PeRiodic Infrared Milky-way VVV Star-catalogue : PRIMVS

Niall MILLER

Contents

List of Figures

List of Tables

Chapter 1

Abstract

We present the PeRiodic Infrared Milky-way VVV Star-catalogue - 'PRIMVS'. We utilise the VVV survey's unique depth and breadth to investigate the variability of astronomical sources within the Galactic bulge and disk. There is a focus on an unbiased and complete identification and classification of periodic variable stars. Employing internal metrics from the VIRAC table for initial selection, we meticulously clean and preprocess light curves to increase reliability and completeness. Care has been taken to address photometric contamination and other sources of uncertainty.

Our approach includes constructing periodograms using Lomb-Scargle, Phase Dispersion Minimisation, Conditional Entropy, and Gaussian Processes to ascertain periodicity.

This above process allowed us to curate a catalogue of 86,507,172 candidate variable sources.

Machine learning techniques, particularly decision trees and autoencoders, facilitated the initial steps in classification of a significant portion of these sources.

Chapter 2

Introduction

The PeRiodic Infrared Milky-way VVV Star-catalogue - 'PRIMVS' aims to provide a thorough and reliable catalogue of all periodic variable stars present within the VVV survey. The identification of these variables is achieved by the use of parameters present in the VIRAC (Smith et al., 2018) database followed by a variability based selection after light curve cleaning. The compute power of the University of Hertfordshire cluster was heavily utilised for both parallelisation (with a high core count of 128) and the use of GPUS. Care was taken to ensure a minimal amount of quasi-periodic, and otherwise difficult to detect, periodic variables were missed. A Quasi-periodic source is a source whose periodicity is irregular. This irregular behaviour can be caused by an aperiodic change in: period, amplitude, average magnitude or some combination of these. Many statistical measures separate from the identification of a period were made (section 6). These statistics serve to provide a full picture of the certainty and reliability of an extracted period and to produce astronomical information, allowing further identification of the source

Chapter 3

Candidate Selection

Due to the uniqueness of the VVV survey, all sources selected for analysis were done so exclusively using internal metrics. An initial selection is made using the variability metrics found in the VIRAC table. These selections are highlighted in table 3.1

Ks detections > 50	Ensure we have at least 50 measurements
Ks detections $> 0.6\,Ks$ observations	Ensure the source is detected at least 60% of the time
$\sigma_{Ks} > 0.01$	Ensure some variability
$\sigma_{Ks}/Ks\_ivw\_err\_mag > 4$	Ensure variability is above some measure of noise

TABLE 3.1: Variability selections performed on VIRAC metadata, prior to light curve cleaning. Where 'σ_{Ks}' is the standard deviation and '*Ks_ivw_err_mag*' is the inverse variance weighted error

These are relatively loose selections aimed at completeness. Due to the high amount of unreliable measurements (as much as 60% in crowded regions) in VIRAC light curves we can't fully rely on variability metrics calculated from the raw light curve. After the initial VIRAC variability selection, the light curve is retrieved and cleaned (see section 4). A second check for variability is then made, selections for which are shown in table 3.2.

K_s detections > 50	Reaffirm we have at least 50 measurements after cleaning
Ks error < 0.5	Ensure we have sensible uncertainty
$KsQ_{99} - KsQ_{01} > 0.1$	Ensure there is a minimum of 0.1 mag variability
$KsQ_{75} - KsQ_{25} > 2\text{median}(Ks\text{error})$	Ensure inter-quartile variability is above twice the uncertainty

TABLE 3.2: Variability selections performed after cleaning the light curve

After this selection the light curve is processed as described in section 5. Figure 3.1 outlines the process for selecting variables in the VVV data. After selection we are left with 86,507,172 candidate variable sources.

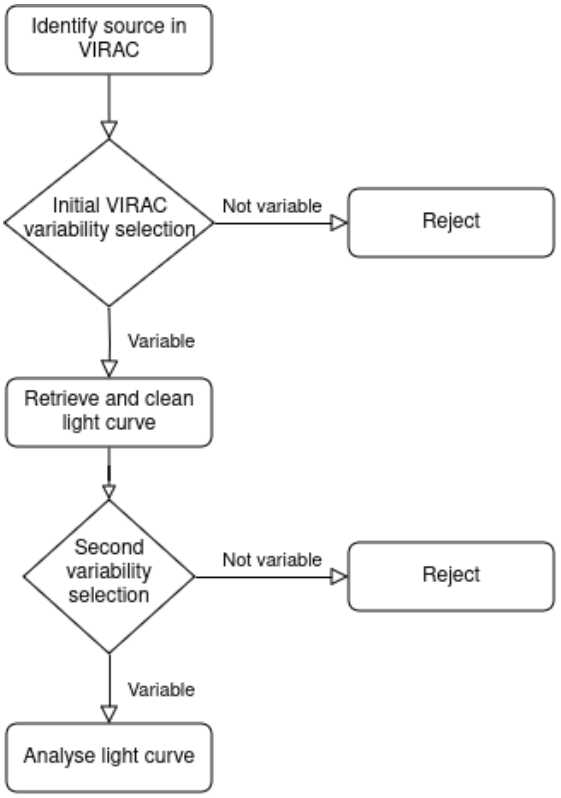


FIGURE 3.1: Flowchart showing the selection process for astronomical sources based on their variability.

Chapter 4

Light Curve Preprocessing

From a practical standpoint, most period-finding methods are relatively simplistic (irrespective of their mathematical complexity). Hence, a large portion of the robustness we achieve in our analysis comes from thoroughly pre-processing a light curve such that a period-finding method will have its effectiveness maximised. This process involves first cleaning the light curve so every measurement is as reliable as possible, and then modifying the light curve, allowing for a more accurate analysis. Due to the depth and survey area of the VVV survey, photometric contamination is a common occurrence.

VIRAC provides multiple metrics of reliability for each measurement in a light curve. Figure 4.1 shows the 'ast_res_chiq' vs 'chi' of a light curve with the dashed line signifying the selection cuts used. Where 'chi' is the DoPhot Chi parameter, representing the quality of the profile fit and 'ast_res_chisq' represents the quality of the 5 parameter astrometric fit to position, proper motion and parallax. It can be seen that the majority of the measurements cluster below the cuts. These cuts were determined with the intention of removing photometry most commonly affected by photometric contamination. This does not serve to remove bad photometry caused by saturation however. It is likely these will also have higher 'chi' values but we do not need to remove them as they will still contribute to any apparent periodicity. If a star is sufficiently saturated such that the photometric error is problematic, both 'chi' and 'ast_res_chisq' should reflect this and flag the point for removal. There exists a trade off between completeness of the light curves and reliability. Through iterations of the PRIMVS pipeline it has been observed that points with high 'chi' help make the catalogue more complete for bright pulsating stars near the saturation limit. This will enable us to extract a likely period even if amplitudes may be under-estimated. A blanket rejection of points with a magnitude error 0.2 is also applied.

We can also utilise the observing pattern of VVV to both increase the reliability of our data and the photometric certainty. The aforementioned 'paw-print pairs' can be used to check if two points taken close together in time are similar. We do this by ensuring that any pair of data

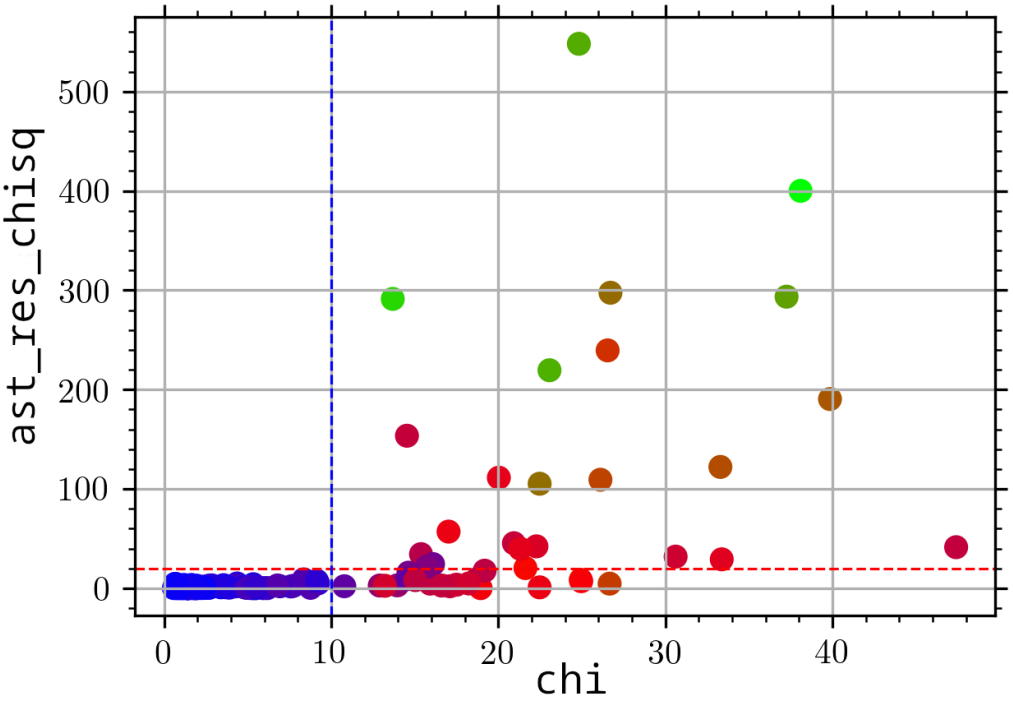


FIGURE 4.1: Showing the range of values taken for a light curve with varying quality of points. Where 'ast_res_chi' and 'chi' are astrometric values taken from the DoPHOT. Colour is proportional to magnitude.

points have similar magnitude errors and are within $2 \times m_{err}$ of each other. If either are not, both are rejected. After this, any data point within 1 hour of each other with $m_{err} > 0.1$ are combined by binning 'N' measurements such that $\sigma_{new} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sigma_i$. Any light curve with fewer than 40 measurements after this process is removed from future processing.

After removing erroneous data and combining close points, we move to modifying the remaining data with the goal of making the periodic signal the only photometric variable. A straight line is fitted and subtracted to the light curve, as can be seen in figure 4.3. The linear model is fitted to ensure there are no linear trends throughout the light curve.

AGB stars can feature strong periodic variability on top of a long term variability (Höfner and Olofsson, 2018). While both of these sources of variability provide useful astrophysical information, we are focused on robust period extraction.

Our period-finding methods assume only one source of variability. This can cause an otherwise correctly phase folded light curve to look messy or even incorrect. For most methods, a sufficiently strong linear trend would render all periodic signal extraction virtually impossible.

To subtract the straight-line we first bin the light curve into 10 bins. The weighted median is calculated for each bin and the straight line is fitted to the resulting 10 points. This is done to

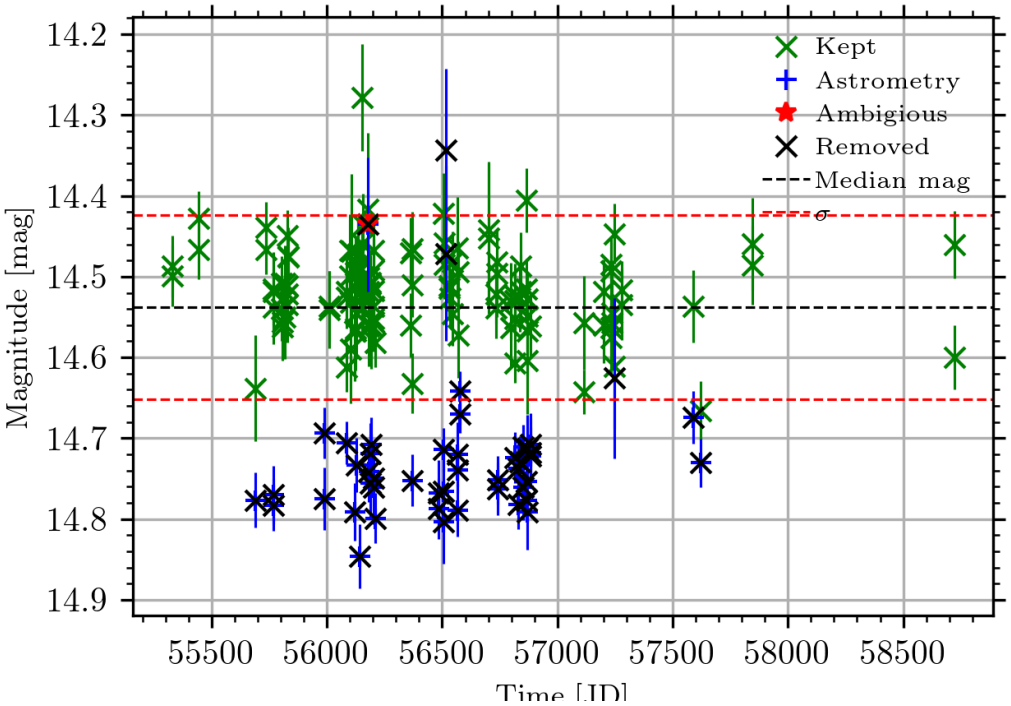


FIGURE 4.2: Showing the points which are removed from a light curve after cleaning. The blue 'Astrometry' points are those that fell outside of the cuts seen in figure 4.1. The red 'Ambiguous' points are determined by a boolean flag which signifies if the source appears blended with a neighbour. The green points are deemed reliable and used for analysis.

best capture an overall trend separate to periodic variability whilst also accounting for erroneous outliers. The straight-line is only subtracted if $dm/dt > 2 \times 10^{-4}$ mag/day and $R^2 > 0.2$, where R^2 is the coefficient of determination ($R^2 = 1 - \frac{RSS}{TSS}$, where 'RSS' is the sum of squares of residuals and 'TSS' is the total sum of squares)

The aperiodic form of variability may not be a linear trend and we could fit a higher-order polynomial. However, fitting a higher-order polynomial runs the risk of potentially removing or modifying the periodic source variability, particularly those of longer periods.

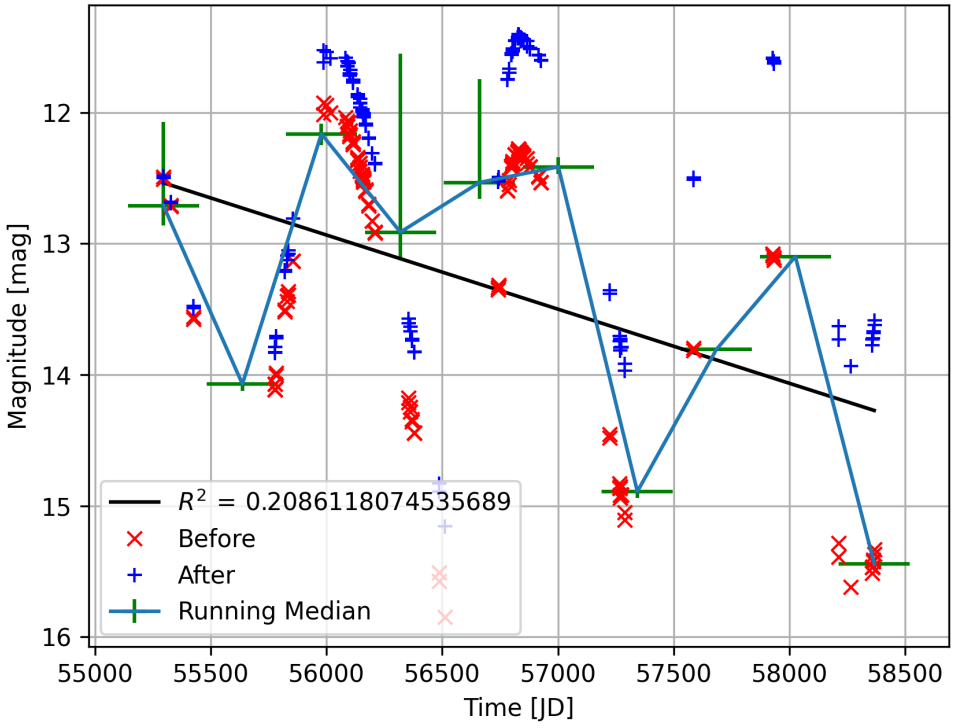


FIGURE 4.3: Showing the measurements before the removal of the linear trend in red crosses along with the fitted black line. The measurements after the removal can be seen as blue plusses '+'.

