

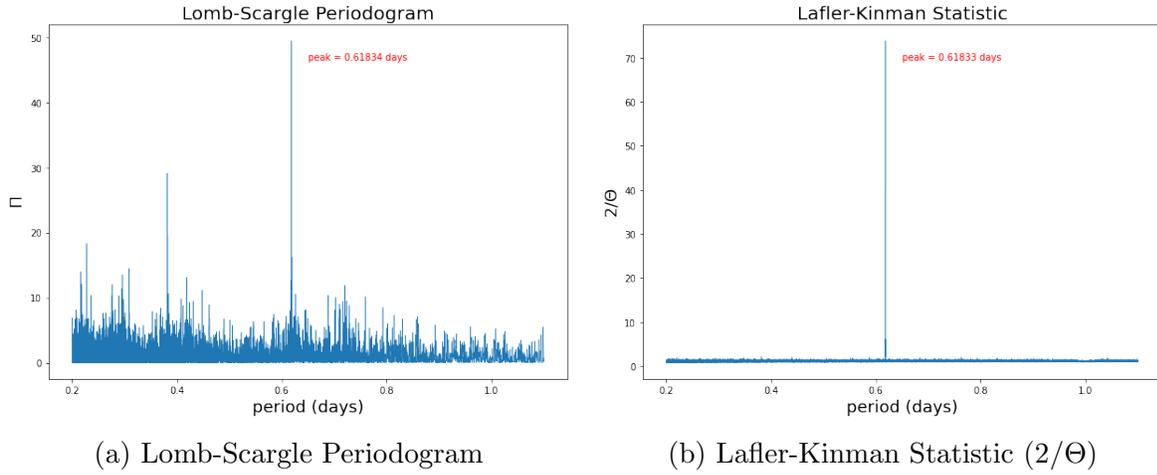


Figure 2.3: An example of periodograms plotted.

is a minimum to be nearest to the correct period. The implementation of this technique we use is the Lafler-Kinman statistic [Lafler and Kinman, 1965], using the test parameter defined by

$$\Theta = \frac{\sum_j^N (m_j - m_{j+1})^2}{\sum_j^N (m_j - \bar{m})^2} \quad (2.8)$$

where \bar{m} is the mean magnitude, m_j represents a magnitude measurement, and N is the number of observations. The numerator is the sum of the differences squared between adjacent measurements, if ordered by phase.

Strengths of the Lafler-Kinman approach include that it can handle irregularly spaced data and it does not require any assumptions about the shape of the light curve. A weakness is that for very sparse data, the size of gaps in successive phases can vary wildly, introducing false structures that can produce spurious minima.

These two methods have differing strengths and weaknesses. By combining them Saha and Vivas [2017] showed that they could each suppress spurious signals in the other while boosting the fundamental frequency.

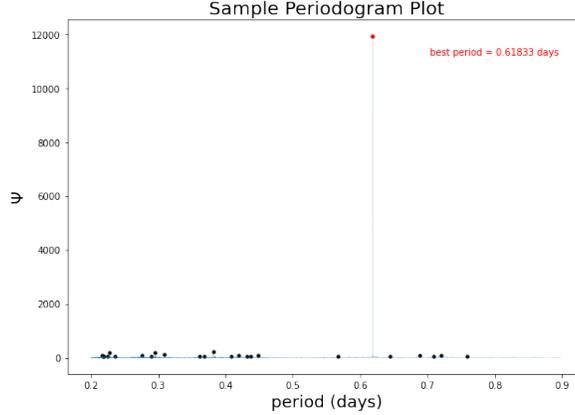
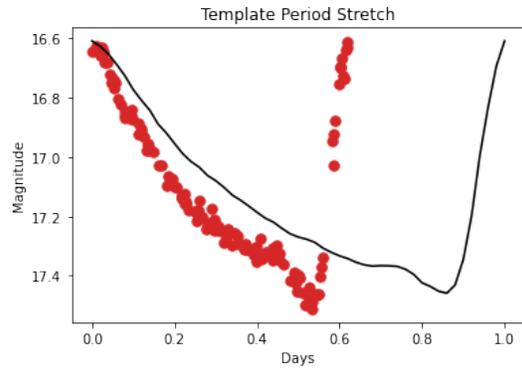


Figure 2.4: Final Ψ periodogram, combining data from all available bands.

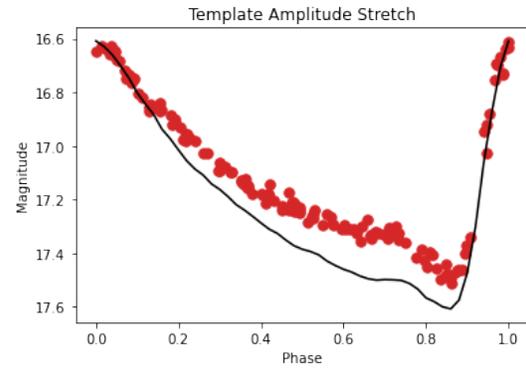
To combine them, we divide the Lomb-Scargle periodogram \mathcal{P} by the Lafler-Kinman statistic Θ , giving us a single value, Ψ . This is shown in Equation 2.9, a normalizing factor of 2 is added because Θ for un-periodic noise tends toward 2, thus the true statistic we consider is $2/\Theta$. Figure 2.3 shows both \mathcal{P} and $2/\Theta$, calculated on the r -band data from the same star as the above phase folding example. We see that for this dataset, \mathcal{P} has some noise but a clear peak is still present while $2/\Theta$ has a strong and unambiguous peak. Combining them in Figure 2.4, the noise in \mathcal{P} is suppressed by Θ .

$$\Psi = 2\mathcal{P}/\Theta \quad (2.9)$$

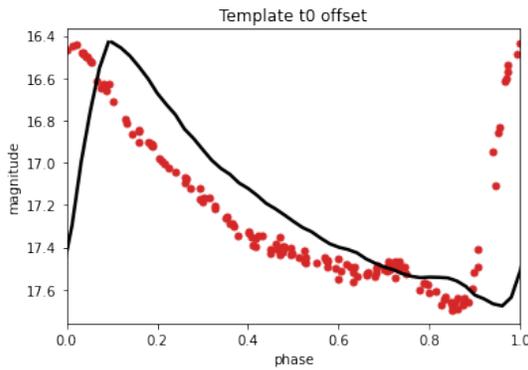
These procedures should be performed on data from a single band, but they can be repeated for each available band and the resulting Ψ s can be summed together to get a multi-band Ψ . Because it is possible for a spike other than the fundamental frequency to become dominant when working with noisy data, we select several peaks and test each by fitting our templates to them to decide which is the best period. Figure 2.4 includes black dots indicating the greatest peak as well as some other detected peaks. In this particular example, the power in the fundamental period is so strong that other peaks are negligible by comparison.



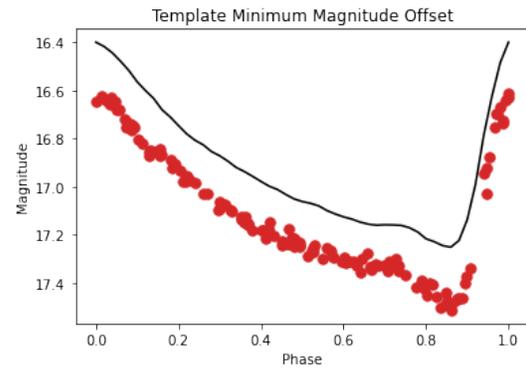
(a) Multiplying template x-axis by period converts it to time.



(b) Multiplying template y-axis by amplitude converts it to magnitude.



