

The ZTF Source Classification Project – II. Periodicity and variability processing metrics

Michael W. Coughlin¹,[★] Kevin Burdge², Dmitry A. Duev³, Michael L. Katz,^{4,5} Jan van Roestel,² Andrew Drake,² Matthew J. Graham³, Lynne Hillenbrand,² Ashish A. Mahabal², Frank J. Masci,⁶ Przemek Mróz,² Thomas A. Prince², Yuhan Yao,² Eric C. Bellm,⁷ Rick Burruss,⁸ Richard Dekany,⁸ Amruta Jaodand,² David L. Kaplan⁹, Thomas Kupfer,¹⁰ Russ R. Laher,⁶ Reed Riddle⁸, Mickael Rigault,¹¹ Hector Rodriguez,⁸ Ben Rusholme⁶ and Jeffry Zolkower⁸

¹*School of Physics and Astronomy, University of Minnesota, Minneapolis, MN 55455, USA*

²*Division of Physics, Math, and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*

³*Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*

⁴*Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA*

⁵*Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA), Evanston, IL 60208, USA*

⁶*IPAC, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125, USA*

⁷*DIRAC Institute, Department of Astronomy, University of Washington, 3910 15th Avenue NE, Seattle, WA 98195, USA*

⁸*Caltech Optical Observatories, California Institute of Technology, Pasadena, CA 91125, USA*

⁹*Department of Physics, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA*

¹⁰*Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106, USA*

¹¹*Université Clermont Auvergne, CNRS/IN2P3, Laboratoire de Physique de Clermont, F-63000 Clermont-Ferrand, France*

Accepted 2020 September 15. Received 2020 September 15; in original form 2020 July 16

ABSTRACT

The current generation of all-sky surveys is rapidly expanding our ability to study variable and transient sources. These surveys, with a variety of sensitivities, cadences, and fields of view, probe many ranges of time-scale and magnitude. Data from the Zwicky Transient Facility (ZTF) yields an opportunity to find variables on time-scales from minutes to months. In this paper, we present the codebase, *ztfperiodic*, and the computational metrics employed for the catalogue based on ZTF’s Second Data Release. We describe the publicly available, graphical-process-unit optimized period-finding algorithms employed, and highlight the benefit of existing and future graphical-process-unit clusters. We show how generating metrics as input to catalogues of this scale is possible for future ZTF data releases. Further work will be needed for future data from the Vera C. Rubin Observatory’s Legacy Survey of Space and Time.

Key words: methods: data analysis – techniques: photometric – catalogues – surveys – stars: statistics.

1 INTRODUCTION

The study of variable and transient sources is rapidly expanding based on the large data sets available from wide-field survey telescopes. Amongst others these include the *Panoramic Survey Telescope* and Rapid Response System (Pan-STARRS; Morgan et al. 2012), the Asteroid Terrestrial-impact Last Alert System (ATLAS; Tonry et al. 2018), the Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009, 2014), the All-Sky Automated Survey for SuperNovae (ASAS-SN; Shappee et al. 2014; Kochanek et al. 2017), the Zwicky Transient Facility (ZTF; Bellm et al. 2018; Masci et al. 2018; Graham et al. 2019; Dekany et al. 2020), the Visible and Infrared Survey Telescope for Astronomy (VISTA; Lopes et al. 2020), and in the near future, the Vera C. Rubin Observatory’s Legacy Survey of Space and Time (LSST; Ivezić et al. 2019). Catalogues of variable sources from these surveys are building upon the highly successful work

identifying periodic variable stars in the Magellanic Clouds and Galactic center; we highlight catalogues such as those from MACHO (Alcock et al. 2000) and OGLE (Udalski 2003; Udalski, Szymański & Szymański 2015) as examples. Coupled with Gaia’s second data release (DR2; Gaia Collaboration 2018) and soon DR3, we are currently experiencing a dramatic expansion of our capabilities for doing time-domain astrophysics. The range of sensitivities and cadences change the parameter space of magnitude ranges and time-scales probed, which can span from years down to minutes for many of these surveys.

While these surveys can be used to follow-up objects of interest, here, we are primarily interested in ‘untargeted’ searches, i.e. searches without *a priori* knowledge about objects, for variable and periodic objects. These searches enable identification of populations of objects, such as cataclysmic variables (Szkody et al. 2020), Be stars (Ngeow et al. 2019), Cepheids and RR Lyrae, useful for studying the processes and properties of galaxy formation (Saha 1984, 1985; Catelan 2009), and measuring the expansion rate of the Universe (Freedman et al. 2001). In addition, they are also sensitive

[★] E-mail: cough052@umn.edu

to rare objects, such as short-period white dwarf binaries (Burdge et al. 2019a,b; Coughlin et al. 2020), large-amplitude radial-mode hot subdwarf pulsators (Kupfer et al. 2019), and ultracompact hot subdwarf binaries (Kupfer et al. 2020).

To highlight one source class, we are discovering and characterizing the population of so-called ultra-compact binaries (UCBs) which have two stellar-mass compact objects with orbital periods $P_{\text{orb}} < 1$ h (Burdge et al. 2019a,b; Coughlin et al. 2020). Many of the UCBs emit gravitational waves in the milliHertz regime with sufficient strain for the upcoming *Laser Interferometer Space Antenna* (*LISA*) to detect (Amaro-Seoane et al. 2017). These *LISA* ‘verification sources’ will serve as crucial steady-state sources of gravitational waves that will not only verify that *LISA* is operating as expected (Kupfer et al. 2018), but also themselves serve as probes of binary stellar evolution (Nelemans & Tout 2005; Antonini, Toonen & Hamers 2017; Banerjee 2018; Kremer et al. 2018), white dwarf structure (Fuller & Lai 2011), Galactic structure (Breivik, Mingarelli & Larson 2020), accretion physics (Cannizzo & Nelemans 2015), and general relativity (Burdge et al. 2019a; Kupfer et al. 2019).