Chapter 5

Time-series Analysis

After the light curve has been cleaned and prepared for analysis, we can construct a periodogram. We compute periodicity using the Lomb-Scargle (LS), Phase Dispersion Minimisation (PDM), Conditional Entropy (CE) and Gaussian Processes (GP) methods. In an effort to increase the flexibility of light curve analysis, much of the completed testing identifies strengths and weaknesses of each of these methods.

Lomb-Scargle For our implementation of the Lomb-Scargle periodogram, we have used `AstroPy` (Astropy Collaboration et al., 2013). Within the `AstroPy Time Series` package there is a pre-built Lomb-Scargle algorithm (Astropy Collaboration et al., 2013)[1]. We also set '`FIT_MEAN = True`' which enables the Lomb-Scargles generalisation (Zechmeister and Kürster, 2009) for help with smaller data-sets and uneven light curves. It takes an average of ≈ 0.4 seconds to compute a 100000 sample periodogram of a source with 200 data points on 1 core and 1 GB of RAM.

Phase Dispersion Minimisation Our version of the PDM periodogram is taken from the original PDM2 source code written in C with edits for efficient interfacing with the python pipeline[2]. We use the updated PDM2 method which has the addition of 'subharmonic sampling'[3]. This uses the PDM window transform to more smoothly bin the phase folded light curve. This allows for clearer differentiation between harmonics of the true period. The python pipeline calls the PDM binary file and passes light curve and periodogram input information via a temporary file. Temporary files are a notable hindrance to the speed of this method as hard disk input/output (IO) operations are orders of magnitude slower than volatile memory operations.

[1] `https://docs.astropy.org/en/stable/api/astropy.timeseries.LombScargle.html`
[2] `https://www.stellingwerf.com/rfs-bin/index.cgi?action=PageView&id=34`
[3] `https://www.stellingwerf.com/rfs-bin/index.cgi?action=GetDoc&id=21&filenum=1`

The unreliability of Python's memory management renders temporary files the only appropriate method of sending and receiving data between an external executable. A future version of the PRIMVS pipeline will seek to fix this obvious bottleneck in compute speed. It takes an average of ≈ 0.2 seconds to compute a 100000 sample periodogram of a source with 200 data points on 1 core and 1 GB of RAM.

Conditional Entropy The CE method is a phase folding technique that uses a very similar method to PDM (Graham et al., 2013). Much like PDM, CE phase folds the light curve for each trial period, then bins the data, and then measures some quantity of the 'scatter' in each of the bins.

This method is fundamentally different from PDM in the way it calculates the 'scatter' of the data points however. Conditional entropy measures the conditional entropy of the phase folded light curve and uses this to quantify the 'scatter' of the data points.

Equation 5.1 describes the conditional entropy $H(m|\phi)$ of a light curve with magnitude 'm' and phase 'ϕ'.

$$H(m|\phi) = \sum_{i,j} p(m_i, \phi_j) ln \left(\frac{p(\phi_j)}{p(m_i, \phi_j)} \right) \tag{5.1}$$

Where $p(m_i, \phi_j)$ is the probability that a data point will occupy the i^{th} magnitude bin of 'm_i' and j^{th} phase bin of 'ϕ_j'. '$p(\phi_j)$' is the probability a data point will occupy the j^{th} bin, which in our case reduces to:

$$p(\phi_j) = \sum_i p(m_i, \phi_j) \tag{5.2}$$

Configuring the amount of bins used for the phase axis is crucial in the Conditional Entropy (CE) method, as it directly affects the sensitivity and accuracy of detecting periodic signals. The jackknifing method from Hogg (2008) is used to determine the number of magnitude bins to use for each light curve. However, it would be computationally expensive to calculate the optimal bins for the phase axis in each trial period. We opted for 10 phase bins based on recommendations from Graham et al. (2013), striking a balance between resolution and noise. Fewer bins may result in a loss of resolution, making it challenging to detect subtle variations in light curves. This will be particularly problematic for sources with complex variability patterns like pulsating stars or eclipsing binaries. Conversely, using too many bins can lead to overfitting, where the conditional entropy becomes dominated by noise rather than genuine signal features. This will result in a noisy representation of the phase distribution. Future work may explore

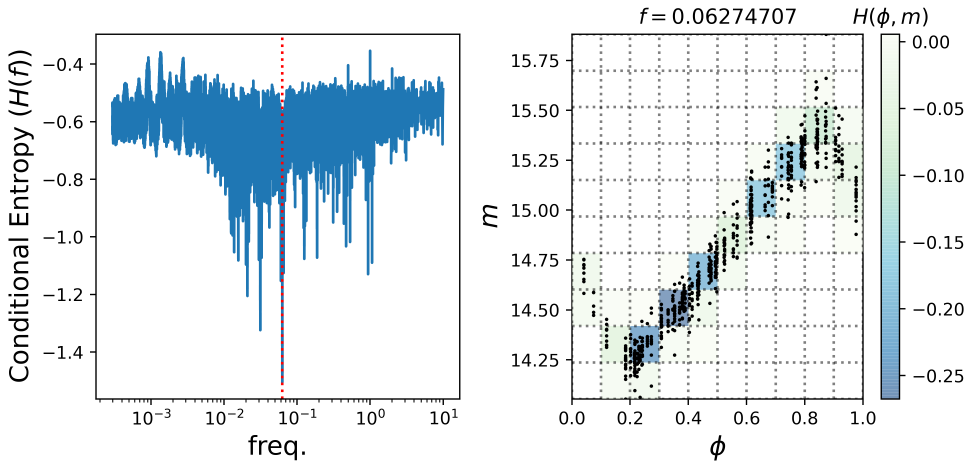


FIGURE 5.1: **Left**: A plot showing conditional entropy as a function of frequency. **Right**: Showing the phase folded light curve that produces the lease total conditional entropy. In this plot we can see each of the bins used as well as the conditional entropy each of them hold.

adaptive binning techniques that adjust the number of bins based on the light curve's characteristics, potentially improving sensitivity across different variable star types and increasing the robustness of period detection.

Figure 5.1 shows the periodogram (left) and the phase folded light curve when plotted at the optimal period (right). This figure highlights how this method works as well as the importance of the bins.

We can see from figure 5.1 that if we were to reduce the number of bins, the resolution of the process would effectively decrease as it would be harder to differentiate between small changes in the shape of the light curve. However, if we increase the number of bins too much, the conditional entropy would be dominated by small changes. This results in a noisy representation of the phase distribution.

For our implementation of CE, we have used the python package '`cuvarbase`'. The CE periodograms were not constructed as part of the main three tests. The computational intensity of CE renders it a GPU bound operation. The University of Hertfordshire High Performance Cluster features 6 GPU nodes, each with at least the computational equivalent of 3 Tesla A100 16GB GPUs. It typically takes ≈ 1 second to recover a period, as opposed to the ≈ 20 minutes for the same process when CPU bound. However, there are only 32 normal cores with 32 GB of RAM for each of these GPU nodes. This is significantly less than the 256 cores and 128 GB of RAM that is used for the combined LS and PDM test. At the time of writing, the CE periodogram is only computed for sources with multiple distinct periods with a low FAP in an attempt to clarify ambiguities. A future version of PRIMVS will be made where significantly more/all of the sources are analysed with CE.

Gaussian Processes A periodic signal observed in astrophysics is rarely a perfect sinusoid. Often periodic signals in this field vary in non-sinusoidal and Quasi-Periodic (QP) ways. To effectively model this behaviour we would ideally have a small number of parameters that are flexible enough to properly describe real astrophysical signals. In Rasmussen and Williams (2006) Gaussian Processes are described as providing a "...principled, practical, probabilistic approach to learning in kernel machines." Gaussian Processes are unique in our comparison of period finding techniques, as their ability to identify a periodic signal is only a product of the kernel used. Through different kernels and different combinations of kernels, Gaussian Processes can model many patterns within data. The flexibility present in Gaussian Processes is from their modelling of the covariant structure of the data, rather than the absolute values of data. This means that a relatively simple kernel is likely to be able to describe the structure of many light curves.

There are many kernels, and combinations thereof, available to use for Gaussian Processes. In Rasmussen and Williams (2006) the Quasi-Periodic kernel is used to measure the concentration of CO_2 on the summit of the Mauna Loa volcano in Hawaii. To achieve this, a product of two basic kernels are used; the squared exponential kernel and the periodic kernel. In Angus et al. (2018) GPs are used to identify the often quasi-periodic nature of stellar rotation periods, the QP kernel is used.

$$k_{i,j} = \mathrm{A}\exp\left[-\frac{(x_i - x_j)^2}{2l^2} - \Gamma^2\sin^2\left(\frac{\pi(x_i - x_j)}{P}\right)\right] + \sigma^2\delta_{i,j} \tag{5.3}$$

Where '$k_{i,j}$' is the covariance between points 'x_i' and 'x_j'. 'A' is the amplitude factor, scaling the overall covariance. '$\exp\left[-\frac{(x_i-x_j)^2}{2l^2}\right]$' is the radial basis function (RBF) which models the smooth variation in the data. 'l' is the length scale of the RBF kernel, controlling how rapidly the similarity between two points decreases as their distance increases.

'$\exp\left[-\Gamma^2\sin^2\left(\frac{\pi(x_i-x_j)}{P}\right)\right]$' is the periodic component of the kernel, where 'P' is the period, and 'Γ' adjusts the relative importance of the periodic versus RBF component.

'$\sigma^2\delta_{i,j}$' represents the noise term, where 'σ^2' is the variance of the noise, and '$\delta_{i,j}$' is the Kronecker delta function, equal to 1 if $i = j$ (i.e., for the diagonal elements representing the variance at each point) and 0 otherwise. This term tries to account for uncorrelated noise measurements, the presence of a Kronecker delta asserts that each measurements noise is independent.

The GP kernel is minimised with both '`scipy minimise`' and '`emcee`' python packages. The '`scipy minimise`' package uses least-squared regression which can struggle with the number of free parameters in the dataset. This is mostly used to provide an initial position for each of the walkers in the Monte Carlo Markov Chain (MCMC) process whic follows. The MCMC minimisation utilised 32 walkers and 500 steps. Statistical analysis is performed on the last 50 steps

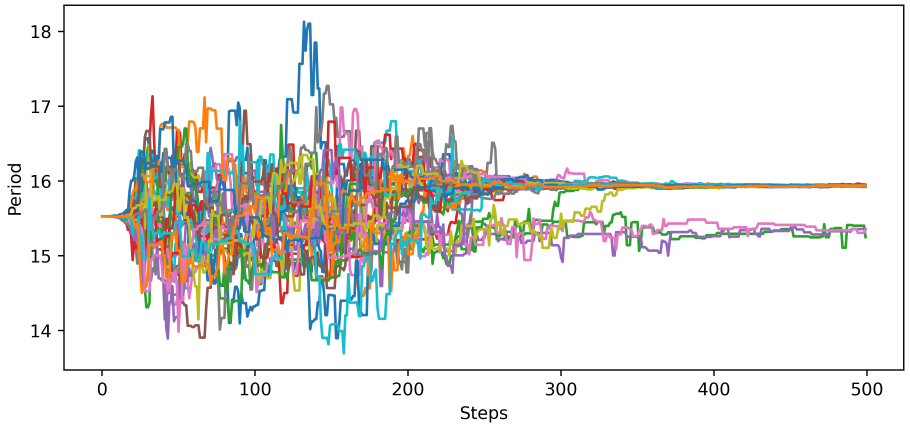


FIGURE 5.2: Showing the path the 32 walkers took in their 500 steps they made within the MCMC process. It can be seen that after $\approx$ 350 steps the majority of the walkers fall into what is approximately the same value for the period with a few which do not.

of the MCMC minimisation process. Figure 5.2 shows the path the walkers took under MCMC minimisation. It can be seen that a small number of walkers deviate from the majority. Without removing the values these walkers represent, any averages drawn from the total output are at risk of being erroneously shifted by these walkers. In the future, it might be possible to model systems with multiple periods by identifying separate clumps of walkers. For this iteration of the PRIMVS pipeline, the GP is run if the straight line fit has $dm/dt > 2 \times 10^{-4}$ mag/day(from the end of section 4), FAP > 0.2, Amplitude > 0.5 and there are more than 100 measurements. This is done to save compute and utilise the GPs ability to identify periodicity with additional trends present. Much like CE, the intention is to analyse most/all of the PRIMVS catalogue with this method.

5.1 Period Searches

A periodogram will be constructed from a list of test periods. Period finding methods will take a set of test periods, apply that period to the light curve (either via phase folding or fitting) and then measure some feature of the light curve (goodness of sine fit, binned scatter...). In order to maximise completeness in a blind search for periodicity we must ensure our set of trial periods does not impart a selection bias. It is common practice to search for periodicity linearly in frequency space (Chen et al., 2020) as doing so in period space will disproportionately compute for longer period variable stars (e.g. a periodic variable star will look fine when phase folded at 50.1 days given a true period of 50 days. This is not the case for a star with a period of 5 days which has been phase folded at 5.1 days).

Test #	Period range
Test 1	1 d <P <500 d
Test 2	0.01 d <P <1 d
Test 3	500 d <P <$T_{lc}/2$

TABLE 5.1: Table showing each of the three successive period scans used

For most completeness we split our period search into three successive searches of 100,000 trial periods, the ranges of which are seen in table 5.1. For each of the three tests we remove any sources with a FAP < 0.1 from future tests to save on compute.

Where '$\frac{T_{lc}}{2}$' is the length of the light curve divided by two, ensuring a minimum of 2 cycles are seen. Figure 5.3 shows the pipeline for the period analysis.

5.2 Periodogram

Another important area for ensuring the thorough analysis of a time-series data set is the treatment of the constructed periodogram. For each periodogram, we first exclude known problematic areas, such as those from the lunar, diurnal and yearly cycles. Then the three most significant peaks are identified. This is not always the 3 highest distinct points on the periodogram. To identify a peak we must also find a trough on either side. This stops us from extracting a particularly wide peak twice or extracting a peak which is instead just the edge of the periodogram. It also allows us to characterise the peak width and height.