(c) Subtracting by epoch, t_0 , shifts time of maximum brightness



(d) Adding the minimum magnitude to the template magnitude shifts it from 0 to the desired value.

Figure 2.5: Demonstration of template stretching and offsets.

Template Fitting

Now that we have a short list of potential periods, we are ready to do the main template fitting. We start with templates that are normalized to the range 0–1 in both magnitude and time, which are then be stretched and shifted to match the data. This is convenient because these stretches and shifts will correspond to parameters of the star; namely the period, peak-to-peak amplitude, minimum magnitude (maximum brightness) and the epoch of maximum brightness (t_0). Figure 2.5 shows the stretches and shifts graphically.

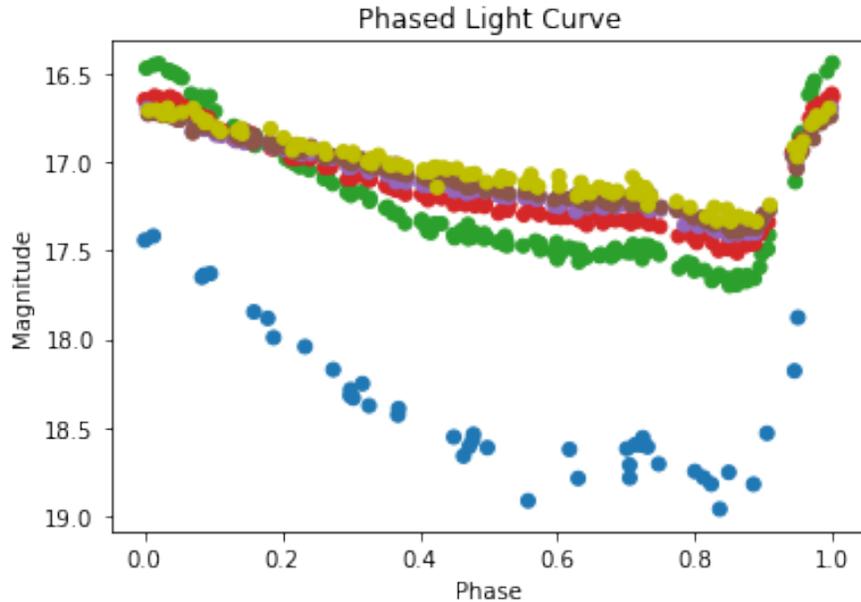


Figure 2.6: Light curve with the standard u , g , r , i , z and Y bands displayed.

Up to this point, the examples have primarily been primarily r -band data. When considering multiple bands, we must realize that some of these parameters may be different between bands. Let us look at all the data we have on this same star to compare how the bands differ, shown in Figure 2.6. These data have been folded by the same period and offset by the same t_0 epoch, thus we see that these two parameters do seem to be the same for all bands and can be fit for all bands simultaneously. The amplitude and magnitude in each band does appear to differ however, thus these parameters may have to be fit separately for each band.

As discussed earlier, the template can be stretched and shifted to fit the data. Since addition and multiplication are reversible (commutative) operations, we can also stretch and shift the data. Figure 2.7 shows these data after having their minimum magnitude shifted to zero and then dividing the magnitude measurements by the amplitude to normalize the data to roughly the 0 – 1 range.

The parameters being fitted are the Period, the epoch of maximum brightness (t_0),

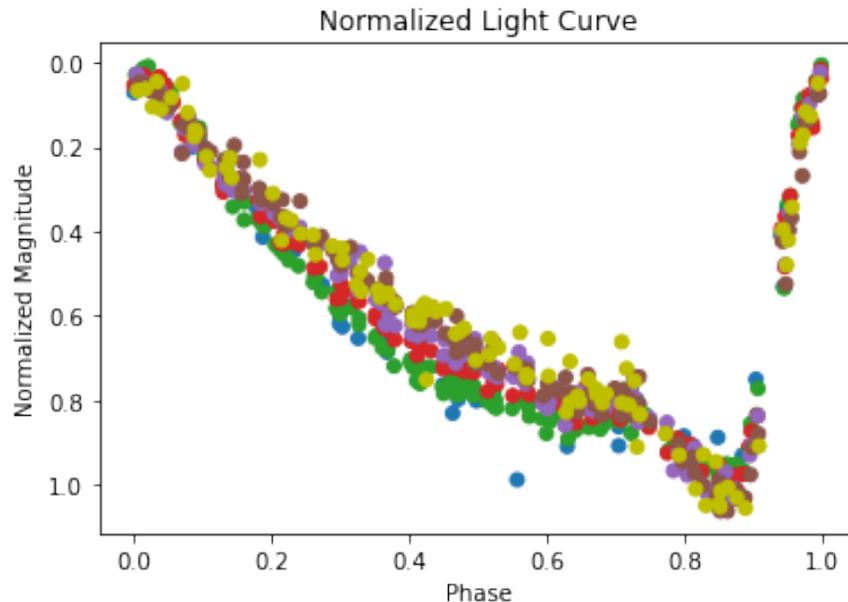


Figure 2.7: Normalized light curve with u , g , r , i , z and Y bands superimposed.

one amplitude per band, and one minimum magnitude per band. The total number of parameters is, therefore, dependent on how many bands there are, being $2 + 2 * k$, where k is the number of bands. This can be reduced further, as the ratio between the amplitude in one band and another has proven to be constant for RR Lyrae in established catalogs that I have investigated. For example, in Figure 2.8, are plotted the amplitudes for g and i bands versus the r amplitude of known RR Lyrae stars, taken from the Sloan Digital Sky Survey and Pan-STARRS1 survey Sesar et al. [2009, 2017]. We can see the amplitudes follow a linear pattern and the ratio is roughly equivalent between catalogs. Therefore, if we know these ratios we can fit just a single amplitude parameter and scale it for each band using the constant amplitude ratios.

Using the two catalogs shown in Figure 2.8 as a starting point, along with amplitudes fitted from NSC data, I measured the ratio of amplitudes for each possible pair of bands contained in the NSC and their inverses. A ratio times its inverse should equal one, to ensure this I set each ratio equal to the square root of the measured ratio divided by the measured

inverse ratio, for example, $(g/r) = \sqrt{\frac{(g/r)'}{(r/g)'}}$. This is essentially finding the geometric mean of the two measurements. I also realized additional information on the ratio between two bands could be found with the ratios of those bands with a third band using the following property, $(g/r) = \frac{(g/z)}{(r/z)}$. All these ratio measurements can then be combined with a geometric mean,

$$\mathcal{R}_{ij} = \left(\frac{R'_{ij}}{R'_{ji}} \prod_k \frac{R'_{ik}}{R'_{jk}} \right)^{1/n} \quad (2.10)$$

where R'_{ij} represents a measured amplitude ratio between bands i and j , the k in the Π product represents each other band, and n is the total number of bands.

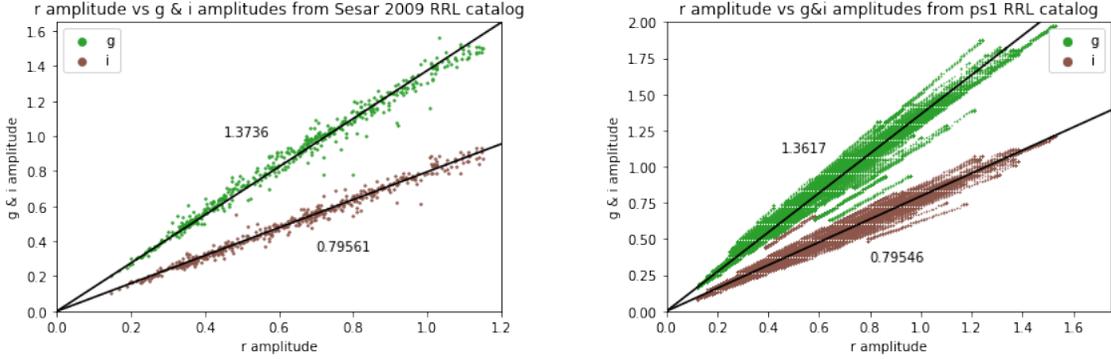
The final amplitude ratios for the NSC bands I have come up with, here expressed relative to the r band, are as follows,

u/r	1.81480
g/r	1.38605
r/r	1.00000
i/r	0.79662
z/r	0.74672
Y/r	0.71875
VR/r	1.05078

My template fitting script, `Leavitt`, includes a parameter to toggle the amplitude ratio mode on or off, as well as the ability to change these ratios if needed.