We are interested in computationally efficient algorithms for measuring the variability and periodicity of large suites of light curves for the purpose of creating catalogues of sources. These catalogues are useful for identifying classes of sources such as those mentioned above, as well as for mitigating the presence of known variable sources in searches for new transient objects, such as short γ -ray burst or gravitational-wave counterparts, e.g. Coughlin et al. (2019a, b). The size of all-sky survey data are large, enabling identification of a large number of sources. However, the computational cost scales linearly with the number of objects. In ZTF’s Second Data Release (DR2),¹ which we will employ here, there are 3153 256 663 light curves, regardless of passband; for comparison, a cross-match to a Pan-STARRS catalogue indicates that ~ 1.3 billion of these sources are unique. In the following, each individual light curve will be processed if they exceed 50 detections after surviving data quality and other cuts described below; we will focus our discussion in this paper on a small fraction of the fields, covering a range of stellar densities, consistent with those chosen for classification studies in the accompanying catalogue paper (van Roestel et al. 2021). We note that this separately treats g -, r -, and i -band observations of the same light curves, and the set includes many objects with hundreds of detections in multiple passbands. In the following, we analyse single-band light curves, and so the same objects in different passbands are analysed separately. We seek to choose variability metrics and periodicity algorithms appropriate for analysing this data set. We have taken inspiration from variability and periodicity codebases, such as FATS (Nun et al. 2015), *astrobases* (Bhatti et al. 2020), and *cesium* (Naul et al. 2016) in choosing the metrics, with a focus on using scalable graphical processing unit (GPU) periodicity algorithms to make the required large-scale processing tractable.

In this paper, we will describe the pipeline `ztfperiodic`,² which we use to systematically identify variable and periodic objects in ZTF; we use a mix of metrics designed to efficiently identify and characterize variable objects as well as algorithms to phase-fold light curves in all available passbands. We use a variety of available arrays of GPUs to period search all available photometry. The arrays of GPUs available are ideal for period finding large data sets and already have shown significant speed-ups relative to central processing units (CPUs).

2 OBSERVATIONAL DATA

We employ data predominantly from DR2 in this analysis that covers public data between 2018-03-17 and 2019-06-30, and private data between 2018-03-17 to 2018-06-30; in our analysis, we also include private data up until 2019-06-30. ZTF’s Mid-Scale Innovations Program in Astronomical Sciences (MSIP) program covers the observable night sky from Palomar with a three-night cadence, and a nightly cadence in both g and r filters in the Galactic plane (Bellm et al. 2019). The ZTF partnership is also conducting a moderate cadence survey over a 3000 square degree field at high declination, visiting the entire field six times per night in g and r filters, resulting in more than 1000 epochs. Within this field, we expect to probe variables at the 20 mmag level for objects at 16th magnitude, and around 100 mmag at 20th magnitude (Masci et al. 2018). Also, ZTF conducts a high cadence survey in the Galactic plane, with galactic latitude $|b| < 14^\circ$ and galactic longitude $10^\circ < l < 230^\circ$, where the camera continuously sits on a field for several hours with a 40 s cadence. Overlaid by stellar densities, we show the ZTF primary field grid in Fig. 1; we note there is also a secondary grid filling in gaps in the primary grid. In this figure, we also include outlined in red the fields targeted in the forthcoming catalogue paper (van Roestel et al. 2021).

The left-hand panel of Fig. 2 shows the probability density function of the number of detections for individual objects passing the data quality and time cuts for four example fields. The locations of these fields are given in Table 1, chosen to span a variety of galactic latitudes and longitudes and therefore stellar densities. This analysis removes any observation epochs indicating suspect photometry according to the *catflags* value in the light curve metadata.³ It also removes any ‘high cadence’ observations, defined as observations within 30 min of one another; for any series of observations with observation times less than 30 min apart, we keep only the first observation in the series. This is to support the periodicity analyses; an over-abundance of observations in very densely observed sets otherwise dominate the period finding statistics, as these observations will have the same weight as all temporally separated observations, but lack the sensitivity to periods longer than the single night in which they are taken. This short-coming will be a point of study moving forward. In addition to those observations removed, some subset of sources have fewer observations within the same field; these occur because faint stars have fewer detections because of lower detection efficiency. We also remove any stars within 13 arcsec of either an entry in the Yale Bright Star Catalog (Hoffleit & Jaschek 1991) and Gaia (Gaia Collaboration 2018) stars brighter than 13th magnitude; we arrived at this value through evaluating the measured variability of stars near cataloged objects, and 13 arcsec was a threshold beyond which the variability was consistent with background. This helps to remove bright blends from the variability selection process. On the right of Fig. 2, we show the probability density function of the weighted standard deviation for the light curves of individual objects in the example fields, showing consistency of this metric across fields.

To query the photometry, we are using ‘Kowalski’,⁴ an efficient non-relational database that uses MongoDB to efficiently store and access both ZTF alert/light curve data and external catalogues (including 230M+ alerts and 3.1B+ light curves) (Duev et al. 2019). We show the read time from the ZTF light curve database ‘Kowalski’ versus the number of objects returned on the left of Fig. 3. We

¹<https://www.ztf.caltech.edu/page/dr2>

²<https://github.com/mcoughlin/ztfperiodic>

³For details, see the DR2 Release Summary <https://www.ztf.caltech.edu/page/dr2>, specifically Section 9b.