After each periodogram has its 3 most prominent peaks extracted, their corresponding periods are measured and compared to each other. The FAP which we used allows for the universal comparison of periodicity regardless of the method used to identify the periodicity. This allows us to mitigate a lot of the biases that are exclusive to either method. For example; LS struggles more than PDM with the identification of periodicity for eclipsing binaries (VanderPlas, 2018) and so in cases where LS will fail, PDM should succeed and be recognised by the lower FAP. This method also allows us to account for some of the effects that sampling, aliasing and other perturbers may have in hiding the true period amongst other high peaks. If the true period is within the top 3 most significant peaks then it should be identified by the neural network FAP, which operates independently of the periodogram.

5.3 Period Alias Check

A common issue with the analysis of periodicity is knowing which alias of the period is correct; that is, if we extract two similarly likely periods at P and $0.5 \times P$ or $2 \times P$, which one is correct?

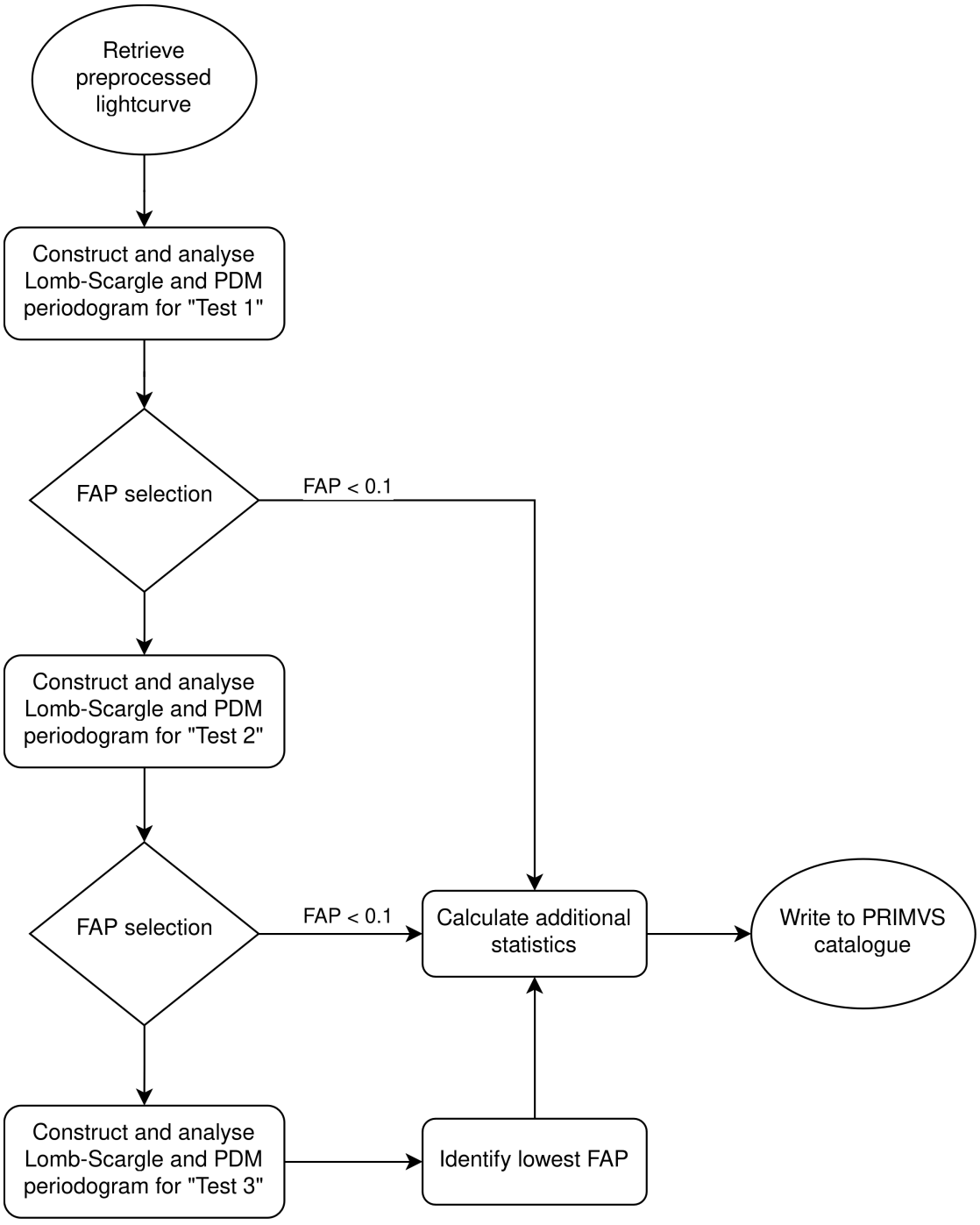


FIGURE 5.3: Flowchart showing the pipeline for processing the light curves through each of the three tests seen in table 5.1. Sources identified with a FAP<0.1 are removed from future tests. Each test consists of 100,000 trial periods.

It is typical to construct a periodogram with a resolution either linear or logarithmic in frequency space (i.e not linear in period space). This is done to sample the frequency space as completely as possible while still being computationally feasible. If we have a source with a period towards the end of our linearly sampled periodogram, it could be possible that we fail to sample at the specific period of the source. However, it could be more likely we sample at half of that period (as the sampling density in period space will increase towards shorter periods). This could erroneously heighten the peak at this period above that of the true period.

While this problem typically hinges on the ability to confidently determine which phase folded period is actually correct (which is often impossible), we can at least ensure a fair test is performed for each trial period. This is achieved by recomputing the periodogram with an increased density of trial periods at previously identified significant peaks. This ensures that we do not miss the true period due to insufficient sampling density. After we have obtained the aforementioned list of unique candidate periods we can recompute the periodogram before calculating a FAP.

At the time of writing, this method computed for all sources with FAP < 0.1 and P > 1000 days (1,028,397 sources). This method is computationally expensive and so was only used on what is expected to be the most secondarily affected sources. The intent is to use this approach of recomputing periodograms for all objects. For the first iteration of PRIMVS, we exclude period ranges that are close to known diurnal and lunar aliases, such as 1 day, 2 days, 29 days, and 60 days. This exclusion applies to these specific period ranges, not to the variables themselves, which might otherwise exhibit periodic behaviour. By avoiding these alias-prone ranges, we aim to reduce the likelihood of misidentifying false periods as genuine. These cuts were selected based on the high susceptibility to observational artefacts within these ranges, which can produce misleading peaks in periodograms. While this approach helps minimise erroneous detections, it also potentially overlooks true periodicities that coincide with these alias ranges. To address this limitation, future iterations will explore the integration of machine learning techniques, similar to those employed by Christy et al. (2023), to better distinguish between genuine and false periods.

The typical readout time for the VIRCAM detectors is around 1 second, with additional overhead for data processing and telescope jitter movements. This results in a timing precision for individual measurements of a few seconds. Given this timing precision, the theoretical minimum period that can be resolved is a few seconds. However, in practice, periods shorter than a few minutes may be challenging to detect reliably due to noise, stochastic sampling, and additional overheads. No uncertainty for time is given in the VIRAC data.

Chapter 6

Further statistics and False Alarm Probability

After the periodograms have been constructed and analysed the FAP is calculated. Other statistics, particularly relating to the magnitude distribution, are also calculated.

The following is a description of each of the (groups of) features in the PRIMVS catalogue. The PRIMVS catalogue has two distinct sets of features: light curve features and periodogram features.

mag_n, time_range, mag_avg and magerr_avg Number of points, time range, median magnitude, and median magnitude error in the cleaned light curve (i.e. the light curve that was used for analysis which is different from the raw light curve)

The mean, standard deviation, skew and kurtosis of the light curve is also calculated. The error weighted counterpart for each of those values is also used.

The following representations of magnitude and error will be used for all further feature definitions; m, $\bar{m}$, σ_m as magnitude, mean magnitude and magnitude standard deviation respectively. Magnitude error is denoted as m_{err} with the same variations used for magnitude ($\bar{m}_{err}$, σ_{merr}).

true_period The most likely period '*true_period*' is the potential period which had the lowest FAP

best_fap The lowest FAP from each of the extracted potential periods. i.e. the FAP which is obtained from the '*true_period*'

Cody_M Measure of asymmetry 'M' from Cody et al. (2014).

$$M = \frac{\bar{m}_p - \text{median}(m)}{\sigma} \tag{6.1}$$

where

$$\bar{m}_p = \frac{1}{N_p} \sum_{m_i \in P} m_i \tag{6.2}$$

and

$$P = \{m_i \,|\, m_i > Q_{90}(m) \text{ or } m_i < Q_{10}(m)\} \tag{6.3}$$

where '$Q_{90}(m)$' and '$Q_{10}(m)$' are the 90^{th} and 10^{th} percentiles of the magnitude distribution. 'N_p' is the number of points in the set 'P', and 'σ' is the overall root mean square of the magnitude distribution.

Stetson_K Robust measure of kurtosis '$Stetson\_K$' (Stetson, 1996)

$$Stetson\_K = \frac{1/N \sum_N^{i=1} |\delta_i|}{\sqrt{1/N \sum_N^{i=1} \delta_i^2}} \tag{6.4}$$

where the relative error 'δ' is defined as

$$\delta = \sqrt{\frac{N}{N-1}} \frac{m - \bar{m}}{m_{err}} \tag{6.5}$$

Where the number of points in the light curve is N.

von Neumann η and η_e The von Neumann variability indices 'η' (Neumann, 1941) and 'η_e' (Kim et al., 2014) was developed as a check for whether successive data points are independent.

$$\eta = \frac{\sum_{i=1}^{N-1} (x_{n+1} - x_n)^2 / (N-1)}{\sigma_m^2} \tag{6.6}$$

However, this assumes we have evenly spaced samples and so we also have

$$\eta_e = \bar{w} (t_{N-1} - t_1)^2 \frac{\sum_{i=1}^{N-1} w_i (m_{i+1} - m_i)^2}{\sigma_m^2 \sum_{i=1}^{N-1} w_i} \tag{6.7}$$

where

$$w_i = \frac{1}{(t_{i+1} - t_i)^2} \tag{6.8}$$

which takes into account the stochastic sampling of our data.

medianBRP The 'median buffer range percentage' (Richards et al., 2011) is the percentage of points within the one tenth of the maximum amplitude.

$$\text{medianBRP} = \frac{|S|}{N} \tag{6.9}$$

where '$|S|$' is the number of points within '$A/10$' of the median magnitude '$\bar{m}$'. '$A/10$' is the amplitude divided by 10.

$$S = \{x \in m | \bar{m} - A/10 < x < \bar{m} + A/10\} \tag{6.10}$$

range_cum_sum The range of a cumulative sum (Ellaway, 1978). The R_{cs} should tend to 0 for symmetric distributions.

$$R_{cs} = \max S - \min S \tag{6.11}$$

where

$$s = \frac{1}{N\sigma_m} \sum_{i=1}^{N} (m_i - \bar{m}) \tag{6.12}$$

max_slope The maximum gradient between two points in the cleaned light curve.

$$\text{Max slope} = \max_{1 \leq i < N} |\frac{m_{i+1} - m_i}{t_{i+1} - t_i}| \tag{6.13}$$

MAD The median absolute deviation 'MAD' of the magnitude distribution.

$$\text{MAD} = \text{median}(|m - \text{median}(m)|) \tag{6.14}$$

mean_var The mean variance '*mean_var*' can be used as a simple indication of variability.

$$mean\_var = \frac{\sigma}{\bar{m}} \tag{6.15}$$

percent_amp The percentage amplitude '*percent_amp*' is the largest percentage difference from the median value.

$$percent\_amp = \frac{\max(m_i - median(mag))}{median(mag)} \tag{6.16}$$

roms The Robust Median Statistic '*roms*' is a metric of variability.

$$roms = \frac{\sum_{i=1}^{N} |m_i - median(m)|/m_{err,i}}{N-1} \tag{6.17}$$

ptop_var The peak-to-peak variability '$ptop\_var$' is effectively the weighted percentage amplitude.

$$ptop\_var = \frac{\max(m - m_{err}) - \min(m - m_{err})}{\max(m - m_{err}) + \min(m - m_{err})} \tag{6.18}$$

lag_auto The lag-1 autocorrelation 'lag_auto' is the dependence of the signal with itself shifted by one. It can be used to represent how similar consecutive points are.

$$lag\_auto = \frac{\sum_{i=2}^{N}(m_i - \bar{m})(m_{i-1} - \bar{m})}{\sum_{i=1}^{N}(m_i - \bar{m})^2} \tag{6.19}$$

AD The Anderson-Darling '*AD*' test is a statistical test for the similarity of a sample with a distribution (Anderson and Darling, 1952). Here, it is used to test for normality where $AD \rightarrow 0.25$ for a normal distribution.

std_nxs The normalised excess variance 'std_nxs' Vaughan et al. (2003) is variability metric commonly used in Active Galactic Nuclei variability (Gliozzi et al., 2002; Vagnetti et al., 2016; Gonzalez et al., 2023).

$$std\_nxs = \frac{\sum_{i=1}^{N}(m_i - \bar{m})^2 - m_{err}^2}{N\bar{m}^2} \tag{6.20}$$

trans_flag The transient flag '$trans\_flag$' is a boolean flag that is used to try to capture potential transients that are misidentified as periodic variable stars. The phase fold of a transient variable, such as a microlensing event, can look clean enough to be identified as periodic by both PDM and the neural network FAP. Figure 6.1 shows the raw and erroneously phase folded light curve for 'OGLE BLG-ECL-292071', which is misclassified as an eclipsing binary in Soszyński et al. (2016).

The transient flag is calculated at the time that the straight line is fitted to the cleaned light curve, figure 4.3. Each of the bins that is used for the straight line fit has the inter-quartile range (IQR) calculated. The median IQR for each of the bins is compared to each individual IQR for each bin. The transient flag is set to 1 if any bin has an IQR one third larger than the median IQR.