Since RRLs should have color indices within a known range, measured as the difference in magnitude between bands, it is conceivable that the magnitude parameter could also be constrained in a similar manner, but I have not attempted to implement this at the time of writing.

I use the `scipy.optimize.curve_fit` package in Python to solve for these parameters.



(a) From Sloan Digital Sky Survey, Sesar et al. [2009]

(b) From Pan-STARRS1 RR Lyrae catalog, Sesar et al. [2017]

Figure 2.8: Comparisons of amplitude ratios.

This is a non-linear least squares method of fitting a series of data to a non-linear model. It attempts to find a local minimum of the sum of squared residuals (Equation 2.11) by small, iterative adjustments to the parameters.

$$S = \sum_j^N (m_j - func(t_j, (\beta_0, \beta_1, \dots, \beta_k)))^2 \quad (2.11)$$

m_j is a magnitude measurement taken at time t_j , $func$ is a function that takes time and a series of parameters $(\beta_0, \beta_1, \dots, \beta_k)$ and returns the template value at the phase calculated from that time, after adjusting the template with these parameters.

To measure the quality of the fit, once parameters that minimize equation 2.11 are found, we use the χ^2 goodness-of-fit test (Eq. 2.12). This is a sum of the square differences between observations and the template divided by the square uncertainties of those observations, σ_j^2 .

$$\chi^2 = \sum_{j=1}^N \left[\frac{(m_{j,obs} - m_{j,tmp})^2}{\sigma_j^2} \right] \quad (2.12)$$

This fitting process is repeated for each trial period, and for each template. At the end,

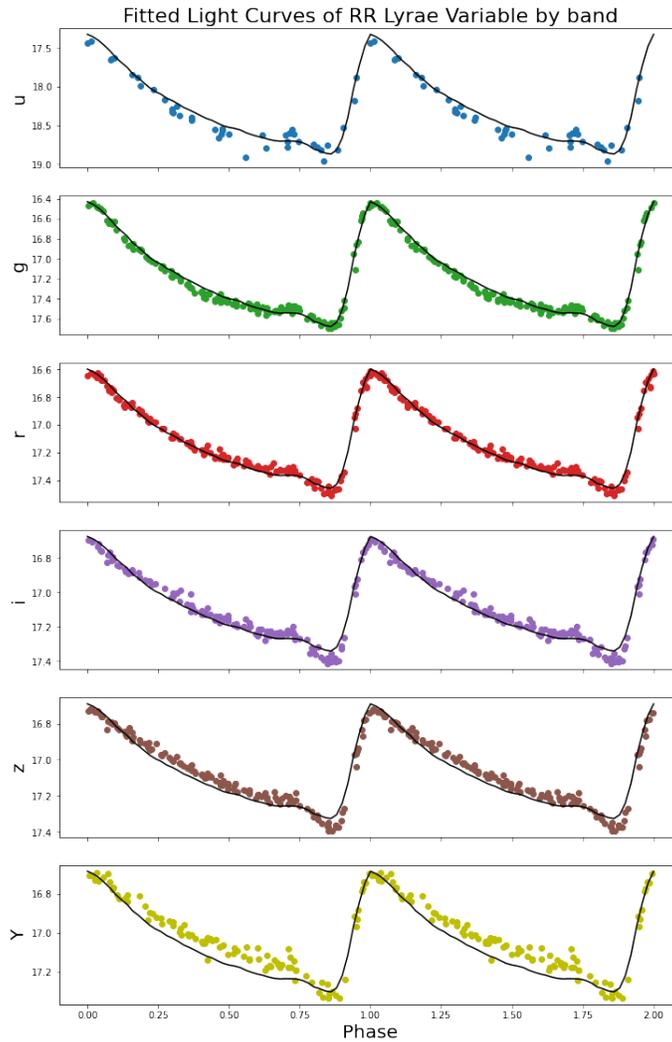


Figure 2.9: Final light curves for each band with best-fitting template.

the parameters and template with the smallest χ^2 are chosen as the best fit. With a set of best-fitting parameters, we can create a graphic with each band plotted separately against the template for easier viewing, see Figure 2.9.

The light curves in one band do not necessarily have the same shape as another for the same star. Sesar et al. [2009] explored the differences in light curve shapes between bands in greater detail, and created a large set of templates for each band. These templates are more complete than the Layden [1998] templates I use, but it is much more computationally

expensive to check them all, due to changing the templates one band at a time. I have created two versions of `Leavitt`, one that attempts to fit data from all bands to a single template, and one that tests different templates for each band. Since the primary goal for this project is to identify RR Lyrae candidates from a very large pool of candidates, I have opted to focus on the single template model as it can often run in less than 5% the time it takes to run the other. I am thus trading some accuracy for computational efficiency. Once candidates are narrowed down we have the option of running the more exhaustive fitter. It has been my experience that the single template model gives reasonably accurate results.

Distance Determination

As discussed in the introduction, RR Lyrae variables follow a period-luminosity-metallicity relationship that enables us to estimate their absolute magnitude. The general form of this relationship is,

$$M_\lambda = \alpha + \beta \text{Log}_{10}P + \gamma \text{Log}_{10}Z \quad (2.13)$$

where M_λ is the absolute magnitude in the λ band, α , β and γ are constants unique to each band, P is the period, Z is the metal abundance ratio. This relationship is strongest in near-infrared, but becomes weak in visible bands, for this reason the i and z bands are preferable to the u , g , and r bands. Cáceres and Catelan [2008] calculated a best fit for Equation 2.13 for i and z and obtained approximately,

	α	β	γ
i	0.908	-1.035	0.220
z	0.839	-1.295	0.211

The period, associated with these values is the fundamental period (RRab), for RRLs that

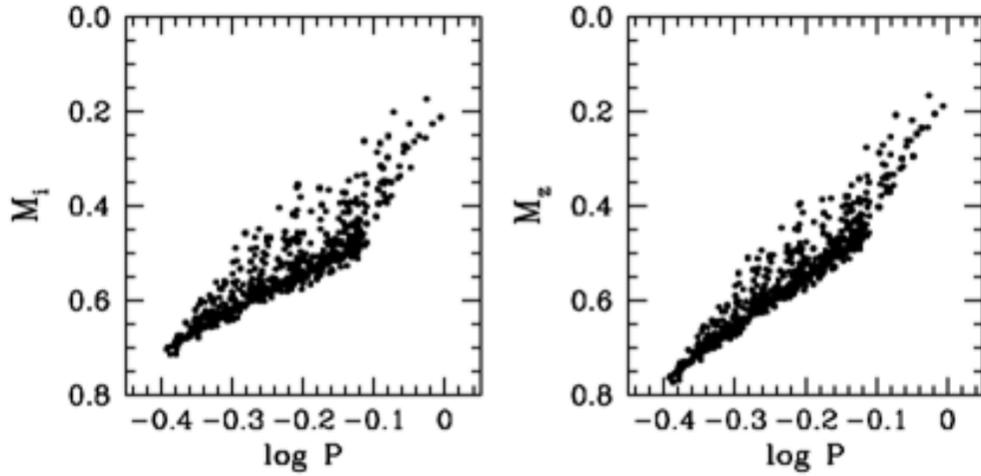


Figure 2.10: Period-Luminosity plots from Cáceres and Catelan [2008].

pulsate at an overtone (RRc). We must fundamentalize the period with the relationship, $\text{Log}_{10}P_f = \text{Log}_{10}P_o + 0.128$.