⁴<https://github.com/dmitryduev/kowalski>

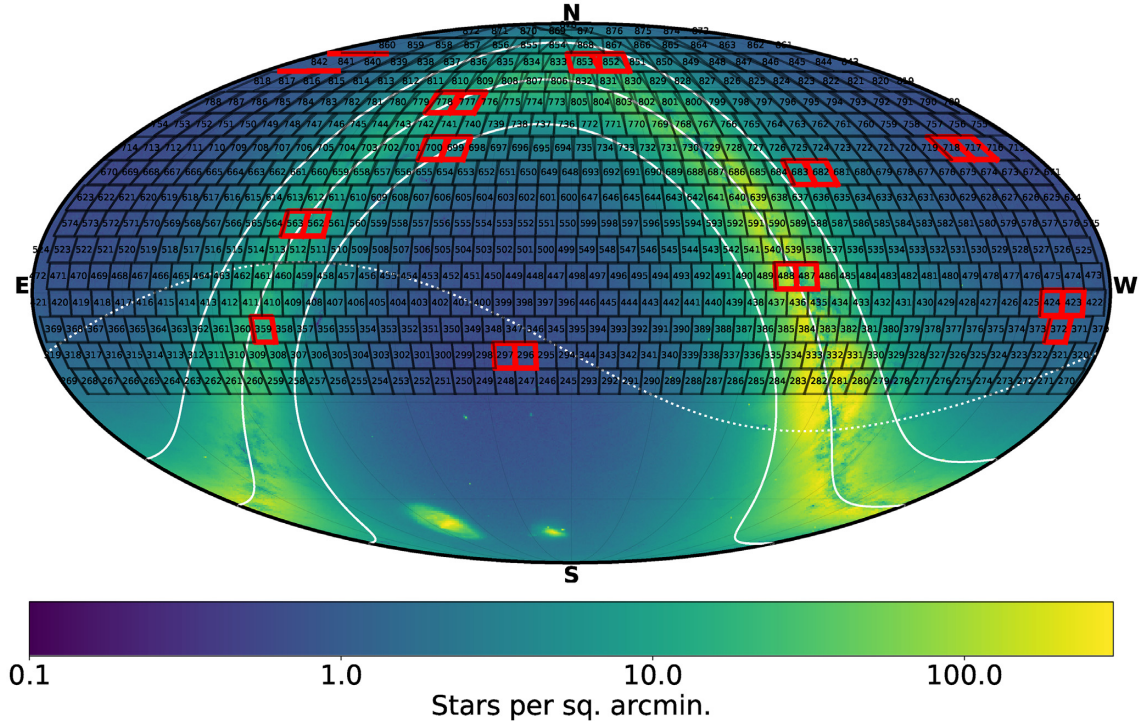


Figure 1. Mollweide projection in equatorial coordinates of the ZTF field coverage as a function of stellar density, using Gaia DR2 (Gaia Collaboration 2018). The black lines correspond to ZTF field boundaries, with the number contained corresponding to the field number. The solid white lines correspond to galactic latitude of -15 , 0 , and 15 , while the dashed line corresponds to the ecliptic. We highlight the 20 fields in the initial catalogue release in red (van Roestel et al. 2021).

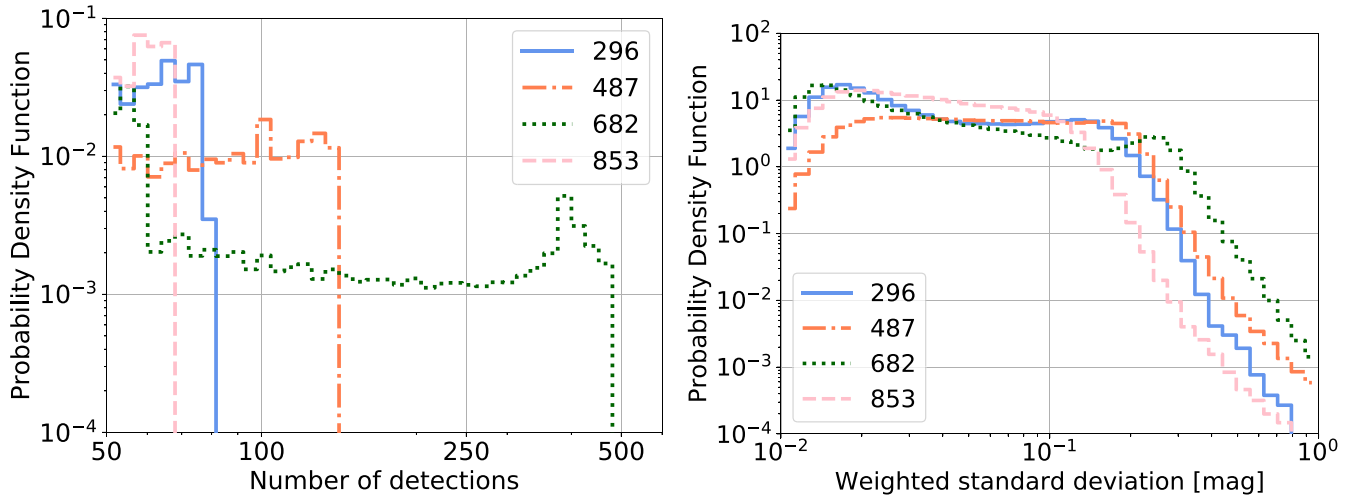


Figure 2. Left: Probability density function of the number of detections for individual objects passing the data quality and time cuts for four example fields. The field IDs are given in the legend, and their locations given in Table 1. Right: Probability density function of the weighted standard deviations for the light curves of individual objects.

Table 1. Locations of the fields (measured at their centres) highlighted in this paper.

Field	RA ($^{\circ}$)	Dec ($^{\circ}$)	Gal. Long. ($^{\circ}$)	Gal. Lat. ($^{\circ}$)
296	15.79	-17.05	141.27	-79.14
487	281.19	4.55	36.55	3.03
682	266.86	33.35	58.61	26.73
851	351.43	69.35	115.73	7.93

note that the typical 100–200 ms includes the typical 54 ms latency between California (where Kowalski resides) and Minnesota (where the test was performed), including ~ 20 ms of light traveltime. As can be seen, the analysis takes ~ 1 s per light curve across the range of light curves analysed at a single time (100–1000 which is the memory limit for the GPUs).