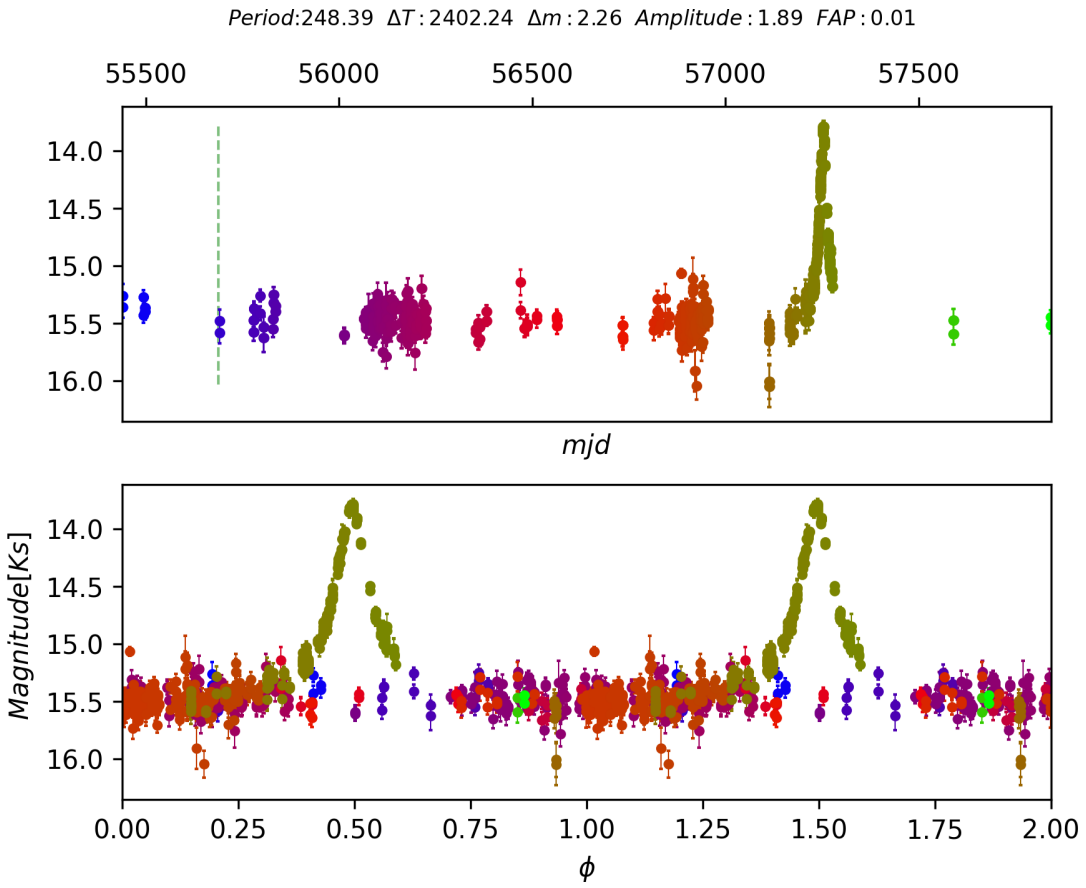


FIGURE 6.1: Top: Raw light cure showing the point of the transient event clearly at ≈ 57250 mjd. The green dashed line represents the period (i.e. $t_0 + Period$). Bottom: The incorrect phase fold at 248.39 days

$$trans\_flag = \begin{cases} 1 & \text{if IQR}_i > 1.33 \times median(IQR) \\ 0 & \text{otherwise} \end{cases} \tag{6.21}$$

ls_bal_fap The Baluev FAP (Baluev, 2008) is a false alarm probability calculated from the analysis of the Lomb-Scargle periodogram.

Periodogram Statistics Each of the periodograms is analysed thoroughly for reliable peak extraction (section 5.2) The same analysis is performed for each of the LS, PDM and CE periodograms (the only difference being LS peaks are at a maximum value whereas both PDM and CE seek to minimise their value).

Each periodogram is analysed and the three most prominent peaks are extracted. From this process we have the period '$LS/PDM/CE\_period\_0,1,2$', peak value '$LS/PDM/CE\_y\_0,1,2$' (height of the peak) and peak width '$LS/PDM/CE\_peak\_width\_0,1,2$' for each of these three

peaks. To allow for future comparison against the peak values, each periodogram has multiple percentiles calculated - $LS/CE/PDM\_0.001, 0.01, 1, 25, 50(median), 75, 99, 99.9, 99.99$

No such analysable periodogram is constructed for GPs and so we instead save all of the fitted metrics;

- 'gp_A' - amplitude factor of the covariance
- 'gp_l' - length scale of the RBF kernel
- 'gp_g' - 'Γ' the relative importance of the RBF kernel
- 'gp_P' - Period

Chapter 7

The Catalogue

The PRIMVS catalogue has 86,507,172 computed sources at the time of writing. If we take a heuristic cut of anything with a FAP less than 0.3 we have 5,161,222 periodic variable stars. The true number of periodic variables is likely to be different than that.

We can compare this to the VIVACE catalogue (Molnar et al., 2022). The VIVACE catalogue is a catalogue of periodic variables in VVV which this catalogue aims to supersede. Virtually all (97%) of the sources found in VIVACE can be found in this catalogue. Those that are not found in PRIMVS are because of different quality cuts

The University of Hertfordshire High Performance Cluster was used for the computation of each of the three tests. Each test was computed with 64 parallel instances with 4 cores and 2 GB of RAM (Totalling 256 cores and 128 GB of parallel computation use).

It is difficult to calculate a compute time for this catalogue as speed improvements and re-runs of tests create uncertainty. It takes an average of $0.9 - 1.1$ seconds per source to compute the whole pipeline inclusive of cleaning and post-processing statistics. This means it would take $\approx$ 16 days to process the catalogue with one test (ignoring cases where extra periodograms are computed).

The Bailey diagram (Bailey et al., 1919)– Logarithmic period versus amplitude–is a fundamental tool for characterising periodic variable stars. Figure 7.1 shows the Bailey diagram for all stars with FAP$<$ 0.3 in the PRIMVS catalogue.

The absence of stars at $\log_{10}(P) \approx 1.4$ is because we currently exclude periods on the diurnal and lunar time scale ($\approx$ 30 days). These will be re-added in the next version of PRIMVS but at the time of writing too many contaminants rendered this period range largely unusable. If we compare to the Bailey diagram constructed from Galactic bulge focused VIMOS data (Kains et al., 2019) we find similarly located densities. We see the same cluster of short period stars

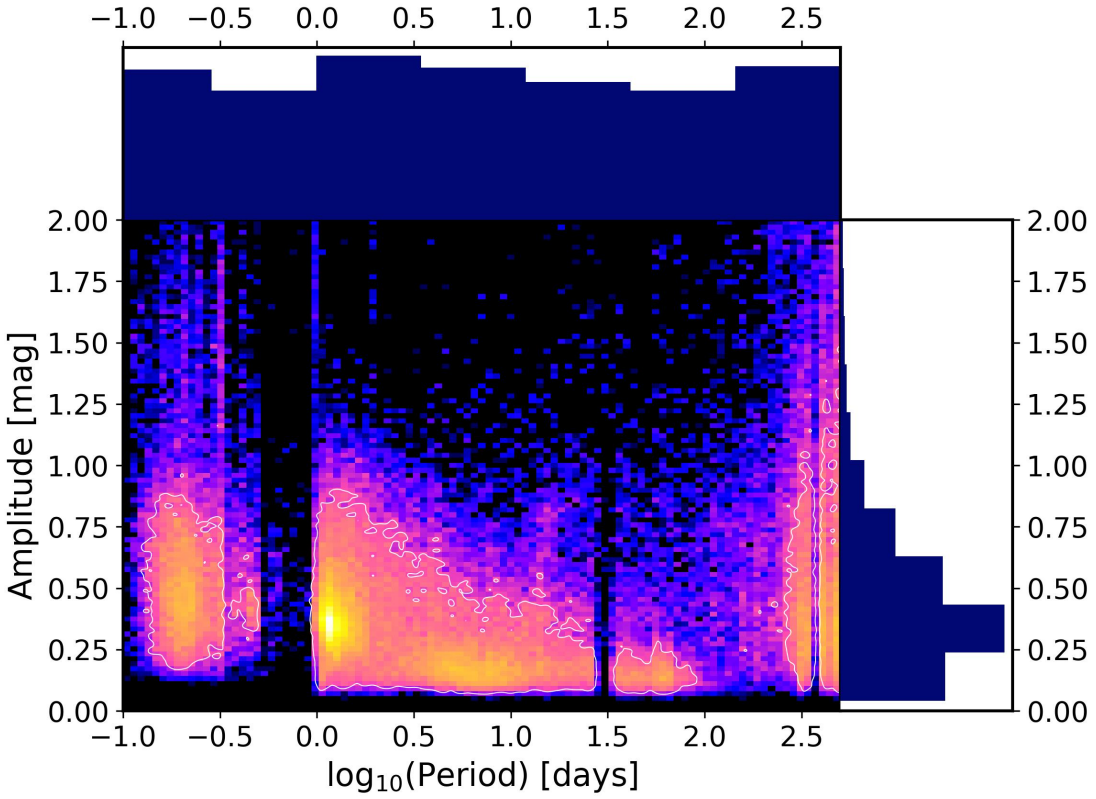


FIGURE 7.1: A plot of $\log_{10}(Period)$ verses Amplitude of all stars with an FAP<0.3. The histograms for $\log_{10}(Period)$ and Amplitude are also displayed on their respective axes. For clarity with the large sample size, a 2D histogram is used with a contour around the 80th percentile of the data. The colour axis (show of density) of the 2D histogram is in log scale.

$(-1 < \log_{10}(P) < -0.5)$ which are suspected contact binaries. We also see a density of stars where we expect to see Cepheids (da Silva et al., 2022; Kains et al., 2019; Bono et al., 2000). We do not see evidence for the typical 'double-peak' distribution caused by the Hertzsprung progression (Hertzsprung, 1926; Christy, 1975; Bono et al., 2000) (we should see a 'V' shape centred at ~ 10 days). As we have not fully classified this catalogue it is likely that the lack of this shape is due to non-Cepheids, such as EBs and RR Lyrae, filling that gap.

We measure a completeness limit of 90% for our periodic variable stars to be at a magnitude of ≈ 14.5. Figure 7.2 shows the magnitude distribution for all stars in the PRIMVS catalogue with a FAP<0.3.

We describe the VVV survey as "*an infrared time-series survey focused on the southern viewable Galactic disk and bulge*" and so it is fitting we check the key parameters of the PRIMVS catalogue against their position in Galactic coordinates.

The bottom two panels of figure 7.3 show light curve amplitude as a function of Galactic latitude and longitude. Most objects are found to have an amplitude < 1. Figure 7.4 shows the magnitude distribution in the same way as figure 7.3. A homogeneous distribution can be seen throughout

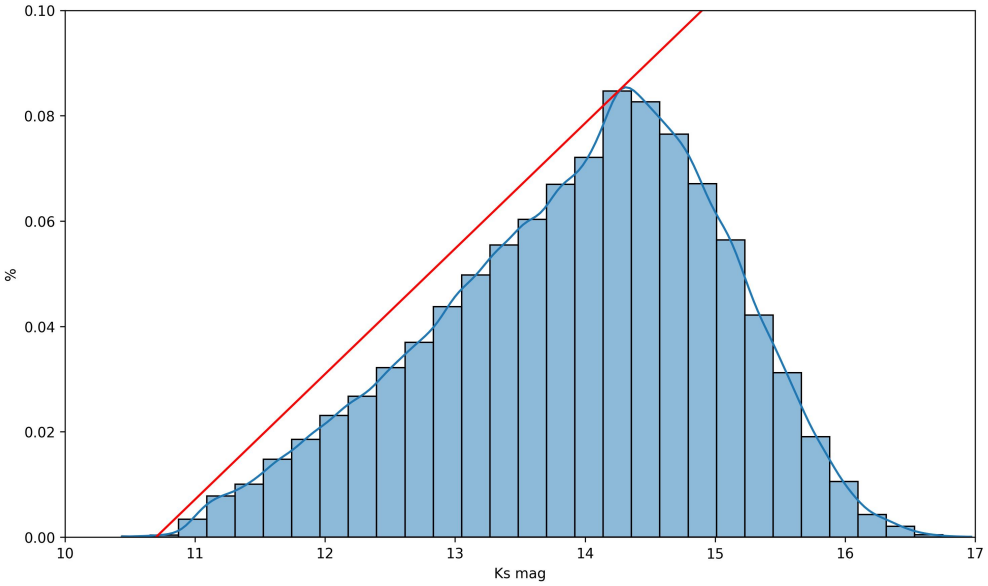


FIGURE 7.2: A histogram of magnitudes of all stars with a FAP$<$ 0.3.

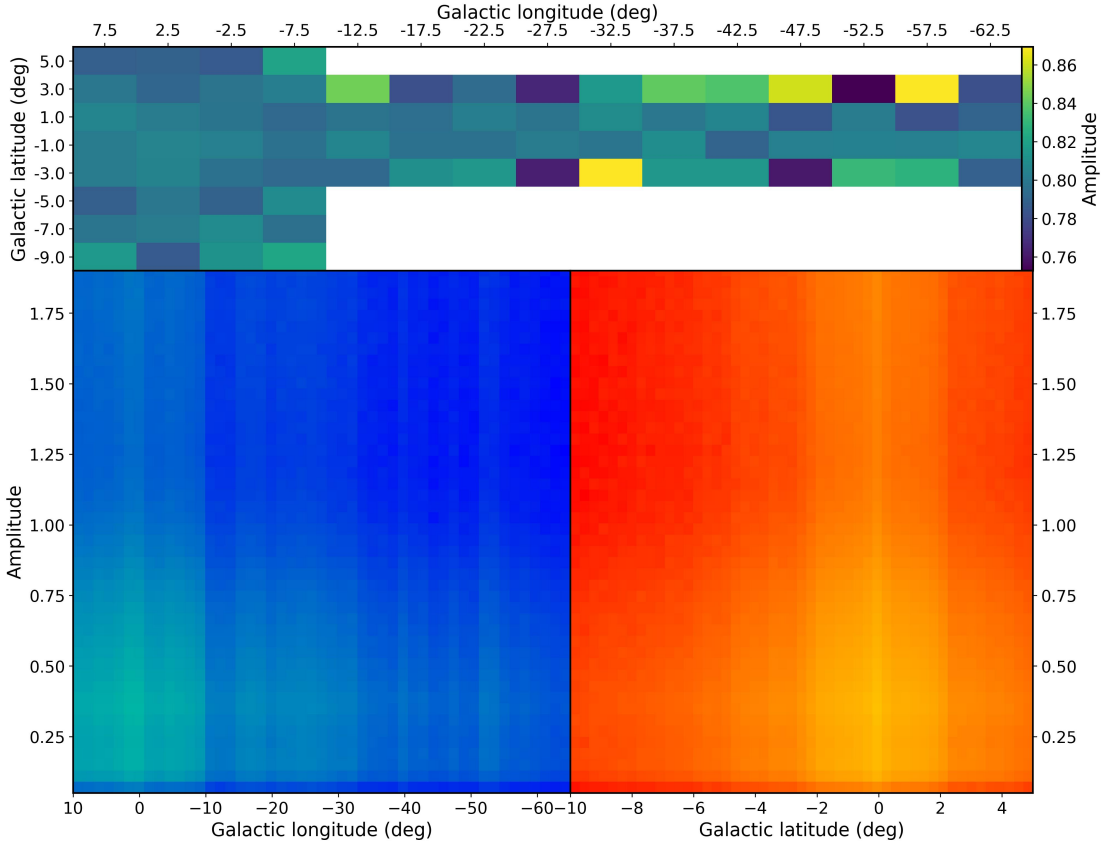


FIGURE 7.3: Light curve amplitude as a function of Galactic coordinates. Top: Histogram showing the median amplitude in each bin with respect to Galactic coordinates. Bottom Left: Density scatter plot showing amplitude as a function of Galactic longitude. Bottom Right: Density scatter plot showing amplitude as a function of Galactic latitude.