With an absolute magnitude, distance can be measured using the distance modulus,

$$m - M = 5\text{Log}_{10}(d/10) - A_\lambda \quad (2.14)$$

where m is the apparent magnitude, M the absolute magnitude, d the distance in parsecs, and A_λ is the extinction term.

RESULTS

We selected a subset of the NSC objects to run our fitting script, `Leavitt`, on. This subset included objects for which 30 or more measurements were available and were flagged as potentially variable by having a median absolute deviation (MAD) greater than 10 standard deviations above the median for all objects in the catalog. These criteria left us with a sample size of 2,031,102 variable star candidates.

$$MAD = \text{median}(|m_i - \text{median}(m_i)|) \quad (3.1)$$

These two million candidates were fit with templates. The results are summarized as a period-amplitude plot in Figure 3.1a. I have rejected stars with large χ^2 values and applied a color cut based on those suggested in Željko Ivezić et al. [2005], shown below, so we can focus on RRL candidates. In this plot, there are two prominent lobes centered around periods of 0.55 and 0.33 days, these are the RRL type ab and type c respectively. For comparison Figure 3.1b shows a period-amplitude plot using the RRL catalog from the second Gaia data release, Gaia Collaboration [2018].

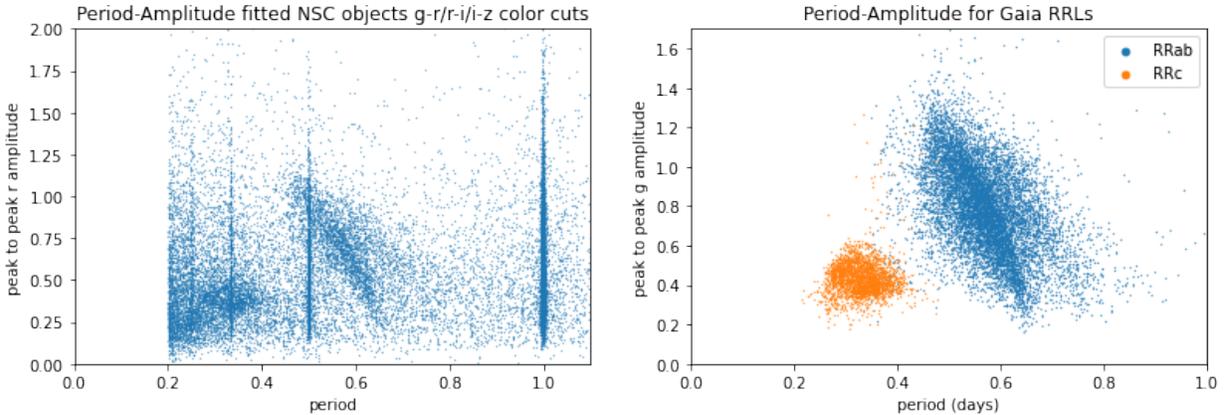
$$0.7 < u - g < 1.36$$

$$-0.15 < g - r < 0.33$$

$$-0.15 < r - i < 0.24$$

$$-0.19 < i - z < 0.23$$

Another prominent feature of Figure 3.1a are a series of vertical spikes located at periods of 1, 1/2, 1/3 and 1/4 days. These are due to the nightly observation cadence caused by the Earth's rotation period and aliases of it, and not due to intrinsic periodic behavior of the star.



(a) Fitted Period-Amplitude of the NSC stars. (b) Gaia RRL catalog Period-Amplitude.

Figure 3.1: Period-Amplitude diagram comparison.

I investigated some of the light curve for objects in these spikes to see if I could recognize any common patterns. Frequently, data from these stars is concentrated in one region of frequency (phase) space, like in Figure 3.2, while the rest of the light curve is poorly covered. By recognizing when data are temporally concentrated like this, we may be able to reject these false period detections as us merely rediscovering the Earth.

I cross-matched the right ascension and declination of the known RRLs in the Gaia catalog shown in Figure 3.1b with the NSC objects in my sample to find which stars are in both, finding 12,881 matches. This way I could compare my results directly with the known results, star by star. Figure 3.3 shows the best-fit NsC periods versus the periods from the Gaia catalog. The central diagonal are the objects for which my results match, but several other prominent lines are seen. These are beat frequencies that were picked up by the Lomb-Scargle periodogram during the period search phase.

A beat frequency is caused when two periodic signals with slightly different frequencies interfere. The classic example of this is when two slightly different sine tones are heard simultaneously, a beat materializes from the waves alternating between constructive and destructive interference. For our failed period detection, the pulsation frequency of the RRL

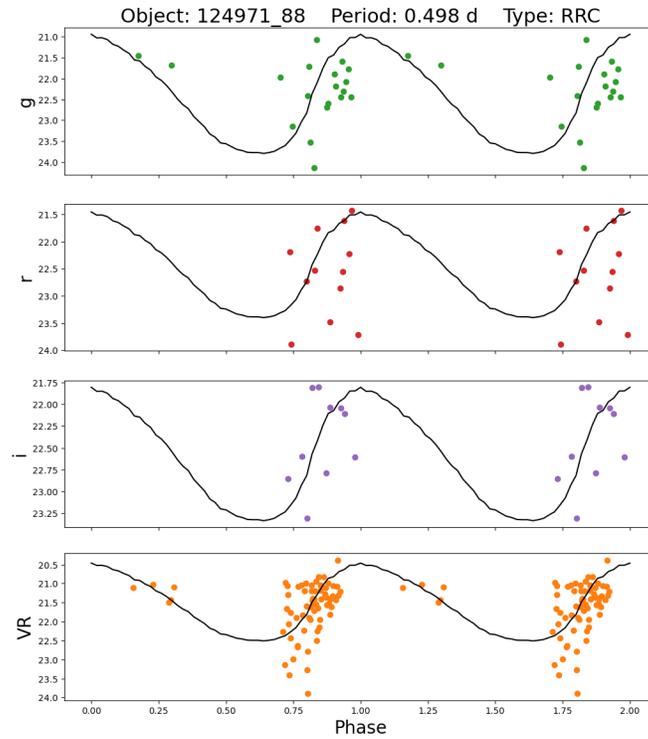


Figure 3.2: Object with a detected period near 1/2 day.

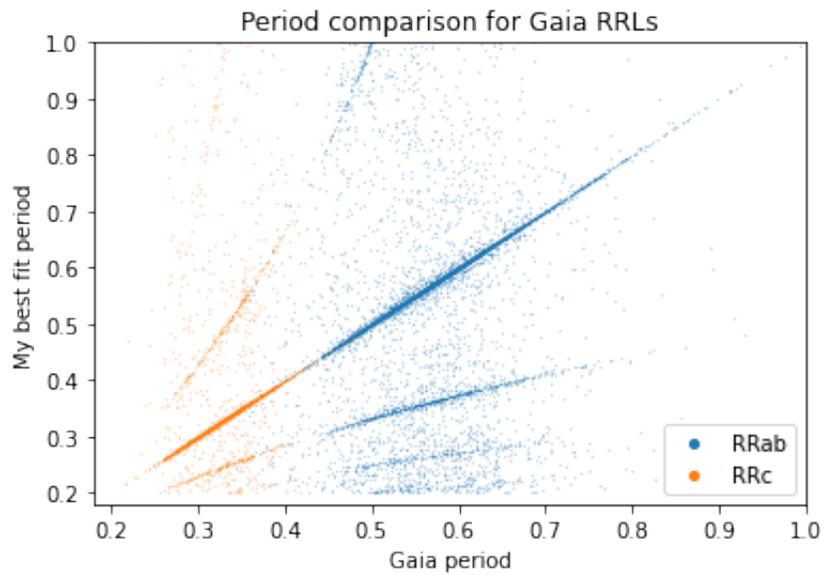


Figure 3.3: Best-fit NSC periods versus Gaia periods.