The light curves are analysed in groups of ~ 1000 to fill out the RAM available on most GPUs employed; when the jobs are being

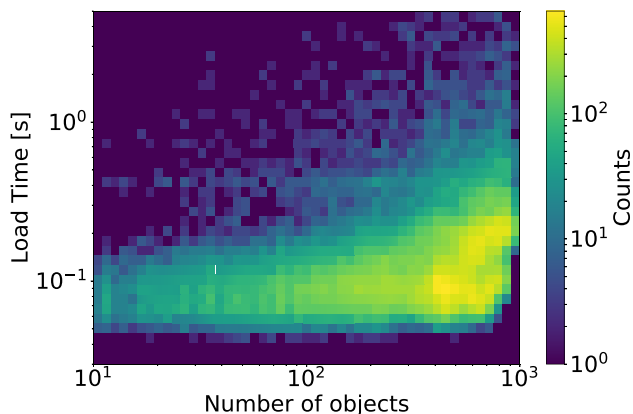


Figure 3. Two-dimensional histogram of the read time from the ZTF light curve database ‘Kowalski’ versus the number of objects returned. The load time is mostly independent of the number of objects returned.

distributed, the total number of light curves in a given quadrant on a particular CCD within a field are queried. This is possible as ZTF has two grids on which all observations are performed, with a repeating pointing accuracy varying ~ 50 – 200 arcsec; therefore, a particular object will generally only appear twice, once on each grid. For reference, there are 64 quadrants, 4 quadrants for each of the 16 CCDs. Based on this, the number of jobs required to analyse the total number of light curves, in chunks of 1000, are computed. To ensure the efficacy of the variability and periodicity metrics we discuss below, we place a threshold of at least 50 detections in a single band for an object to be analysed. This requirement means fewer than 1000 light curves are sometimes analysed because some will not meet the 50 observation limit required for analysis. Because `ztfperiodic` analyses the data in chunks like these, it is simple to parallelize across GPUs, with different GPUs running on different sets of ~ 1000 light curves; an hdf5 file is written at the end to disc containing the statistics for a particular job (see below), each chunk receiving a different name based on a simple convention (field, CCD, quadrant, job index).

3 VARIABLE SOURCE METRICS

Due to the significant amount of astronomical time series data provided by ZTF and other all-sky surveys, it is important to have robust selection criteria and algorithms to find variable objects. In ground-based surveys like ZTF, light curves are irregularly sampled, have gaps, and can have large statistical errors, which means that these algorithms must be robust in order to efficiently find true signals. The catalogue includes two main types of metrics: variability and periodicity. These metrics are typically useful for machine learning algorithms to group light curves into categories through feature extraction based on light curve data. The goal of these features is to encode numerical or categorical properties to characterize and distinguish the different variability classes, such that machine learning algorithms can distinguish between classes of light curves.

There are a variety of computationally cheap variability metrics we use, ranging from relatively basic statistical properties such as the mean or the standard deviation to more complex metrics such as Stetson indices (Stetson 1996). As shown by Pashchenko, Sokolovsky & Gavras (2017), many of the commonly used features are strongly correlated. We therefore chose the set of features suggested by Pashchenko et al. (2017), with the addition of robust measurements

of the amplitude. We also performed a Fourier decomposition to characterize the shape of the folded light curve better. We detail the choices we have made below, which include the number of measurements, weighted mean and median magnitudes (RMS and percentile-based), kurtosis, skewness, variance, chi-square, Fourier indices, amongst many others. The statistics simply require three vectors for each light curve, encoding the time, magnitude, and magnitude error.

We provide a summary of the metrics employed in Table 2. While the accompanying catalogue paper (van Roestel et al. 2021) will discuss their use extensively, to demonstrate their efficacy, Fig. 4 shows the probability density for a subset of the features in the analysis for a ‘variable’ and ‘non-variable’ set of objects. We define ‘variable’ objects as those that statistically change from observation to observation inconsistent with their assigned error bars; ‘non-variable’ objects do not display these traits at the signal-to-noise reached by ZTF. A number show clear differentiation between the two sets, indicating their suitability as metrics for this purpose. We note that while some metrics have similar marginalized distributions, this marginalized format masks potential higher order correlations between parameters, and therefore we choose to use all metrics in our classifier.

4 PERIOD FINDING

We use period-finding algorithms to estimate periods for all objects; phase-folded light curves, such as those in Fig. 10, are used when scanning and classifying sources, typically for those with significant periodicities. There are a variety of period-finding algorithms in the literature (see Graham et al. 2013b for a comparison and review), including those based on least-squares fitting to a set of basis functions (Zechmeister & Kürster 2009; Mortier et al. 2015; Mortier & Collier Cameron 2017). By far the dominant source of computational burden is in the period finding. The algorithm we employ is hierarchical, with two period-finding algorithms supplying candidate frequencies to a third.

The first algorithm is a conditional entropy (CE; Graham et al. 2013a) algorithm. CE is based on an information theoretic approach; broadly, information theory-based approaches improve on other techniques by capturing higher order statistical moments in the data, which are able to better model the underlying process and are more robust to noise and outliers. Graphically, CE envisions a folded light curve on a unit square, with the expectation that the true period will result in the most ordered arrangement of data points in the region. More specifically, the algorithm minimizes the entropy associated with the system relative to the phase, which naturally accounts for the non-trivial phase space coverage of real data. The CE version we use is GCE (Katz et al. 2020).⁵ It is implemented on Graphics Processing Units (GPU) in CUDA (Nickolls et al. 2008) and wrapped to python using Cython (Behnel et al. 2011) and a special CUDA wrapping code (McGibbon & Zhao 2019).

The software GCE is a PYTHON code that prepares light curves and their magnitude information for input into the Cython-wrapped CUDA program. Therefore, the user interface is entirely PYTHON-based. The period range searched varies between 30 min and half of the baseline, T ; 30 min was chosen as a lower bound for computational considerations. We use a frequency step df of $\frac{1}{3 \times T}$, oversampled by a factor of 3 in order to account for the irregular sampling. This oversampling term was measured empirically by

⁵<https://github.com/mikekatz04/gce>

Table 2. Statistics calculated based on the light curves and period finding. N is the number of observations, m_i is the i th observation magnitude, and σ_i is the i th observation magnitude error. The variables c_1 , c_2 and the a_i 's and b_i 's are constants to be fit for in statistics 23–35 and 37. In addition to these statistics, we save the best-fitting period based on the hierarchical analysis as well as its significance.