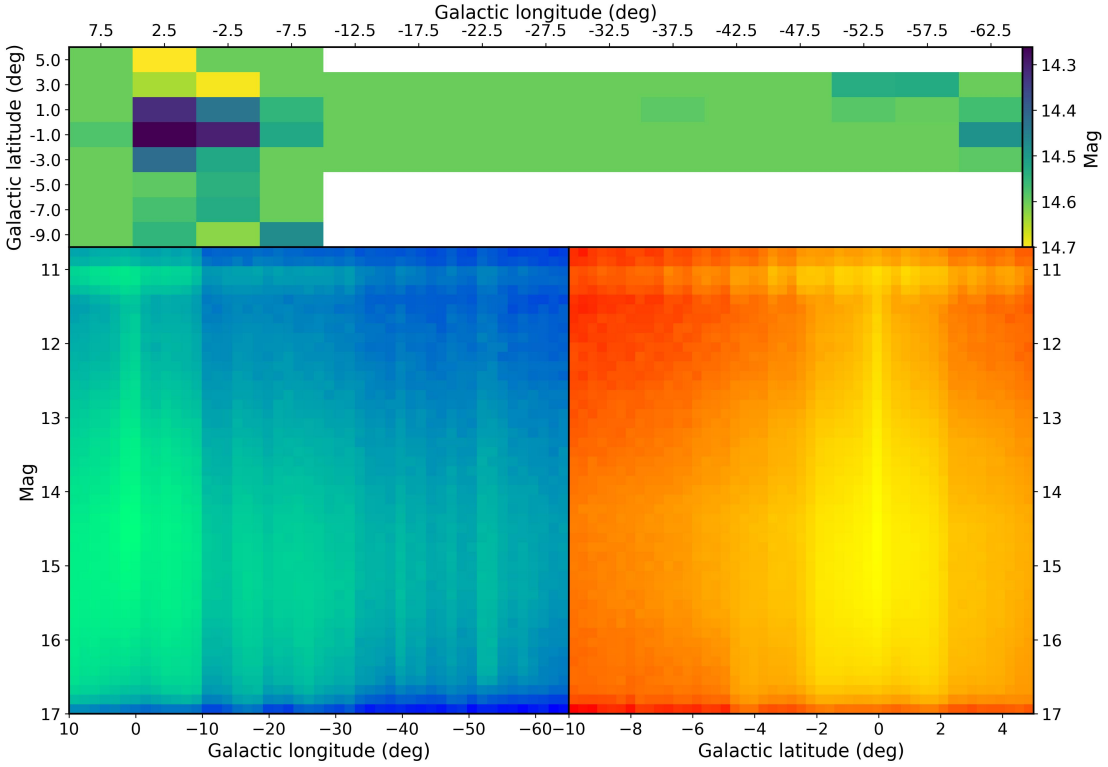


FIGURE 7.4: Light curve magnitude as a function of Galactic coordinates. Top: Histogram showing the median magnitude in each bin with respect to Galactic coordinates. Bottom Left: Density scatter plot showing magnitude as a function of Galactic longitude. Bottom Right: Density scatter plot showing magnitude as a function of Galactic latitude.

except for the Galactic bulge where the photometric depth increases by 0.2 mag. This same pattern of brighter sources can also be seen in figure 7.4. This is imparted from the array of detectors used in the VISTA telescope.

The observing pattern of the VVV survey reveals a correlation between position in Galactic coordinates and the specific detector used for measurement. This is apparent as the observing pattern for the VVV survey is based in Galactic coordinates. The VISTA telescope employs an array of 16 Raytheon VIRGO HgCdTe 0.84-2.5 micron detectors (Bornfreund, 2005). As these detectors utilise relatively early-stage technology, they can exhibit differences in sensitivity, linearity, and particularly in saturation limits. This variability across the detectors affects the precision and reliability of measurements, as some detectors reach saturation at lower brightness levels than others. Such discrepancies can lead to areas of the VVV survey region which are probed to greater depth and/or brighter magnitude.

We can look for other features as a function of location that help us to begin to verify the completeness of PRIMVS. Due to the nature of period finding techniques, light curves with uneven magnitude distributions/non-sinusoidal shapes are often underrepresented in periodic variable catalogues. Eclipsing binaries are ubiquitous and largely homogeneous throughout the galaxy (Mowlavi et al., 2023). Due to the nature of an eclipsing binary light curve, they are

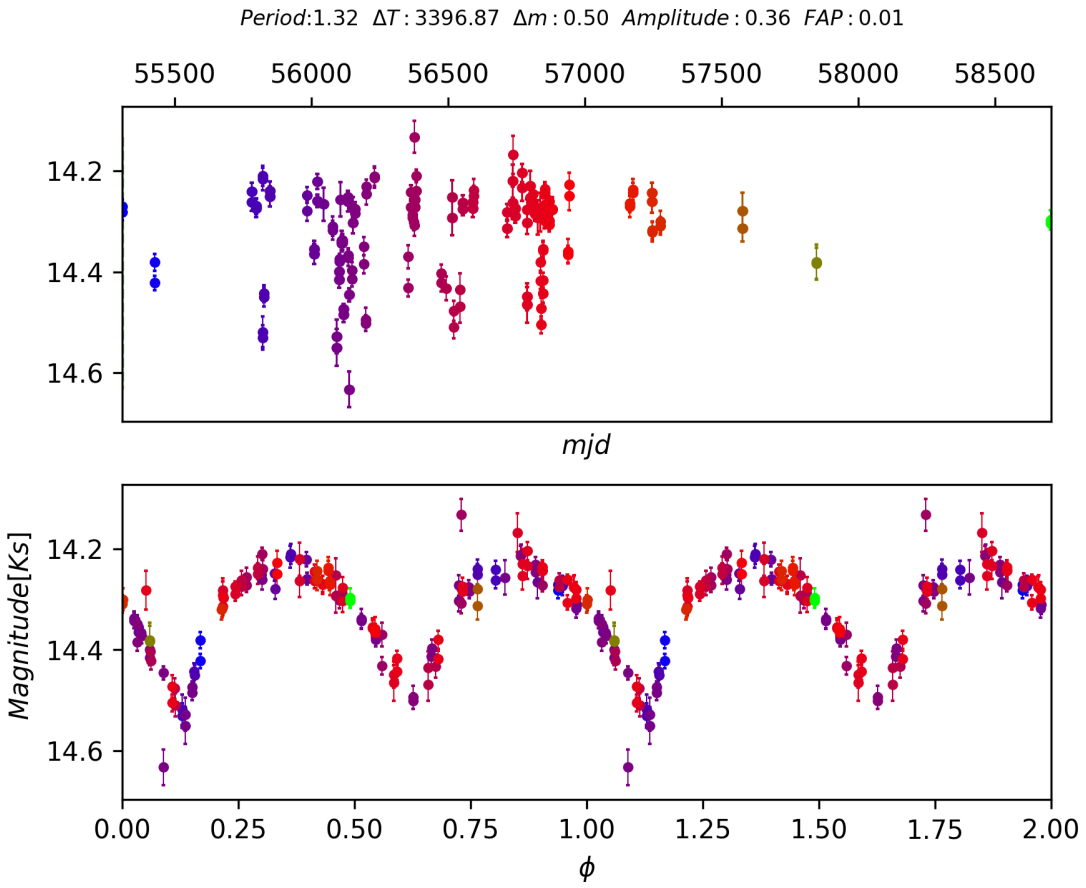


FIGURE 7.5: Light curve of eclipsing binary 'V* V2679 Sgr'. Top: Light curve as a function of time. Bottom: phase folded light curve. The colours are correlated with time.

largely unique in their light curve morphology and resulting magnitude distribution. This is exemplified by figure 7.5 where a typical EB (β − lyrae) light curve can be seen.

For an EB, a large distribution of the points are in the brighter stages of the light curve (either the lack of an eclipse or a relatively minor reflection) with fewer points tracing out the two eclipses. This results in a uneven distribution of points, unlike how a Cepheid, RR-Lyra or AGB light curve would have[1]. EBs are therefore likely to be the largest contributor to any measured skew deviating from 0 in the catalogue. Figure 7.6 shows the measured skew in PRIMVS as a function of Galactic coordinates. The median skew is greater than 0.2 across the whole of the PRIMVS catalogue but also appears largely homogeneous throughout the Galactic disk and bulge. This helps to indicate that we have not preferentially selected EBs in either the Galactic disk or bulge, regions with different densities of stars.

[1] assuming no other perturbation

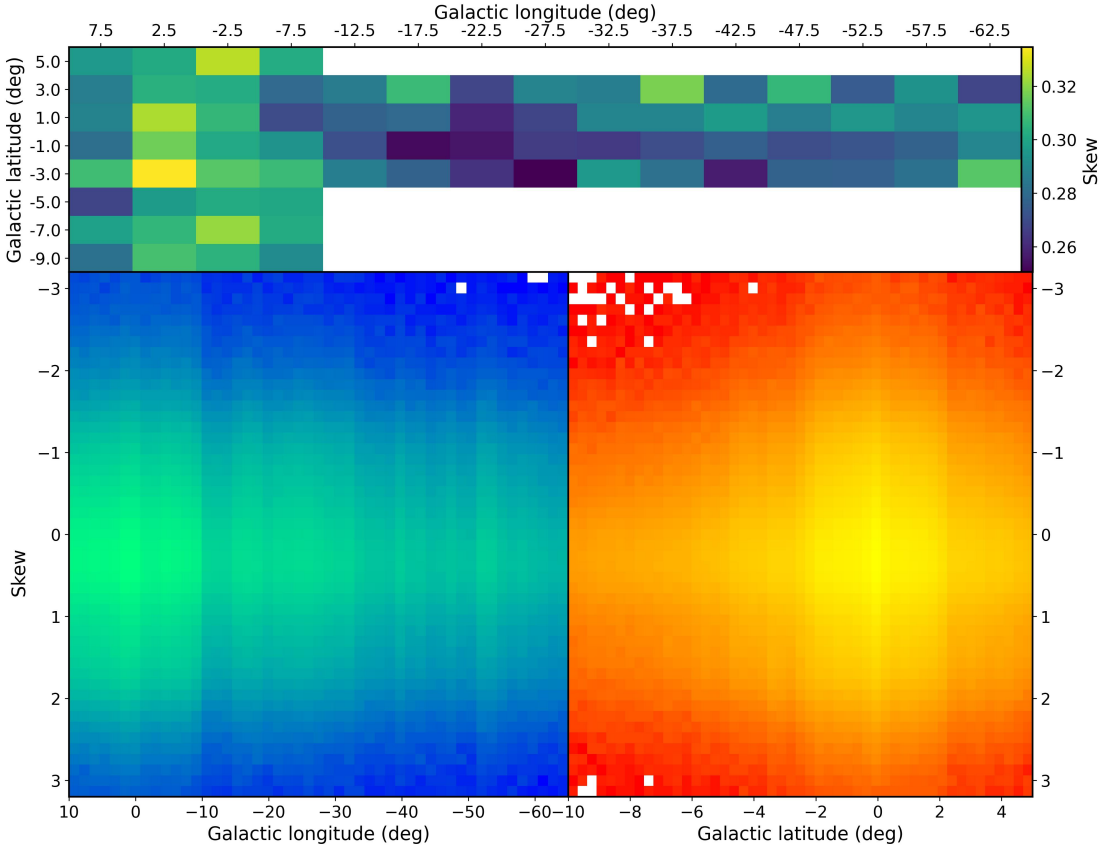


FIGURE 7.6: Light curve skew as a function of Galactic coordinates. Top: Histogram showing the median skew in each bin with respect to Galactic coordinates. Bottom Left: Density scatter plot showing skew as a function of Galactic longitude. Bottom Right: Density scatter plot showing skew as a function of Galactic latitude.

7.1 Quasi-periodic sources

The variability of a source can be from a multitude of non-mutually exclusive intrinsic and extrinsic reasons. It follows that there are many variable sources which feature some apparent quasi-periodicity. This quasi-periodicity could be from the combination of a periodic variability with some other aperiodic variability (as is commonly seen in YSOs) or single causes of quasi-periodicity (such as star spots). The treatment of quasi-periodicity was considered in the construction of this catalogue by the combination of feature that are outlined in section 6. The neural network FAP acts as something analogous to a measurement of structure in the phase folded light curve (similar to PDM). There exists cases where quasi-periodic sources feature strong structure within their phase fold. This will lead to the neural network based FAP to erroneously proscribe a FAP indicative of periodicity. However, other features that are calculated for the light curve will indicate a deviation from periodicity. Most notably is the neural network FAP and the Baluev FAP will likely disagree as Baluev FAP has a dependency on the similarity to a sinusoidal wave. It is likely that quasi-periodic sources live in a latent space described by each of these features that is exclusively different from the periodic sources. A future work for

this project will be to identify this latent space region and assign a flag of quasi-periodicity to any sources that are within it.

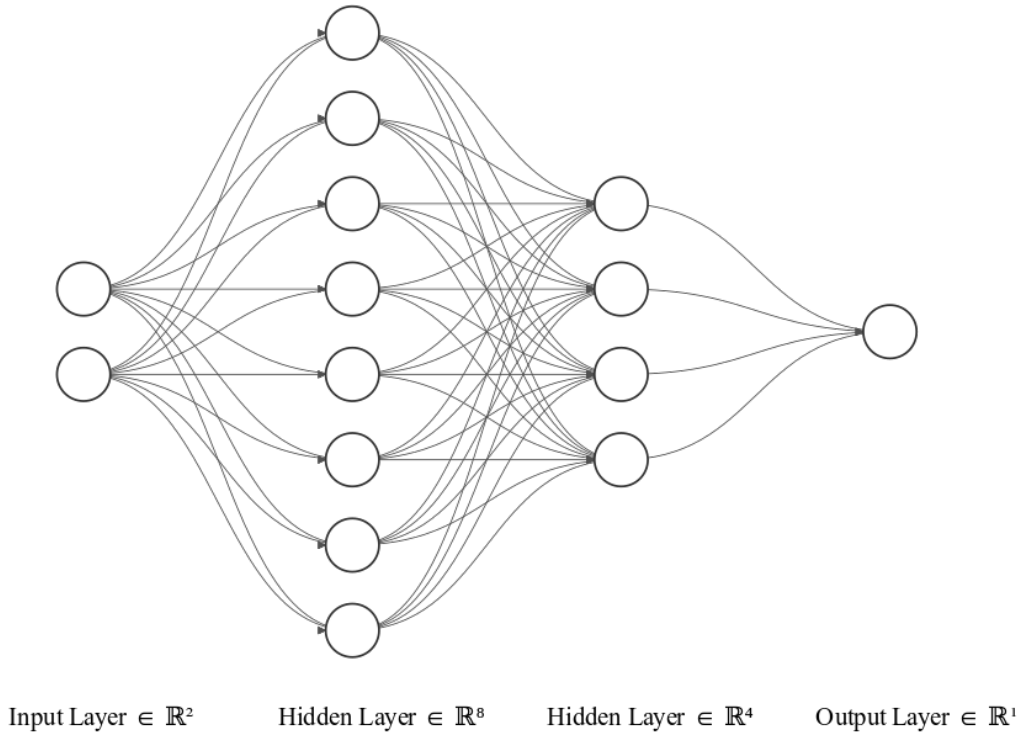


FIGURE 7.7: Showing architecture of the autoencoder that was used. For easy visualisation, each node here represents 16 actual nodes in the network.

7.2 PRIMVS Embedding

To continue with the overarching goal of unbiased exploration of the VVV data set we can make latent space representations of the PRIMVS catalogue to highlight potential groups. Figure 7.7 shows the architecture of the autoencoder (which takes the form of an MLP) that was used for this process.

The network was trained for 469 epochs with early stopping on the plateau of validation loss. The initial learn rate of 0.1 was halved each time the validation loss did not reduce for more than 10 steps. Only features with discernible physical meaning were used (i.e. features that an astronomer would use to begin to classify a star). Table 7.1 shows each of the features that were used. Features with a '*' are taken from the VIRAC catalogue. A future improvement for any methods using this selection of features would be to identify important and useless features. There is a lot of shared information between many of the features as a large portion of them are describing the magnitude distribution of the light curve.