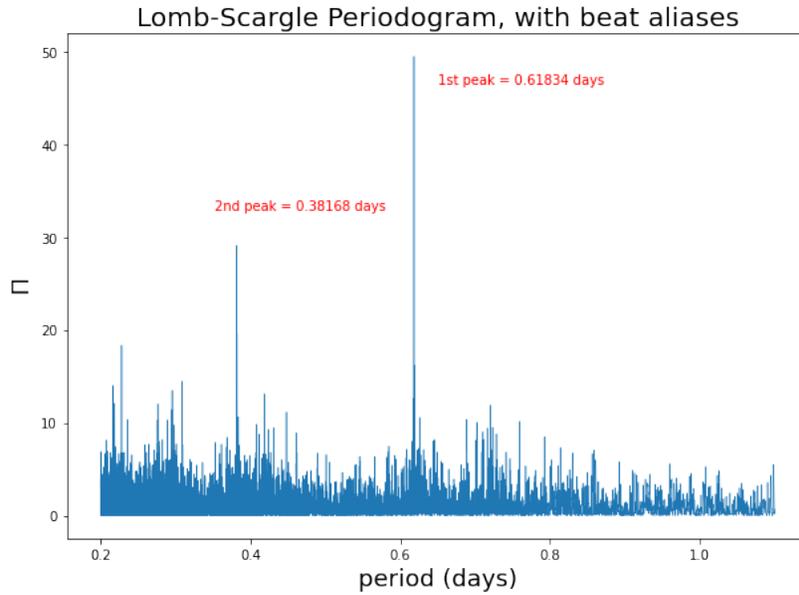


Figure 3.4: Lomb-Scargle Periodogram with a beat period shown.

and the rotation of the Earth are the frequencies interfering. Since a beat is the difference between two frequencies, we can easily express the failed period detection in terms of the correct period and the rotation period of the Earth (1 day).

$$\frac{1}{p_{obs}} = \frac{1}{p_{real}} - \frac{n}{1 \text{ day}} \quad n = (\dots, -2, -1, 0, 1, 2, \dots) \quad (3.2)$$

In the Lomb-Scargle periodogram example in Figure 2.3a, we notice a secondary peak to the left of the primary peak. Figure 3.4 shows this again, with the secondary peak labeled. This peak is associated with the beat pattern, being at approximately $(1/0.61833+1)^{-1} = 0.38208$ days. In that example above, this secondary peak was suppressed by the Lafler-Kinman statistic, but as data becomes more sparse the periodogram becomes more noisy and secondary signals like this can become dominant.

I have provided another example of a star from the Gaia catalog, for which Leavitt fit a beat frequency at a period of 0.34683 days, shown in Figure 3.5a. I reversed Equation 3.2 to solve for the corresponding period, $(1/0.34683 - 1)^{-1} = 0.53100$ days which is close

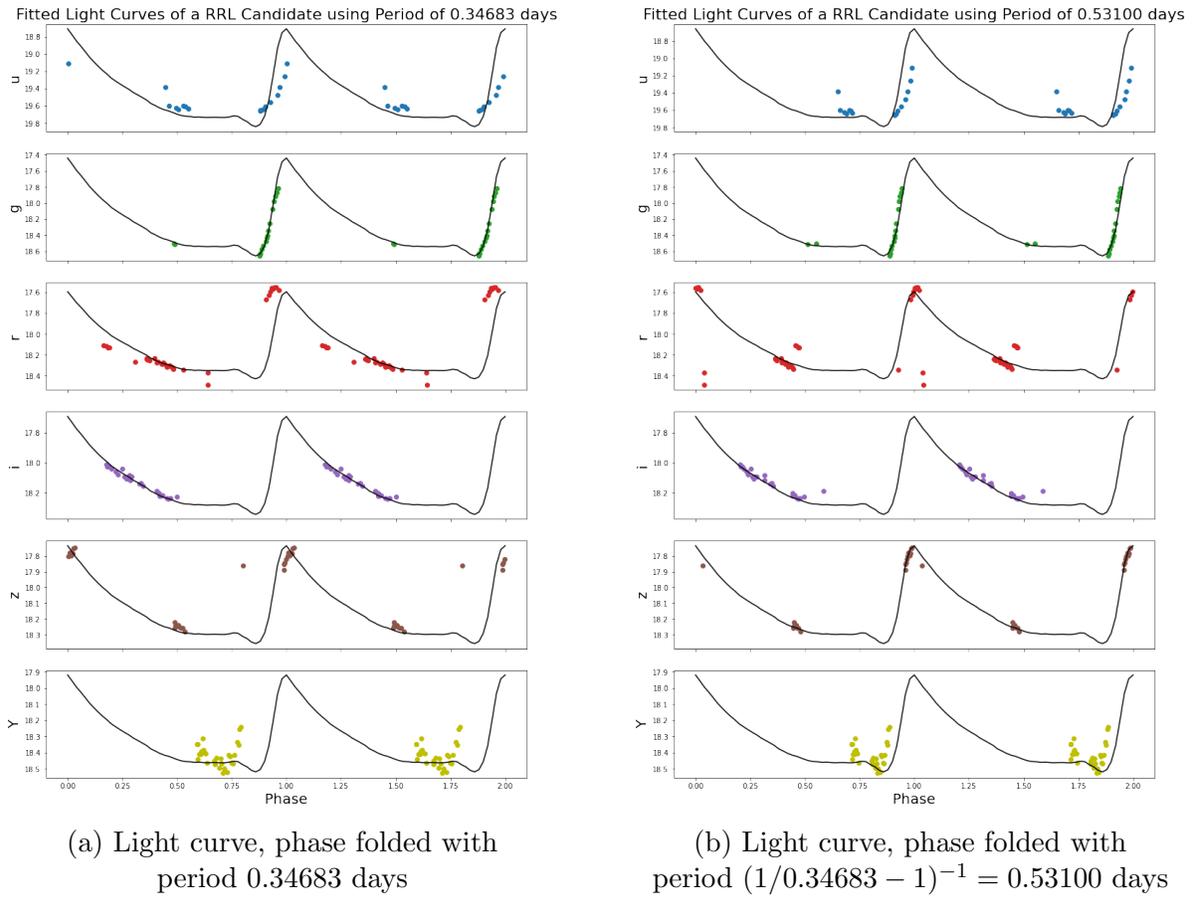


Figure 3.5: Light Curve example of an object with a Beat Period.

to the period listed in the Gaia catalog, 0.53174 days. The difference is small, so it makes sense that the wrong period could be detected here.

If the periodogram does not find a peak close enough to the true period, the template fitter will not be able to find it. For this reason, I recommend checking beat frequencies for the major peaks using Equation 3.2.