Index	Statistic	Calculation
1	N	Number of observations passing cuts
2	m_{median}	Median magnitude
3	m_{mean}	Weighted mean
4	m_{var}	Weighted variance
5	χ^2	$\frac{1}{N-1} \sum_i \frac{(m_{\text{median}} - m_i)^2}{\sigma_i^2}$
6	RoMS	$\frac{1}{N-1} \sum_i \frac{ m_{\text{median}} - m_i }{\sigma_i}$
7	Median absolute deviation	$\text{median}(m - m_{\text{median}})$
8	Normalized peak-to-peak amplitude	$\frac{\max(m - \sigma) - \min(m + \sigma)}{\max(m - \sigma) + \min(m + \sigma)}$
9	Normalized excess variance	$\frac{1}{Nm_{\text{mean}}^2} \sum_i (m_{\text{mean}} - m_i)^2 - \sigma_i^2$
10–14	Ranges	Inner 50 per cent, 60 per cent, 70 per cent, 80 per cent, and 90 per cent range
15	Skew	$\frac{N}{(N-1)(N-2)} \sum_i \frac{(m_{\text{mean}} - m_i)^3}{\sigma_i^3}$
16	Kurtosis	$\frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_i \frac{(m_{\text{mean}} - m_i)^4}{\sigma_i^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$
17	Inverse Von Neumann statistic	$\eta = \left(\frac{1}{\sum_i \left(\frac{1}{\Delta t_i} \right)^2 m_{\text{var}}} \right) \sum_i \left(\frac{\Delta m_i}{\Delta t_i} \right)^2$ where $\Delta t_i = t_{i+1} - t_i$ and $\Delta m_i = m_{i+1} - m_i$
18	Welch/Stetson I	$\frac{N}{N-1} \sum_i \left(\frac{m_i - m_{\text{mean}}}{\sigma_i} \right) \left(\frac{m_{N-i} - m_{\text{mean}}}{\sigma_{N-i}} \right)$
19	Stetson J	$\sqrt{\frac{N}{N-1}} \sum_i \text{sgn}((m_i - m_{\text{mean}})(m_{N-i} - m_{\text{mean}})) \sqrt{\left \left(\frac{m_i - m_{\text{mean}}}{\sigma_i} \right) \left(\frac{m_{N-i} - m_{\text{mean}}}{\sigma_{N-i}} \right) \right }$
20	Stetson K	$\sqrt{\frac{1}{N}} \sum_i \left \left(\frac{m_i - m_{\text{mean}}}{\sigma_i} \right) \left(\frac{m_{N-i} - m_{\text{mean}}}{\sigma_{N-i}} \right) \right / \sqrt{\sum_i \left(\frac{m_i - m_{\text{mean}}}{\sigma_i} \right) \left(\frac{m_{N-i} - m_{\text{mean}}}{\sigma_{N-i}} \right)^2}$
21	Anderson–Darling test	Stephens (1974)
22	Shapiro–Wilk test	Shapiro & Wilk (1965)
23–35	Fourier decomposition	$y = c_1 + c_2 \times t + \sum_{i=1}^5 (a_i \cos(\frac{2\pi t i}{P}) + b_i \sin(\frac{2\pi t i}{P}))$
36	Bayesian information criterion	χ^2 likelihood keeping different number of Fourier harmonics
37	Relative χ^2	$y = c_1 + c_2 \times t$ to full Fourier decomposition

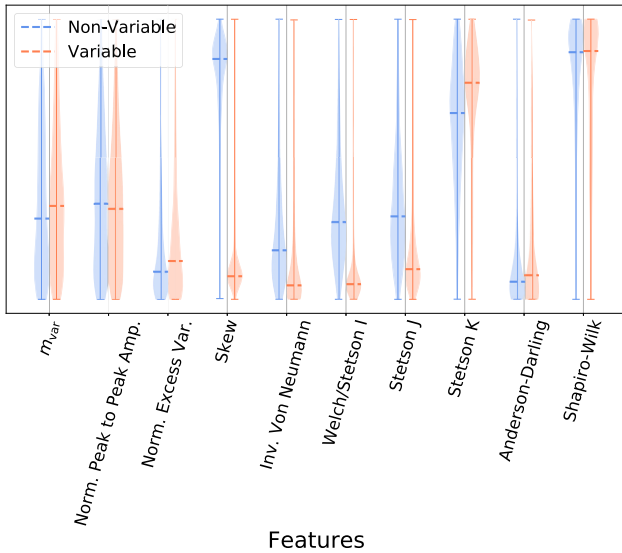


Figure 4. Probability density for a subset of the features in the analysis for a ‘variable’ and ‘non-variable’ set of objects (van Roestel et al. 2021), as assessed by a machine learning algorithm XGBoost (Chen & Guestrin 2016). The dashed lines correspond to the 50th percentile for the features. For simplicity, we have normalized all features such that they appear on the same plot, with values increasing in the upwards direction on the otherwise arbitrary y-axis.

trying the frequency grid on a variety of test cases, and shown to be the minimum factor that gave the correct period for a small sample of eclipsing binaries and RR Lyrae as determined by a higher resolution analysis. To evaluate the efficacy of this choice, Fig. 6 shows a two-dimensional histogram of the relative difference between an analysis of variance computation with an oversampling factor of 3, as used in the main analysis, and an oversampling factor of 10. We plot this relative difference as a function of the highest significance period identified by the oversampling factor of 10 analysis. We plotted any relative difference values below that of 10^{-3} in the bottom row of the histogram, indicating agreement at the 0.1 per cent level. The main feature in this histogram, beyond the build-up of support of equal periods at the bottom, is the approach of a curve to a relative difference of 1.0, indicating those sources that disagree by a factor of 2 are modulated by the effect of diurnal sampling. There is a less dense horizontal line at those sources at half of the frequency. This analysis indicates that a more refined period grid, perhaps as a secondary step in the analysis, may be useful in the future. We note that phasing agreement to better than the 0.1 per cent level is already achieved, and accuracies at this level are required to, for example, detect small but detectable period changes present in a variety of astrophysical processes.