Figure 7.8 shows the Uniform Manifold Approximation and Projection (UMAP; McInnes et al.,

*z med mag-ks med mag	Median z band magnitude - median ks mag
*y med mag-ks med mag	Median y band magnitude - median ks mag
*j med mag-ks med mag	Median j band magnitude - median ks mag
*h med mag-ks med mag	Median h band magnitude - median ks mag
*l	galactic longitude
*b	galactic latitude
Cody M	AM Cody 'M' value
stet k	Stetson 'K' value
eta e	Von Neumann 'eta e' vallue
med BRP	Median buffer range percentage
range cum sum	Range of a cumulative sum
max slope	Maximum slope between two points
MAD	Median Absolute Deviation
mean var	Mean Variance
percent amp	Percentage Amplitude
true amplitude	Amplitude
roms	RObust Median Statistic
p to p var	Peak-to-peak variability
lag auto	Lag-1 autocorrelation
AD	Anderson-Darling
std nxs	Normalized excess variance
weight mean	Weighted Mean
weight std	Weighted Standard deviation
weight skew	Weighted Skew
weight kurt	Weighted Kurtosis
mean	Mean
std	Standard deviation
skew	Skew
kurt	Kurtosis
true period	Period

TABLE 7.1: Table showing each of the features used in the embedding process. Features with a '*' are taken from the VIRAC catalogue.

2018) of these features. UMAP is a machine learning technique used for dimensionality reduction. It is particularly effective at preserving both the local and global structure of the data, making it useful for visualisation of high-dimensional datasets, similar to t-SNE (van der Maaten and Hinton, 2008). UMAP works by constructing a high-dimensional graph representing the data, then optimising a low-dimensional graph to be as structurally similar as possible, hence reducing dimensions while retaining the data's original structure.

Figure 7.9 shows the Principle Component Analysis (PCA) projection of these features. PCA is a statistical technique used for dimensionality reduction while preserving as much of the variance in the high-dimensional data as possible. It works by identifying the directions (principal components) along which the variance in the data is maximised. The data is $\mathbb{Z}$-normalised and the covariance matrix is across each feature is calculated. Eigenvectors of the covariance matrix are computed and sorted with respect to the magnitude of their eigenvalues, this forms the

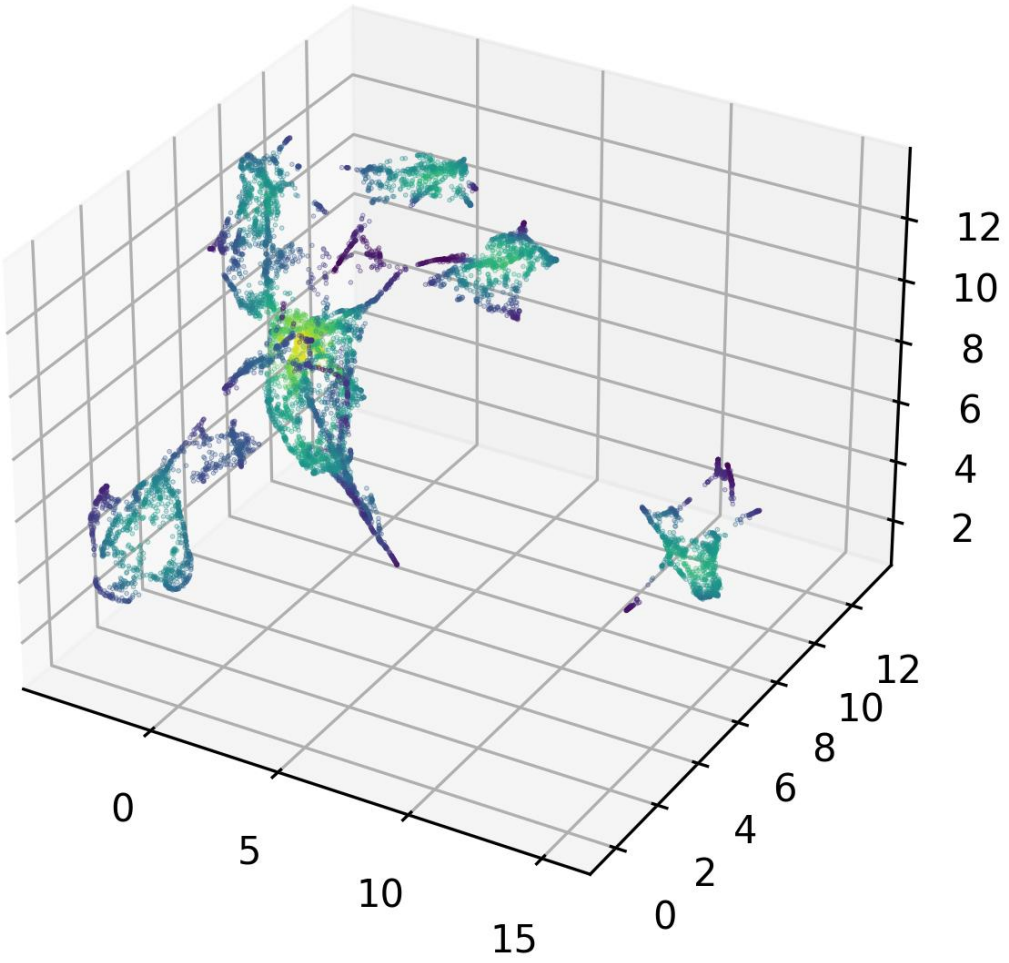


FIGURE 7.8: A 3 dimensional UMAP representation of the features from table 7.1

principle components. The original data is projected onto the principal components selected in the previous step, resulting in a new dataset with reduced dimensions. Figures 7.8 & 7.9 show 20,000 points with the colour representing density.

Both of these projections show similar features, most notably the smaller isolated group that can be seen in both. Comparing both projections shows that the same set of stars are found in both the PCA and UMAP isolated groups. These groups comprise $\approx 2\%$ of the total distribution in both projections. Comparing these groups to the rest of the distribution we find features which can parameterise it: amplitude > 1, Lag-1 autocorrelation < 0.2, $0.2 <$ FAP < 0.3 and, period $>$ 1000 days. Figure 7.10 shows the raw and phase folded light curves of 16 sources selected from this isolated group. These objects appear to be high amplitude, quasi-periodic variable stars. This is expected as the above parameterisation appropriately descibes such objects. It is also expected that the objects found in such an isolated group would not be as cleanly periodic.

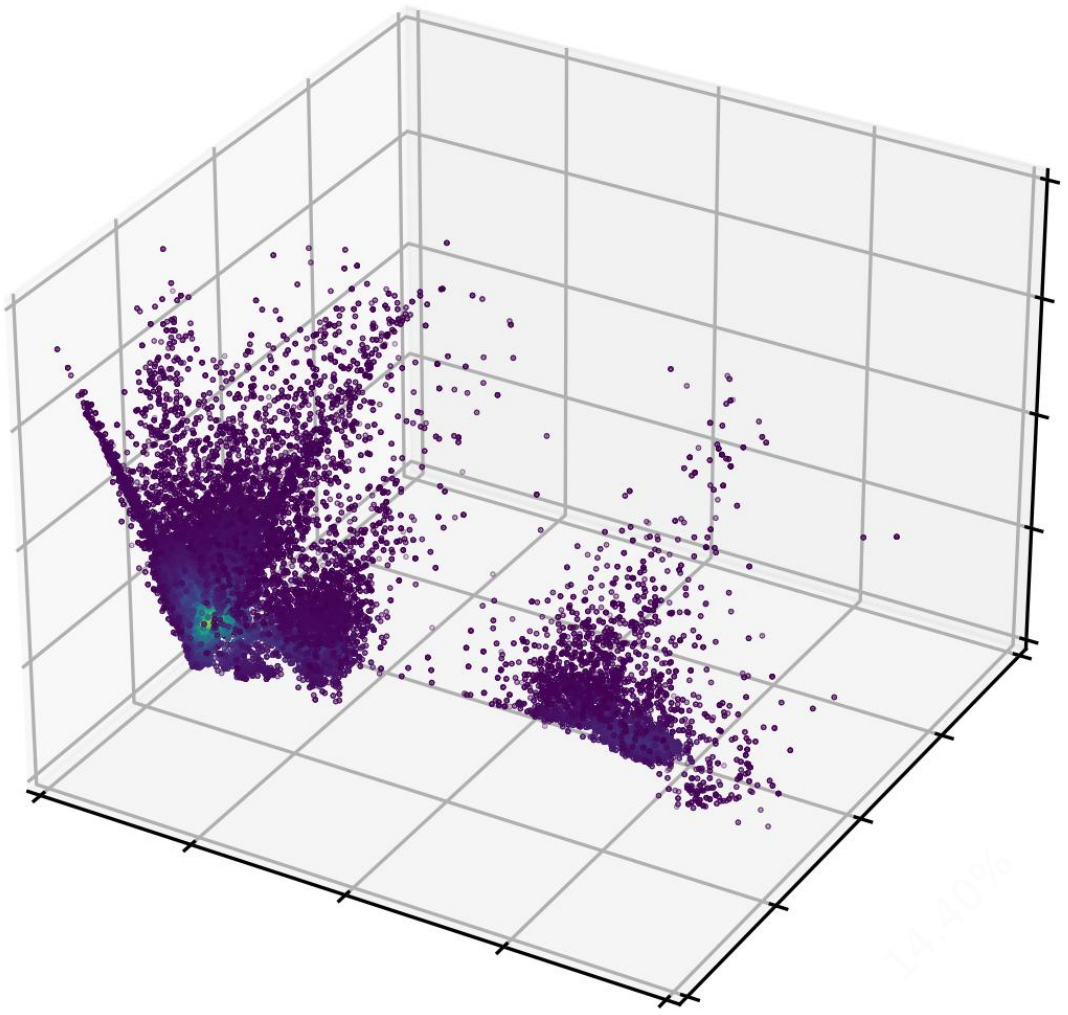
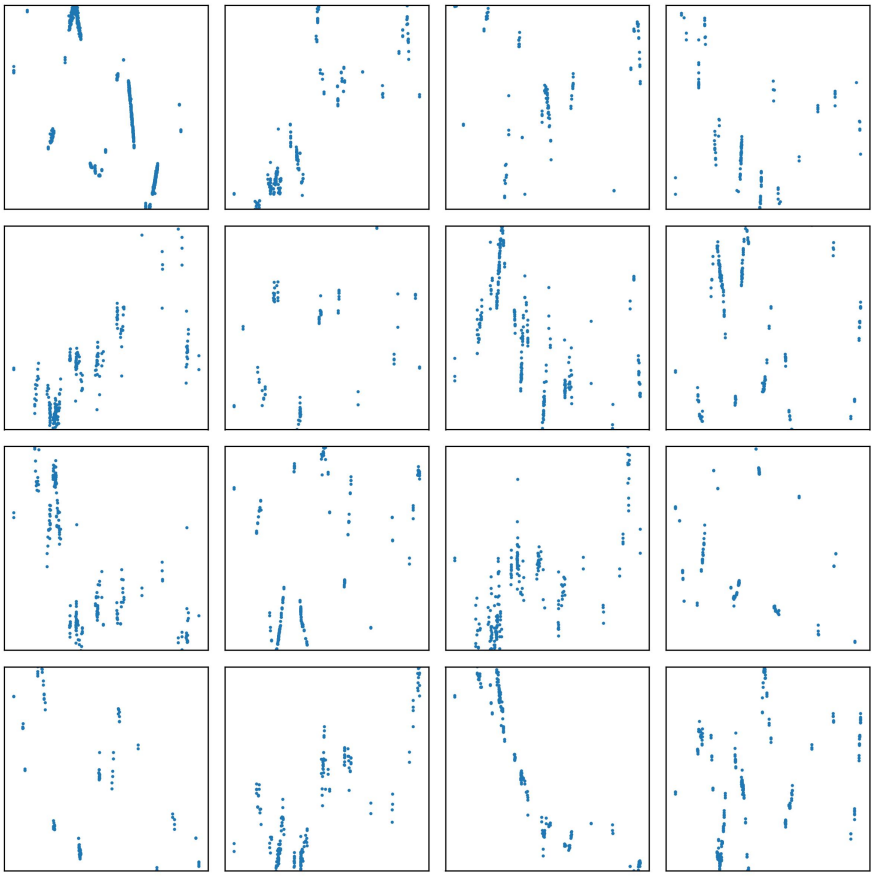


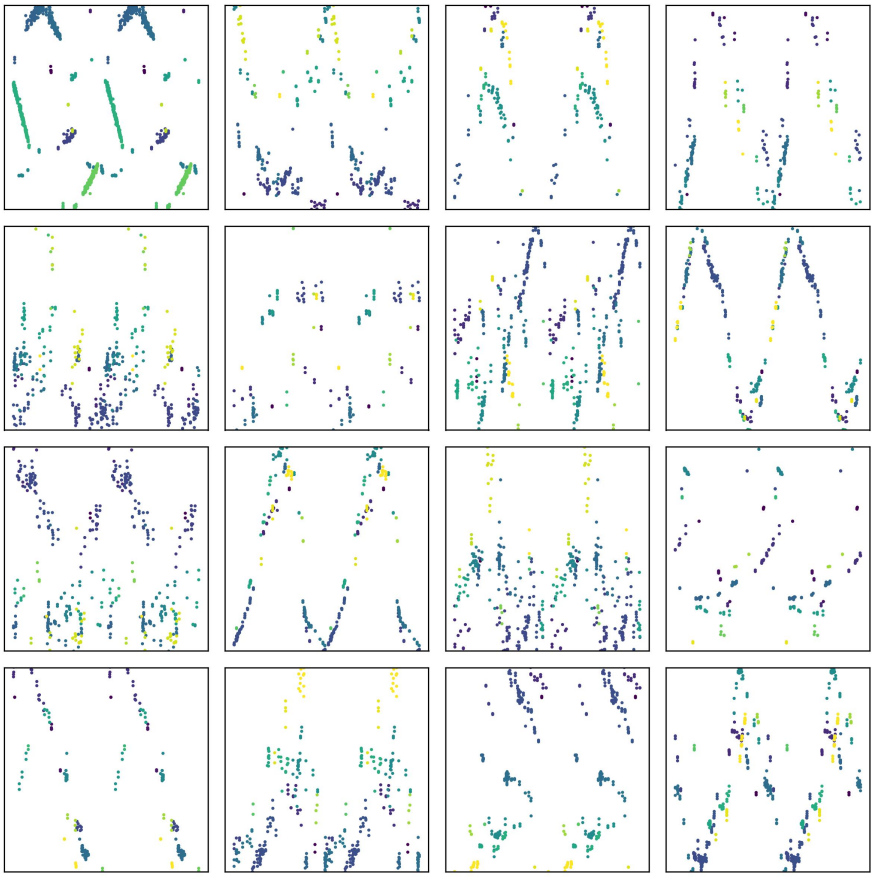
FIGURE 7.9: A 3 dimensional PCA representation of 31 features from table 7.1. The importance of the dimensions is; 32.58%, 14.40%, 13.68%

A Quasi-periodic (or Aperiodic) object would possess features more distinctly different than individual classes of periodic variable stars.