Finally, I have rerun the objects with $1/n$ days periods, rejecting these periods in the initial period search step, and reran the objects fit to beat frequencies, using 3.2 to test for the real period. This leaves me with results shown in Figure 3.7a. From this period-amplitude plot, I have selected the objects within the two prominent lobes, using the Gaia

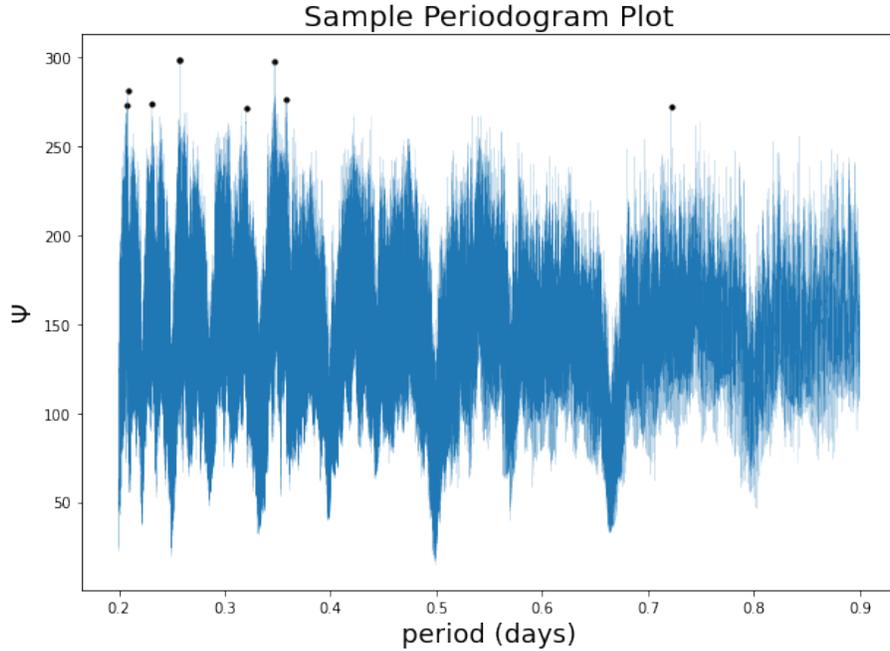


Figure 3.6: Periodogram of sparsely sampled object where beat period was detected.

catalog as a reference, (Figure 3.7b). The left lobe are RRL type C, the right lobe are RRL type AB. This gives us 17,843 candidate RRLs; 14,381 type AB, 3,462 type C. Of these, 9,198 were matched with RRLs in the Gaia catalog.

Distances can now be found for these objects. First, type C periods were fundamentalized using, $\text{Log}_{10}P_f = \text{Log}_{10}P_o + 0.128$. These periods can now be used with Equation 2.13 to obtain absolute magnitudes in i and z , a metal abundance of $Z = 0.001$ was assumed. The mean observed magnitudes were estimated using the minimum magnitude fit plus half the peak-to-peak amplitude. The difference between these two magnitudes (the distance modulus) can tell us the distance using Equation 2.14.

Figure 3.8a shows a histogram of the distances measured. The other three plots in Figure 3.8 show the 3-D spatial distribution of the RRLs by incorporating their Right Ascension and Declination with distance to calculate Galactocentric Cartesian coordinates. Here, Z points toward the celestial north pole, X points to the vernal equinox, and Y points

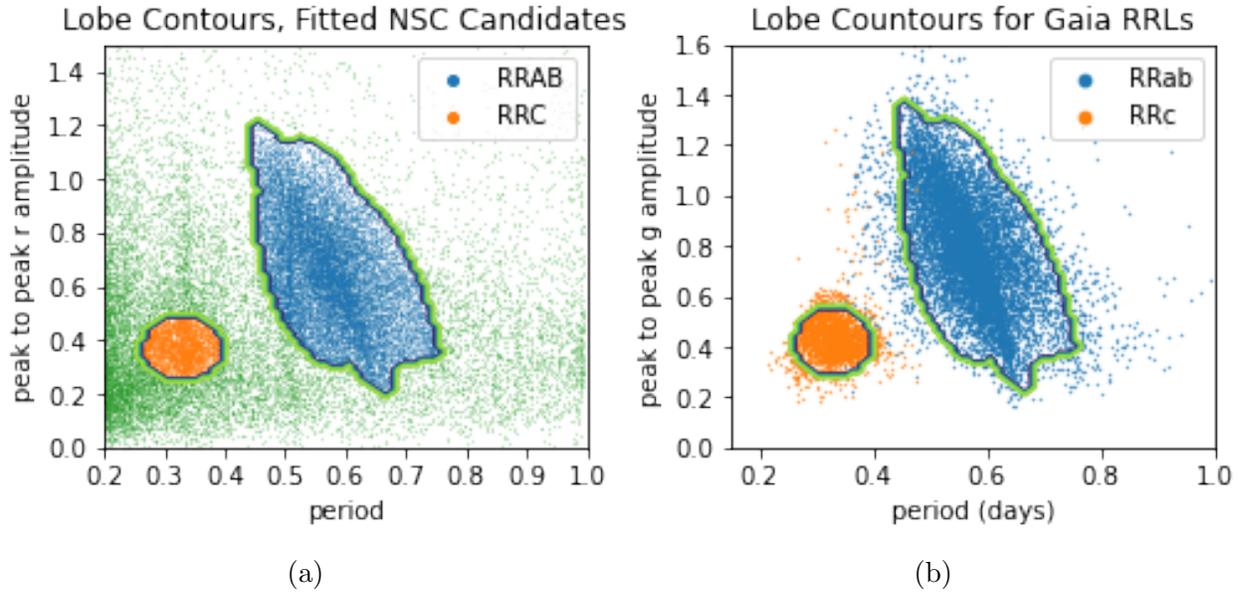


Figure 3.7: Period-Amplitude plot showing lobe selection.

toward a right ascension of 6 hours on the celestial equator.

I have highlighted a large grouping of RRLs in this sample that are distant and clustered. Based on their location, these must represent the Large and Small Magellanic Clouds. The Large Magellanic Cloud is approximately 50 kiloparsecs away from the Earth, and it is reassuring to see the cluster of stars centered at approximately 52 kiloparsecs. This also shows up prominently in the distance histogram in panel (a). The most distant RRL are at 70–80 kiloparsecs, indicating that the RRL in the NSC can be used to probe the outer reaches of our Milky Way galaxy.