We note that this baseline will change for each chunk analysed; the frequency array is then simply $f = f_{\text{min}} + df \times [0 \dots \text{ceil}((f_{\text{max}} - f_{\text{min}})/df)]$. We use 20 phase bins and 10 magnitude bins in the conditional entropy calculation, which amounts to the size of the

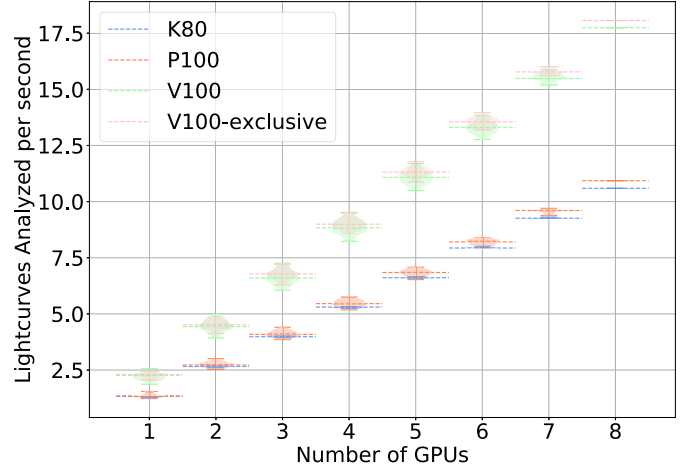
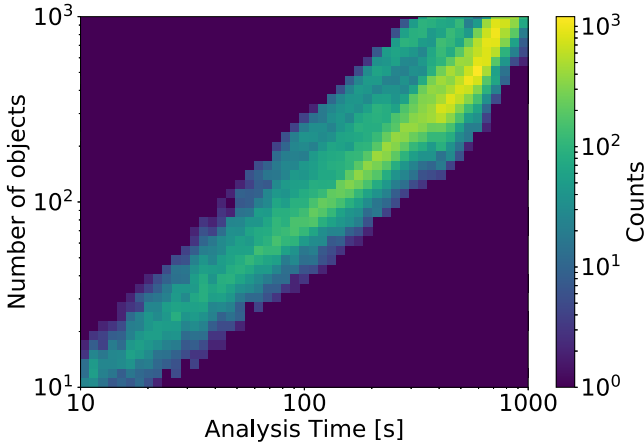


Figure 5. Left: Two-dimensional histogram of the number of objects analysed versus analysis time for the ZTF light curves. Right: Scaling data based on GPU resources for the algorithm discussed in the text. We note that the V100- and V100-exclusive analysis (where jobs are restricted to run on their own on the GPU) are closely overlapping.

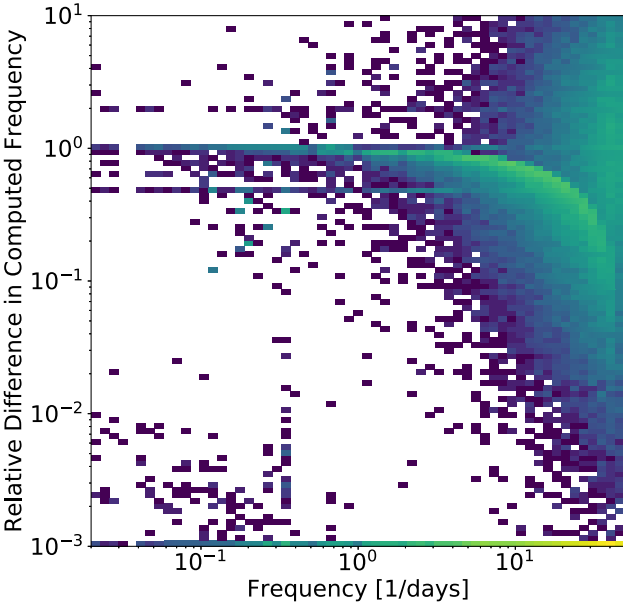


Figure 6. Two-dimensional histogram of the relative difference between an analysis of variance computation with oversampling factors of 3, as used in the main analysis, and an oversampling factor of 10 versus highest significance period identified by the oversampling of 10 analysis.

phase-folded, two-dimensional histogram. We note that conditional entropy, as a two-dimensional binning algorithm, does not currently have weights implemented. In principle, this could be done similarly to the case of Lomb–Scargle below, although careful consideration should be taken for how to handle, for example, eclipsing systems where downweighting fainter points with correspondingly larger error bars could be detrimental to identifying eclipses.

The second algorithm is a Lomb–Scargle (LS; Lomb 1976; Scargle 1982) implementation; these algorithms are similar to Fourier transforms, but for irregularly sampled data. In this version, it decomposes the time series into the frequency domain using a linear combination of sine waves $y = a \cos \omega t + b \sin \omega t$. If we define T as the period

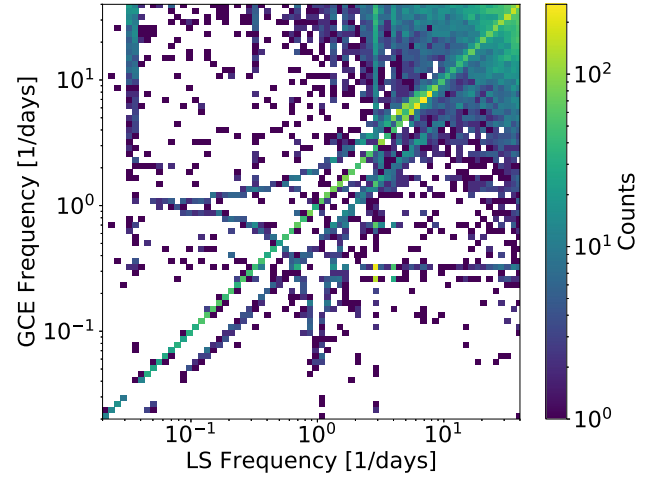


Figure 7. Two-dimensional histogram comparing LS (x-axis) and GCE (y-axis).