Interactive plots and data visualisations that are not suitable for the pdf format can be found at: https://nialljmiller.com/projects/PRIMVS/PRIMVS.html.



(A) mjd vs mag plot



(B) Phase vs mag plot (colour as mjd)

FIGURE 7.10: Light curves of 16 objects identified from the isolated group seen in figures 7.8 & 7.9

7.3 Decision Trees

The use of a pre-classified star catalogue as a training set for a machine learning algorithm is not new. The General Catalogue of Variable Stars (GCVS, (Samus' et al., 2017)) exemplifies such a catalogue, having compiled variable objects since 1946. This approach has been applied to data from VVV (Molnar et al., 2022), ASAS-SN (Jayasinghe et al., 2018, 2019), Gaia DR2 (Rimoldini et al., 2019) and DR3 (Rimoldini et al., 2023), EROS-II light curves (Kim et al., 2014), and in the identification of microlensing events (Husseiniova et al., 2021).

It follows that we can use this approach to classify the variable stars in the PRIMVS catalogue. This method is not without caveats however, most notably the biases inherited from the training set.

The cross match of PRIMVS with the Gaia DR3 all-sky classification catalogue (Rimoldini et al., 2023) yielded 118,172 sources with an FAP < 0.3 and a 'best class score' > 0.7. This selection forms the training set which was then used with a gradient boosted decision tree classifier 'XGBoost' (Chen and Guestrin, 2016). Figure 7.11 shows the distribution of classes found in the cross-matched data. This distribution is not even across all classes and reflects the selection bias of both the Gaia DR3 all-sky classification catalogue and the VVV PRIMVS catalogue. This is a notable caveat, especially as the differences in the data (e.g. optical vs near-IR) likely leads to different selection biases.

Figure 7.12 shows the confusion matrix achieved from using the Gaia DR3 all-sky classification catalogue to form our training set for classifying PRIMVS. It can be seen that the majority of classes are correctly identified with a high completeness. All of the White Dwarf and RCB variables are misclassified as EBs and LPVs respectively. RCB variables are hydrogen-poor, carbon/helium-rich, high-luminosity stars. Their variability is characterised by high amplitude (1-9mag) aperiodic changes on the order of hundreds of days. This is superimposed by periodic pulsations up to several tenths of a magnitude on the time scale of tens of days (Clayton, 1996). The light curves of RCBs are therefore complex and likely span a large range in any feature space. Considering this with their scarcity, it is not surprising we misclassify all of them as LPVs, a much more common class with similar features. Given this, RCBs and White Dwarfs were removed from the data.

We can calculate how confident the model is with the highest probable class. Figure 7.13 shows the 'Entropy' versus the 'Confidence metric' for each class with a probability> 0.5 for its most likely class. where 'Entropy' is the entropy across the classes, (i.e. $S = \sum P_{class} \ln(P_{class})$). Therefore, a lower Entropy suggests that the model's predictions are more certain because the probability distribution across classes is less uniform – One class has a much higher probability compared to others, indicating a strong preference by the model for that class. The 'Confidence metric' is the difference between the most likely and next most likely class.

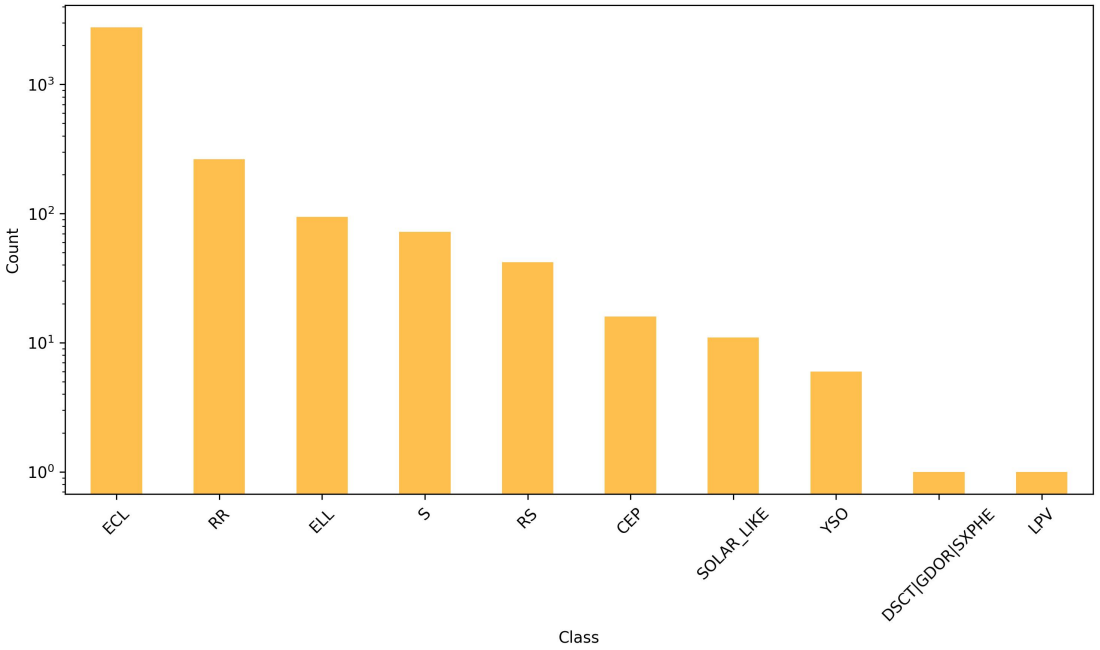


FIGURE 7.11: Training set of cross-matched VVV-Gaia data. Where; 'ECL' is eclipsing binaries, 'RR' is RR Lyraes, 'ELL' is Ellipsoidals, 'S' is short time-scale objects, 'RS' is RS Canum Venaticorum variables, 'CEP', is Cepheids, 'SOLAR_LIKE' is for Solar-like objects, 'YSO' is young stellar objects, 'DSCT||GDOR||SXPHE' is for delta-scuti like objects, and 'LPV' is for long period variables.

A Bailey diagram is an excellent tool for versifying how sensible our classifications are. Figure 7.14 shows the Bailey diagram constructed from the highest probability sources for each class. For both figure 7.14 and figure 7.15 we select only sources with a probability > 0.7, entropy < 0.2 and confidence metric > 0.9. The same colours and markers are also used to represent each class throughout figures 7.13,7.14, and 7.15 The Cepheid population is clearly visible and takes the expected form on the plot. The expected bimodal distribution of Cepheids can be seen as a loose 'V' shape at 10 days (Bono et al., 2000). We also see good agreement with Kains et al. (2019) in terms of our LPV, Delta Scuti and RR Lyrae placements.

Figure 7.15 shows the stellar classifications across the VVV survey region in relation to their positions within the Milky Way. This plot provides insights into the typical locations of different stellar populations. Cepheids, marked as red dots, are young, luminous stars commonly found in the thin disk throughout the galaxy Skowron et al. (2019). Figure 7.15 shows our sample of Cepheids throughout the disk mostly within $|l| < 1.5$, with an increased density at $|b| < 6$. Long-period Variables, such as Miras and semi-regulars are typically older, evolved stars and thus are more prevalent in the Galactic bulge and halo, where older stellar populations dominate (Wood and Bessell, 1983). Figure 7.15 shows these objects homogeneously spaced throughout the disk with significantly higher densities towards the inner bulge. RR Lyrae stars, yellow plus signs, are old, metal-poor stars found mainly in the Galactic bulge and halo, highlighting regions with

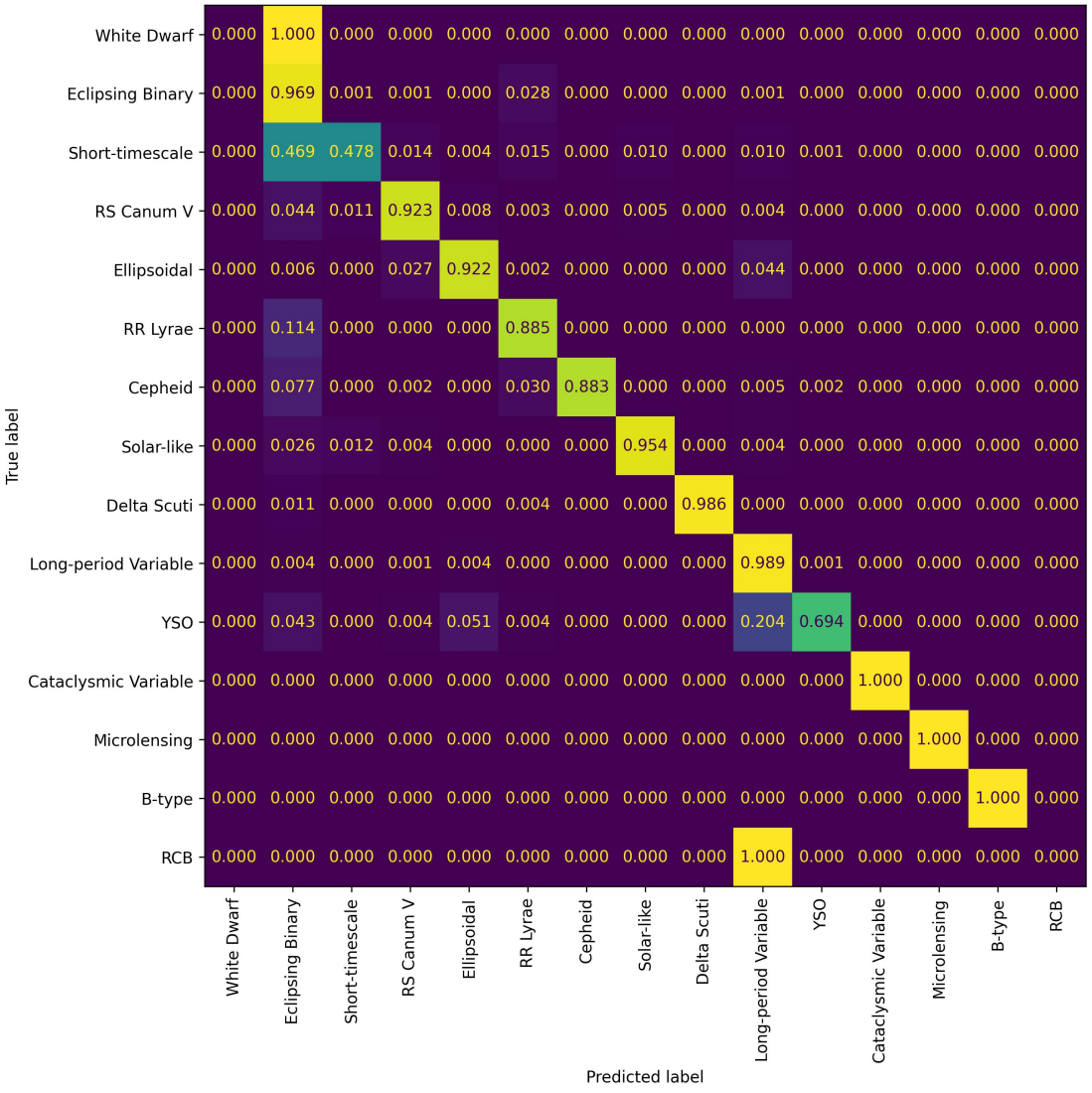


FIGURE 7.12: Confusion matrix of grouped classes from the Gaia Data Release 3 All-sky classification (Rimoldini et al., 2023). 'RCB' is R Coronae Borealis variables and YSO is Young Stellar Objects.

ancient star populations ,(Cabrera Garcia et al., 2023; Ramos et al., 2018). This seems to be in agreement with figure 7.15.

Similar to the autoencoder, interactive plots and data visualisations that are not suitable for the pdf format can be found at: https://nialljmiller.com/projects/PRIMVS/PRIMVS.html.

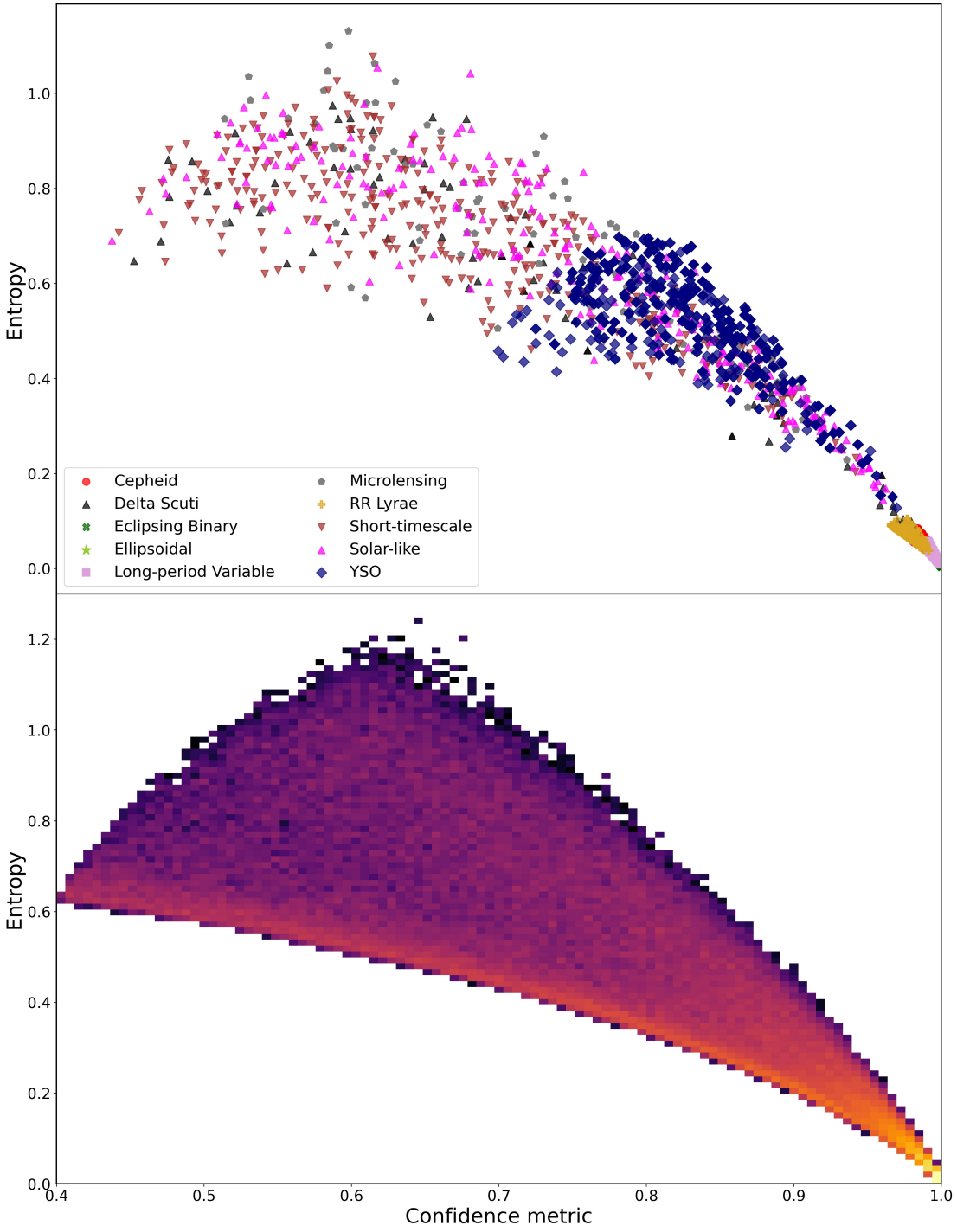


FIGURE 7.13: Top: A scatter plot of 'Entropy' versus the 'Confidence metric' for each each class with a probability>0.5. Bottom: A 2D histogram of 'Entropy' versus the 'Confidence metric' for the same data