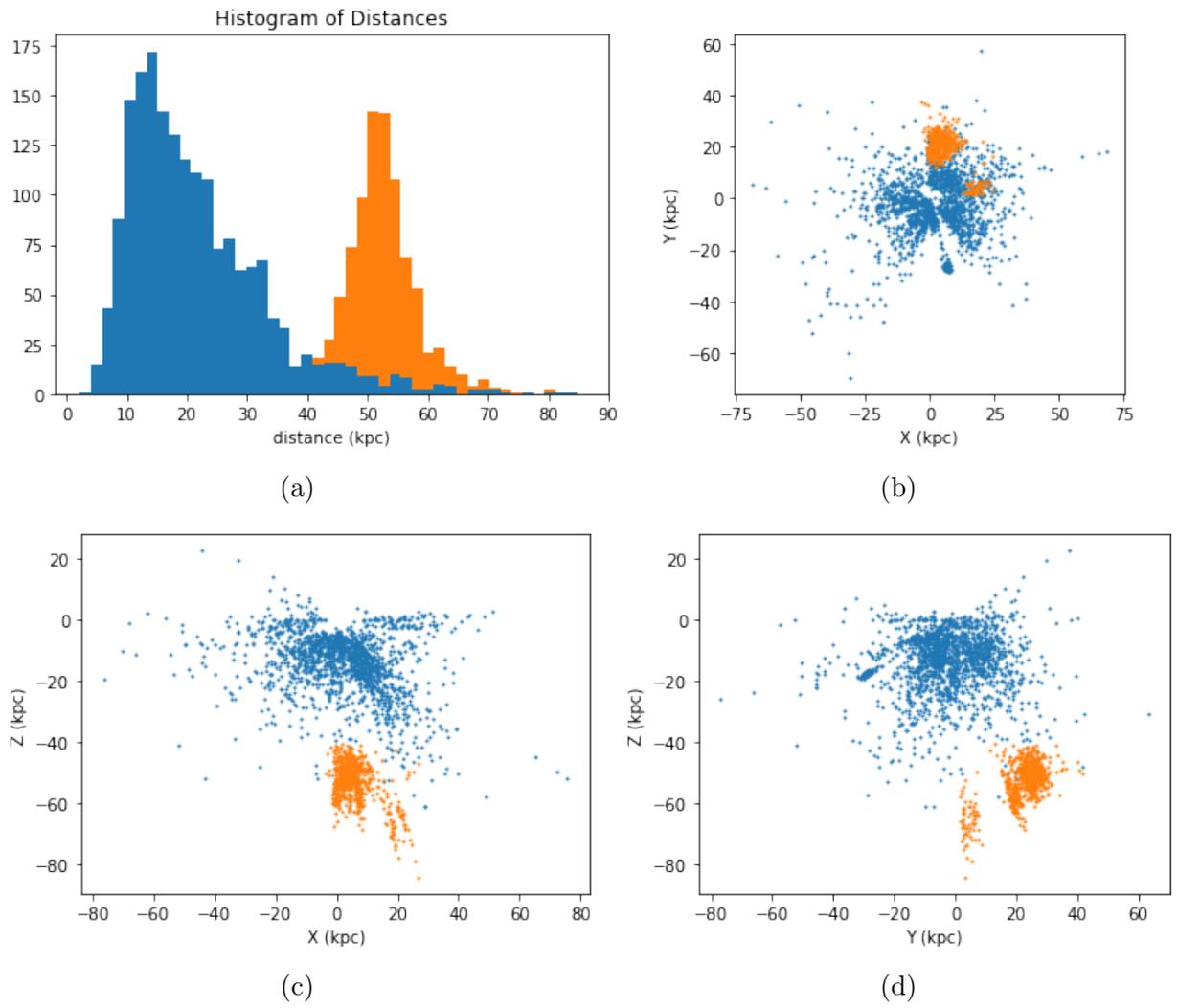


Figure 3.8: Spatial distribution of RRL sample. The Magellanic Clouds are highlighted in orange.

CONCLUSIONS AND FUTURE DIRECTIONS

The template fitting script, `Leavitt`, is promising. The fitted parameters for known RRLs in the Gaia catalog match the known values. The period and apparent magnitude parameters were used to derive reasonable distance measurements as evidenced by the matching Magellanic Cloud distance.

Classification between type AB and type C was also done based on best fitting template, these best fit classes match Gaia's classification well. Of the 10,015 type AB stars in the Gaia catalog and the NSC shortlist, `Leavitt` correctly classified 7,983. Of the 2,610 type C stars, 2,076 were correctly classified.

These results can be improved. In the Period-Luminosity-Metallicity relationship used in my distance estimate I assumed a universal metal mass fraction of $Z = 0.001$, but RRLs can range between 0.002 and 0.0001 which means actual distances can range from 5% closer to 10% further than determined here.

I only used the 7 Layden templates for this project, and the same were used for each band. As discussed in the methods section larger sets of templates exist, and the light curve in each band can be different, so different templates should be used for each band. Unfortunately, to test a different set of templates on each band requires repeating the fitting process exponentially more times. An interesting future area of study would be to determine what relationship there is for the light curve shape in different bands of the same object. If we can narrow the number of templates that have to be checked for one band based on which templates are being used in other bands, we could significantly cut down on the number of tests needing to be done.

Some variable stars have multiple modes they pulsate in, for example, RRL type D and many δ Scuti. Multi mode pulsators prove challenging for template fitting since there is not one but two periods. When folding the light curve by one period, the other continues to

vary the luminosity for the same phase. This has the effect of making the phased light curve messy. Therefore, no matter the template used it will always have large residuals and thus large χ^2 .

I briefly discussed the $1/n$ day periods detected. Many of these are caused by the observations happening around the same time each night and thus in the phased light curve, observations occur in a small section of the phase space while the rest is relatively empty. I currently detect this in a simple way by separating the phase values into bins and comparing the number of measurements within each bin. Professor Neil Cornish recommended a method described by Bruce Allen (source 2005) that was used for gravitational wave template fitting that discriminates based on measuring the χ^2 in a series of bins and seeing if the response in each is consistent with expectations. I was unable to fully understand and implement the method in time for the thesis. I believe this is still worth investigating to see if it can be used to reject periods when detections all line up in phase space because this likely means that the period was detected based on the schedule of observations and not the star's variability.

REFERENCES CITED

- C. Cáceres and M. Catelan. The period-luminosity relation of rr lyrae stars in the sdss photometric system. *The Astrophysical Journal Supplement Series*, 179(1):242, nov 2008. doi: 10.1086/591231. URL <https://dx.doi.org/10.1086/591231>.
- Arthur S. Eddington. On the Cause of Cepheid Pulsation. *Monthly Notices of the Royal Astronomical Society*, 101(4):182–194, mar 1941. ISSN 0035-8711. doi: 10.1093/mnras/101.4.182. URL <https://doi.org/10.1093/mnras/101.4.182>.
- Gaia Collaboration, Brown, A. G. A., Vallenari, A., Prusti, T., de Bruijne, J. H. J., Babusiaux, C., Bailer-Jones, C. A. L., Biermann, M., Evans, D. W., Eyer, L., and ... Gaia data release 2 - summary of the contents and survey properties. *A&A*, 616:A1, 2018. doi: 10.1051/0004-6361/201833051. URL <https://doi.org/10.1051/0004-6361/201833051>.
- J. Laffler and T.D. Kinman. Least-squares frequency analysis of unequally spaced data. *Astrophysical Journal Supplement*, 11:216, jun 1965. doi: 10.1086/190116. URL <https://doi.org/10.1086/190116>.
- Andrew C. Layden. Rr lyrae variables in the inner halo. i. photometry. *The Astronomical Journal*, 115(1):193, jan 1998. doi: 10.1086/300195. URL <https://dx.doi.org/10.1086/300195>.
- Henrietta S. Leavitt and Edward C. Pickering. Periods of 25 Variable Stars in the Small Magellanic Cloud. *Harvard College Observatory Circular*, 173:1–3, March 1912.
- N. R. Lomb. Least-Squares Frequency Analysis of Unequally Spaced Data. , 39(2):447–462, February 1976. doi: 10.1007/BF00648343.
- N. R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophys Space Sci*, 39:447–462, may 1976. doi: 10.1007/BF00648343. URL <https://doi.org/10.1007/BF00648343>.
- David L. Nidever, Arjun Dey, Katie Fasbender, Stéphanie Juneau, Aaron M. Meisner, Joseph Wishart, Adam Scott, Kyle Matt, Robert Nikutta, and Ragadeepika Pucha. Second data release of the all-sky noirlab source catalog. *The Astronomical Journal*, 161(4):192, mar 2021. doi: 10.3847/1538-3881/abd6e1. URL <https://dx.doi.org/10.3847/1538-3881/abd6e1>.
- Abhijit Saha and A. Katherina Vivas. A hybrid algorithm for period analysis from multiband data with sparse and irregular sampling for arbitrary light-curve shapes. *The Astronomical Journal*, 154(6):231, nov 2017. doi: 10.3847/1538-3881/aa8fd3. URL <https://dx.doi.org/10.3847/1538-3881/aa8fd3>.
- J. D. Scargle. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. , 263:835–853, December 1982. doi: 10.1086/160554.