with the angular frequency $\omega = 2\pi/T$, the periodogram is defined as:

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

where

$$w_n = \frac{1}{\sigma_n^2} \bigg/ \sum_{n=1}^N \frac{1}{\sigma_i^2}; \tau = \tan(2\omega\tau) = \frac{\sum_{n=1}^N \sin(2\omega t_n)}{\sum_{n=1}^N \cos(2\omega t_n)}. \quad (1)$$

The algorithm we use is also implemented on GPUs.⁶

Fig. 7 shows a two-dimensional histogram comparing LS (x-axis) and GCE (y-axis). For many objects, the peak frequencies identified

⁶<https://github.com/johnh2o2/cuvarbase>

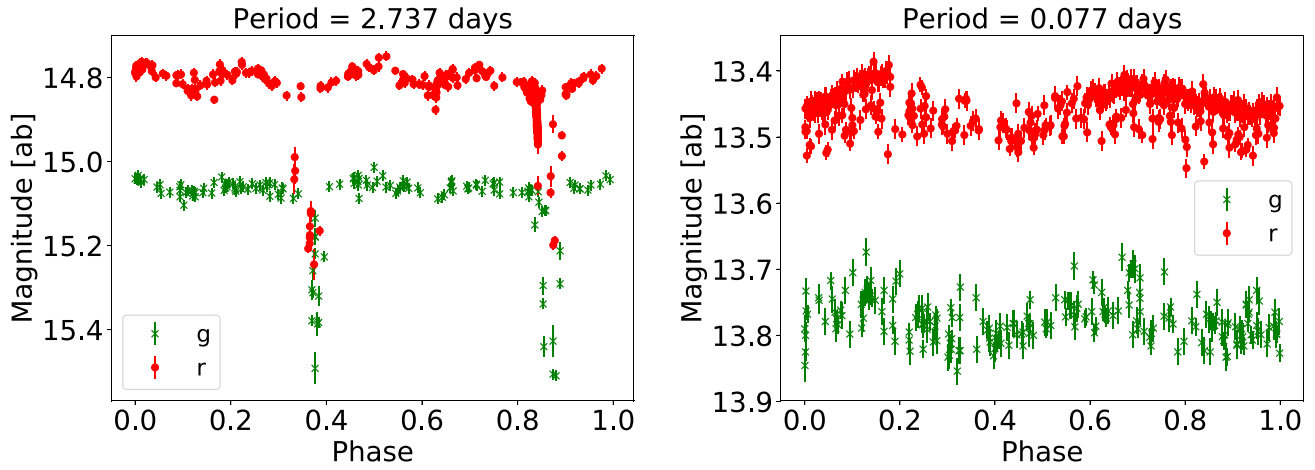


Figure 8. Example periodic variables (their estimated periods are given in the titles), identified by GCE (left) and LS (right) as high significance, while the other period-finding algorithm finds marginal significance.

are in agreement, with a subset identified as differing by half or twice the period. When this occurs, LS preferentially finds the shorter period, which occurs in a number of common scenarios. This includes when analysing eclipsing binaries whose primary and secondary eclipses do not differ greatly in depth; in this case, LS tends to find a period equal to half the true value, while CE will find the correct value. For a further subset, the peak frequencies identified are different, requiring, in principle, a tie-breaker algorithm to determine the ‘best’ choice. A further, interesting feature in these histograms are curved ‘lines,’ symmetrical about the diagonal, that have vertical and horizontal asymptotes at periods corresponding to 1 d, which arise from the true period ‘beating’ with a 1-d period arising from diurnal aliasing.

Fig. 8 shows example periodic variables identified by GCE and LS as high significance, while the other period-finding algorithm finds marginal significance; the GCE example has both eclipsing and underlying modulating behaviour, more difficult for LS to recover, while the LS example identifies low-amplitude modulation in otherwise noisy data, difficult for GCE to recover within its two-dimensional histograms. To address this issue, we take the top 50 frequencies identified by each of the algorithms, with no use of harmonic summing or otherwise period-combining techniques, and run each through a CPU-based multiharmonic analysis of variance (AOV; Schwarzenberg-Czerny 1998) code in the 200 frequency bins, covering a frequency range between peak frequency – $100 \times df$ to peak frequency + $100 \times df$.⁷ To determine the best period, we evaluate their *ad hoc* ‘significance’ using the statistic array s , which is the same length as the frequency array:

$$\text{significance} = (\max(s) - \text{mean}(s)) / \text{std}(s). \quad (2)$$

We note that s , which is parameterized by the period array, corresponds to, for example, the entropy in the phased two-dimensional histograms for conditional entropy, or an estimate of the Fourier power for LS. While other examples of ‘significance’ are also possible, such as approximate significance estimates for LS (Baluev 2008), we find this is sufficient for simply rank-ordering the frequencies across all algorithms and make simple comparisons between objects. We also point out that this technique throws away information from