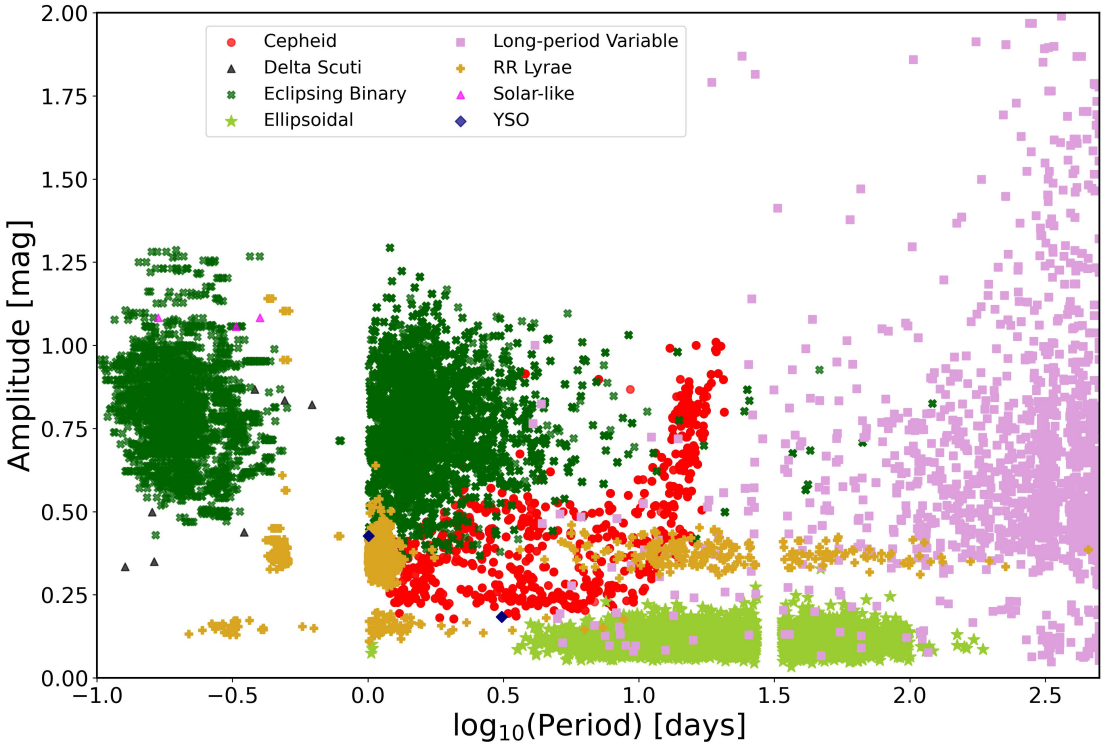


FIGURE 7.14: A plot of $\log_{10}(Period)$ verses Amplitude for the most confident predictions (top 10%) of each class from our decision tree which was trained using the Gaia DR3 all-sky classification catalogue.

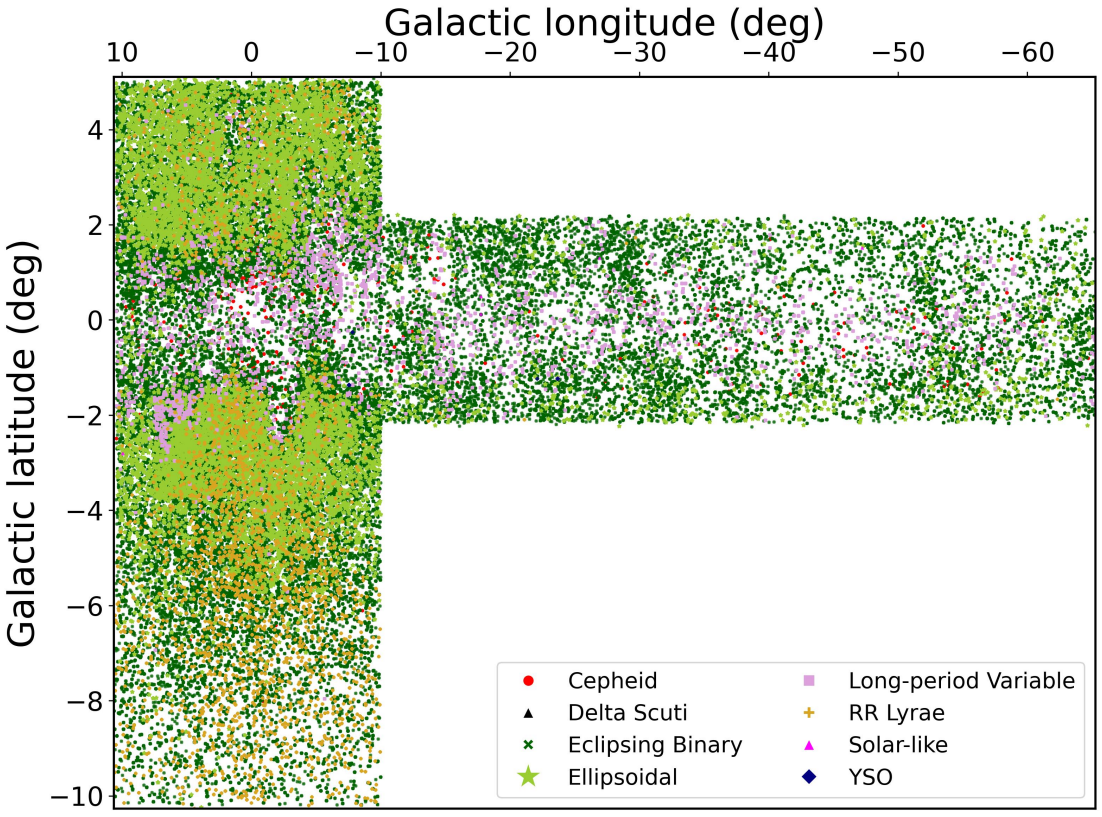


FIGURE 7.15: Spatial distribution of stellar classes across the VVV survey region in the context of the Milky Way. The decision tree based classification uses the Gaia DR3 all-sky classification catalogue as its training set.The absence of YSOs here is due to the conservative cuts shown in figure 7.13

Chapter 8

Conclusions

This work introduces the PeRiodic Infrared Milky-way VVV Star-catalogue (PRIMVS), leveraging the VVV survey's depth and breadth to investigate the variability of astronomical sources within the Milky Way's Galactic bulge and disk. Through meticulous data cleaning and preprocessing, alongside modern analysis techniques, PRIMVS highlights the efforts towards an unbiased and complete identification and classification of periodic variable stars.

Our analysis employed various period-finding methods, demonstrating their strengths and weaknesses, and utilised a novel FAP method to enhance reliability in period identification. The catalogue includes over 86 million candidate variable sources and $\approx$ 5 million periodic variable stars.

Machine learning techniques, notably decision trees, have been shown as viable in classifying a substantial portion of PRIMVS sources. Cross-matched data from Gaia DR3 and the Simbad database has proven effective at identifying known and expected classes of stars. This approach, however, introduces its own set of challenges, notably the potential biases from the training sets and the limitation posed by Gaia's optical depth compared to the near-IR capabilities of VVV.

PRIMVS not only advances our understanding of variable stars within the Milky Way but also showcases the potential of combining traditional astronomical analysis with modern data science techniques to explore and categorise astronomical sources effectively. Future work will aim to refine these classifications, expand the catalogue's scope, and further integrate deep learning approaches for a more thorough understanding of the stellar demographics and population of the Milky Way.

Bibliography

Anderson, T.W. and Darling, D.A., 1952. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 .

Angus, R., Morton, T., Aigrain, S., et al., 2018. Inferring probabilistic stellar rotation periods using Gaussian processes. *MNRAS*, 474(2):2094.

Astropy Collaboration, Robitaille, T.P., Tollerud, E.J., et al., 2013. Astropy: A community Python package for astronomy. *A&A*, 558:A33.

Bailey, S.I., Leland, E.F., Woods, I.E., et al., 1919. Variable stars in the cluster Messier 15. *Annals of Harvard College Observatory*, 78:195.

Baluev, R.V., 2008. Assessing the statistical significance of periodogram peaks. *MNRAS*, 385(3):1279.

Bono, G., Marconi, M., and Stellingwerf, R.F., 2000. Classical Cepheid pulsation models — VI. The Hertzsprung progression. *A&A*, 360:245.

Bornfreund, R.E., 2005. Large format short-wave hgcdte detectors and focal plane arrays. *Raytheon Vision Systems*.

Cabrera Garcia, J., Beers, T.C., Huang, Y., et al., 2023. Probing the Galactic halo with RR Lyrae stars – V. Chemistry, kinematics, and dynamically tagged groups. *Monthly Notices of the Royal Astronomical Society*, 527(3):8973.

Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *arXiv e-prints*, arXiv:1603.02754.

Chen, T., Kornblith, S., Norouzi, M., et al., 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv e-prints*, arXiv:2002.05709.

Christy, C.T., Jayasinghe, T., Stanek, K.Z., et al., 2023. The ASAS-SN catalogue of variable stars X: discovery of 116 000 new variable stars using G-band photometry. *MNRAS*, 519(4):5271.

Christy, R.F., 1975. The Hertzsprung progression in Cepheid calculations. In *NASA Special Publication*, volume 383, pages 85–98.

Clayton, G.C., 1996. The R Coronae Borealis Stars. *PASP*, 108:225.

Cody, A.M., Stauffer, J., Baglin, A., et al., 2014. CSI 2264: Simultaneous Optical and Infrared Light Curves of Young Disk-bearing Stars in NGC 2264 with CoRoT and Spitzer—Evidence for Multiple Origins of Variability. *AJ*, 147(4):82.

da Silva, R., Crestani, J., Bono, G., et al., 2022. A new and Homogeneous metallicity scale for Galactic classical Cepheids. II. Abundance of iron and α elements. *A&A*, 661:A104.

Ellaway, P., 1978. Cumulative sum technique and its application to the analysis of peristimulus time histograms. *Electroencephalography and Clinical Neurophysiology*, 45(2):302.

Gliozzi, M., Brinkmann, W., Räth, C., et al., 2002. On the nature of X-ray variability in Ark 564. *A&A*, 391:875.

Gonzalez, A.G., Gallo, L.C., Miller, J.M., et al., 2023. Characterizing X-ray, UV, and optical variability in NGC6814 using high-cadence Swift observations from a 2022 monitoring campaign. *Monthly Notices of the Royal Astronomical Society*, 527(3):5569.

Graham, M.J., Drake, A.J., Djorgovski, S.G., et al., 2013. Using conditional entropy to identify periodicity. *MNRAS*, 434(3):2629.

Hertzsprung, E., 1926. On the relation between period and form of the light-curve of variable stars of the δ Cephei type. , 3:115.

Höfner, S. and Olofsson, H., 2018. Mass loss of stars on the asymptotic giant branch. Mechanisms, models and measurements. , 26(1):1.

Hogg, D.W., 2008. Data analysis recipes: Choosing the binning for a histogram. *arXiv e-prints*, arXiv:0807.4820.

Husseiniova, A., McGill, P., Smith, L.C., et al., 2021. A microlensing search of 700 million VVV light curves. *MNRAS*, 506(2):2482.

Jayasinghe, T., Kochanek, C.S., Stanek, K.Z., et al., 2018. The ASAS-SN catalogue of variable stars I: The Serendipitous Survey. *MNRAS*, 477(3):3145.

Jayasinghe, T., Stanek, K.Z., Kochanek, C.S., et al., 2019. The ASAS-SN catalogue of variable stars - II. Uniform classification of 412 000 known variables. *MNRAS*, 486(2):1907.

Kains, N., Calamida, A., Rejkuba, M., et al., 2019. New variable stars towards the Galactic Bulge - I. The bright regime. *MNRAS*, 482(3):3058.

Kim, D.W., Protopapas, P., Bailer-Jones, C.A.L., et al., 2014. The EPOCH Project. I. Periodic variable stars in the EROS-2 LMC database. *A&A*, 566:A43.

McInnes, L., Healy, J., and Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, arXiv:1802.03426.

Molnar, T.A., Sanders, J.L., Smith, L.C., et al., 2022. Variable star classification across the Galactic bulge and disc with the VISTA Variables in the Vía Láctea survey. *MNRAS*, 509(2):2566.

Mowlavi, N., Holl, B., Lecoeur-Taïbi, I., et al., 2023. Gaia Data Release 3. The first Gaia catalogue of eclipsing-binary candidates. *A&A*, 674:A16.

Neumann, J.V., 1941. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *The Annals of Mathematical Statistics*, 12(4):367.

Ramos, R.C., Minniti, D., Gran, F., et al., 2018. The vvv survey rr lyrae population in the galactic center region*. *The Astrophysical Journal*, 863(1):79.

Rasmussen, C.E. and Williams, C.K.I., 2006. *Gaussian Processes for Machine Learning*.

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. *The Astrophysical Journal*, 733(1):10.

Rimoldini, L., Holl, B., Audard, M., et al., 2019. Gaia Data Release 2. All-sky classification of high-amplitude pulsating stars. *A&A*, 625:A97.

Rimoldini, L., Holl, B., Gavras, P., et al., 2023. Gaia Data Release 3. All-sky classification of 12.4 million variable sources into 25 classes. *A&A*, 674:A14.

Samus', N.N., Kazarovets, E.V., Durlevich, O.V., et al., 2017. General catalogue of variable stars: Version GCVS 5.1. *Astronomy Reports*, 61(1):80.

Skowron, D.M., Skowron, J., Mróz, P., et al., 2019. A three-dimensional map of the Milky Way using classical Cepheid variable stars. *Science*, 365(6452):478.

Smith, L.C., Lucas, P.W., Kurtev, R., et al., 2018. VIRAC: the VVV Infrared Astrometric Catalogue. *MNRAS*, 474(2):1826.

Soszyński, I., Pawlak, M., Pietrukowicz, P., et al., 2016. The OGLE Collection of Variable Stars. Over 450 000 Eclipsing and Ellipsoidal Binary Systems Toward the Galactic Bulge. , 66(4):405.

Stetson, P.B., 1996. On the Automatic Determination of Light-Curve Parameters for Cepheid Variables. *PASP*, 108:851.

Vagnetti, F., Middei, R., Antonucci, M., et al., 2016. Ensemble X-ray variability of active galactic nuclei. II. Excess variance and updated structure function. *A&A*, 593:A55.

van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579.

VanderPlas, J.T., 2018. Understanding the Lomb-Scargle Periodogram. *ApJ*, 236(1):16.

Vaughan, S., Edelson, R., Warwick, R.S., et al., 2003. On characterizing the variability properties of X-ray light curves from active galaxies. *MNRAS*, 345(4):1271.

Wood, P.R. and Bessell, M.S., 1983. Long-period variables in the galactic bulge : evidence for a young super-metal-rich population. *ApJ*, 265:748.

Zechmeister, M. and Kürster, M., 2009. The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms. *A&A*, 496(2):577.