- Jeffrey D. Scargle. Studies in astronomical time series analysis. ii. statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 265:835–853, dec 1982. doi: 10.1086/160554. URL <https://dx.doi.org/10.1086/160554>.
- Branimir Sesar, Željko Ivezić, Skyler H. Grammer, Dylan P. Morgan, Andrew C. Becker, Mario Jurić, Nathan De Lee, James Annis, Timothy C. Beers, Xiaohui Fan, and ... Light curve templates and galactic distribution of rr lyrae stars from sloan digital sky survey stripe 82. *The Astrophysical Journal*, 708(1):717, dec 2009. doi: 10.1088/0004-637X/708/1/717. URL <https://dx.doi.org/10.1088/0004-637X/708/1/717>.
- Branimir Sesar, Nina Hernitschek, Sandra Mitrović, Željko Ivezić, Hans-Walter Rix, Judith G. Cohen, Edouard J. Bernard, Eva K. Grebel, Nicolas F. Martin, and ... Machine-learned identification of rr lyrae stars from sparse, multi-band data: The ps1 sample. *The Astronomical Journal*, 153(5):204, apr 2017. doi: 10.3847/1538-3881/aa661b. URL <https://dx.doi.org/10.3847/1538-3881/aa661b>.
- Željko Ivezić, A. Katherina Vivas, Robert H. Lupton, and Robert Zinn. The selection of rr lyrae stars using single-epoch data. *The Astronomical Journal*, 129(2):1096, feb 2005. doi: 10.1086/427392. URL <https://dx.doi.org/10.1086/427392>.

APPENDIX A

TEMPLATE FITTER CODE

The following is the template fitting Python code used for this project.

```

import numpy as np
from scipy.optimize import curve_fit
from scipy.interpolate import interp1d
from astropy.table import Table

class tmpfitter:
    """
    Object used to fit templates to data.
    Initialize with the templates you plan to compare against and lists
    of the bands and amplitude ratios for those bands.
    Templates assumed to be in an Astropy table with one column for
    phase and additional columns for each unique template.
    Column names are used as template names. Filter names and amplitude
    ratios can be changed if different than these defaults.
    """
    def __init__(self, tmps, fltnames= ['u', 'g', 'r', 'i', 'z', 'Y', 'VR'],
                 ampratio=[1.8148, 1.38605, 1.0, 0.79662, 0.74672, 0.71875, 1.05078],
                 mode='RRL'):
        # constants
        self.tmps = tmps # Table containing templates
        self.fltnames = fltnames # list of names of usable filters
        self.Nflts = len(fltnames) # number of usable filters
        self.ampratio = np.array(ampratio)

        # model variables
        self.fltsiz = [] # list of filter index values
        self.tmpind = 1 # index of template (1,2,...,N) being used.
        self.period = 1
        self.mode = mode # 'RRL' or 'generic'

    def model(self, t, *args):
        """
        Modifies the template value using peak-to-peak amplitude and y
        offset to get magnitudes.
        Input times, t, are folded by period and phase shifted to match
        template.
        """
        t0 = args[0]

        # set amplitudes and y offsets based on mode
        if self.mode == 'RRL':
            amplist = (args[1] * self.ampratio)[self.fltsiz]
            yofflist = np.array(args[2:])[self.fltsiz]

```

```

elif self.mode == 'generic':
    amplist = np.array(args[1:-self.Nflts])[self.fltinds]
    yofflist = np.array(args[-self.Nflts:])[self.fltinds]

# fold time values to get phase then get template values
ph = (t - t0) / self.period %1
template = interp1d(self.tmps.columns[0],
                    self.tmps.columns[self.tmpind])(ph)

# Scale and offset templates to match magnitude measurements
mag = template * amplist + yofflist
return mag

def tmpfit(self, time, mag, err, fltinds, plist, initpars=None, mode=None,
           verbose=False):
    """
    Function that fits measurements, mag, to templates.
    time - time of measurements (in days)
    mag - magnitude measurements
    err - uncertainty
    fltinds - list of filter indices
    plist - list of periods to test
    """
    self.fltinds = fltinds

    if mode is not None:
        self.mode = mode

# namps and noffs used throughout to access parameters,
# representing the number of fitted amplitudes and offsets,
# which may differ by mode.
if self.mode == 'RRL':
    namps = 1
    noffs = self.Nflts
elif self.mode == 'generic':
    namps = self.Nflts
    noffs = self.Nflts

if isinstance(plist, (int, float)):
    plist = [plist]

# Set initial parameter guess if not given.
if initpars is None:
    initpars = np.zeros(1 + namps + noffs)
    ampests = np.zeros(self.Nflts)

```

```

initpars [0] = time [np.argmaxin (mag)]

for f in np.unique (fltinds):
    initpars [1 + namps + f] = min (mag [fltinds==f])
    ampests [f] = (max (mag [fltinds==f]) - min (mag [fltinds==f]))

if self.mode == 'RRL':
    r_apest = ampests / self.ampratio
    r_apest [r_apest <= 0] = np.nan
    initpars [1] = np.nanmean (r_apest)
    if verbose:
        print ("Initial_Parameters:")
        print ("t0:", initpars [0])
        print ("r_amp:", initpars [1])
        print ("y_offset:", initpars [2:])

elif self.mode == 'generic':
    initpars [1:-noffs] = ampests
    if verbose:
        print ("Initial_Parameters:")
        print ("t0:", initpars [0])
        print ("amps:", initpars [1:self.Nflts+1])
        print ("y_offset:", initpars [-self.Nflts:])

# Set boundaries to ensure no negative values
bounds = ( np.zeros (1+namps+noffs), np.zeros (1+namps+noffs) )
bounds [0][0] = 0.0
bounds [1][0] = np.inf
bounds [0][1:namps+1] = 0.0
bounds [1][1:namps+1] = 50.0
bounds [0][-noffs:] = 0.0
bounds [1][-noffs:] = 50.0

# Set boundaries of unused filters to near 0
for i in set (range (self.Nflts)) - set (self.fltinds):
    initpars [1 + namps + i] = 0
    bounds [1][1 + namps + i] = 10**-6
    if self.mode == 'generic':
        initpars [1+i] = 0
        bounds [1][1+i] = 10**-6

minx2 = 2**99
bestpars = np.zeros ( 1 + namps + noffs )
besttmp = -1

```

```

besterr = 0
bestprd = 0
for p in plist:
    self.period = p

    for n in range(1, len(self.tmps.columns)):
        self.tmpind = n

        # Fit template n using period p
        try:
            pars, cov = curve_fit(self.model, time, mag,
                                  bounds=bounds, sigma=err,
                                  p0=initpars, maxfev=9000)
        except RuntimeError:
            if verbose:
                print(f'curve_fit failed on template {n} and
                    period {p}')
            continue

        # x2 is used to measure the goodness of fit.
        x2 = sum((self.model(time, *pars) - mag)**2 / err**2)
        if x2 < minx2:
            minx2 = x2
            bestpars = pars
            besterr = np.sqrt(np.diag(cov))
            bestprd = p
            besttmp = n

    self.period = bestprd
    self.tmpind = besttmp

    if verbose:
        print("Results:")
        print("t0:", bestpars[0])
        print("amp:", bestpars[1:namps + 1])
        print("y_offset:", bestpars[-noffs:])
        print("Period:", bestprd)
        print("Best Template:", self.tmps.colnames[besttmp])
        print("Chi Square:", minx2)

return bestpars, bestprd, besterr, besttmp, minx2

```