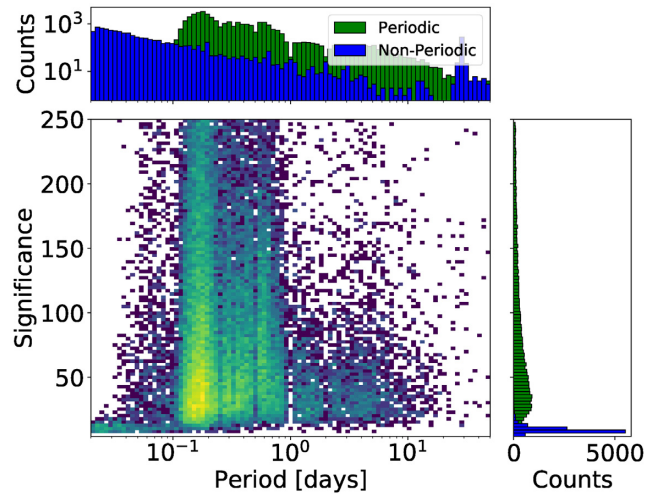


Figure 9. Two-dimensional histograms of the significance versus period for the ‘periodic’ set of objects (van Roestel et al. 2021). The one-dimensional histograms are marginalized versions of the two-dimensional histogram (green). For comparison, we include a ‘non-periodic’ set of objects in blue (van Roestel et al. 2021).

sub-dominant periods in a light curve, and so for objects where multiple periods are important, this technique will be suboptimal. We can evaluate the threshold for significance appropriate for evaluating a confident periodic source; in Fig. 9, we plot the significance versus period for the ‘periodic’ set of objects (van Roestel et al. 2021). For comparison, we show marginalized one-dimensional histograms for both the periodic and non-periodic sets of objects that show distinct differences in both. The non-periodic set show a distinct peak at low frequencies and around the lunar cycle, while the periodic set peaks distinctly in the range 0.1–1 d due to the high rate of RR Lyrae and Delta Scuti variables in this set. For scanning purposes, we can also compare the distribution of significances for these sets; for example, the second percentile for periodic objects is ~ 14.4 , which corresponds to the 95th percentile of non-periodic objects. This would mean that a significance threshold of 14.4 would yield a false dismissal probability of 2 per cent while having a false alarm

⁷<https://github.com/joshuazd/lssfds>

probability of 5 per cent. For further comparison, the relative rate of periodic to non-periodic sources is ~ 0.13 per cent.

There were two main advantages to this otherwise complicated method. The first is that there are known advantages to both CE and LS (Graham et al. 2013a), with one particularly successful for generic light curve shapes and the second for low-amplitude variables, as demonstrated in Fig. 8. To provide a single set of metrics based on these two algorithms, we needed a way to ‘choose’ between CE and LS, from which came the choice of AOV, which has some of the benefits of both (generic light curve shapes with the potential for multiple harmonics). These motivate the second point, which is that the CPU-based version of AOV across the whole frequency grid would be computationally intractable, and so some of the benefits of the most suitable period-finding algorithm are achieved despite its computational expense. The GPU implementations of CE and LS were essential here, as CPU implementations of CE, AOV, and LS take ~ 10 s per light curve each, two orders of magnitude slower than their GPU counterparts. We highlight a few example periodic objects coming out of this analysis in Fig. 10.

We show the analysis speed as a function of number of objects on the left of Fig. 5. As can be seen, the analysis takes ~ 1 s per light curve across the range of light curves analysed at a single time (~ 100 to ~ 1000 samples). At this point, we are dominated by the CPU cost of the AOV analysis, having put all other expensive computations on the GPU. Design of a GPU-based AOV analysis is ongoing, as well as breaking out the AOV process into separate processes (perhaps for objects that meet specific variability requirements). We now perform scaling tests for the numerical setup described above. The right of Fig. 5 shows the scaling test results on 10 GPUs running concurrently. These tests have been performed at a number of facilities: (i) the Minnesota Supercomputing Institute, with both K40 (11 GB of RAM) and V100 GPUs (16 GB of RAM), (ii) XSEDE’s SDSC COMET cluster (Townsend et al. 2014) with K80 (12 GB of RAM) and P100 GPUs (16 GB of RAM), and (iii) NERSC’s CORI cluster with V100’s (16 GB of RAM). As expected, the number of analysed light curves is linear with the number of GPUs allocated.

Fig. 11 shows the two-dimensional histogram of the period significance based on the analysis of variance computation versus highest significance period identified for field 296 (top left), 487 (top right), 682 (bottom left), and field 853 (bottom right). We can identify a number of cases where a non-astrophysical signal is found due to the sampling pattern. It is clear based on the histograms that fractions and multiples of a sidereal day, as well as near sidereal months with the lunar cycle, induce false period estimates. We remove at the period finding level bands around fractions and multiples of a sidereal day to help mitigate this. Specifically, we remove frequencies (in cycles/day) of 47.99–48.01, 46.99–47.01, 45.99–46.01, 3.95–4.05, 2.95–3.05, 1.95–2.05, 0.95–1.05, 0.48–0.52, and 0.03–0.04; this removes ~ 1 percent of the frequency range. We chose these frequencies based on an iterative procedure; we would evaluate a subset of objects, examine histograms of period excesses identified at these frequencies, remove them, and then reevaluate. This will be convolved with longer term trends, including seasonal and annual variations. We note that removal of these frequencies may have removed the true variables covering these period ranges. The marginalized histograms in Fig. 11 indicate that a small excess of sources at the edge of the removed frequency bands remains due to the diurnal and lunar aliases, showing that a broader removal could be useful.

The top left of Fig. 12 shows a comparison between period finding analyses of field 700 using the setup used in this analysis and one where we do not remove these bands; in general, most of the objects are assigned the same period. In addition to some differences due

to GPU errors (see below) and the differences that arise in the significance calculation where different frequencies reach the AOV stage, a significant fraction receive periods in narrow frequency bands that are otherwise cut, indicating the importance of these cuts. The top right of Fig. 12 shows a comparison between period finding analyses of field 700 using the setup used in this analysis and one where periods are searched down to 3 min; similar to before, in general, many of the objects are assigned the same period. Some objects are assigned a period that corresponds to the second harmonic of the original analysis. Also, some objects have very short periods assigned, as typically occurs for marginal periodic signals.

We also characterized GPU-based transient errors that affect a small fraction of objects when running on the HPC clusters used here (Tiwarei et al. 2015). The bottom row of Fig. 12 shows a comparison between two identical period finding analyses of field 700 using the setup used in this analysis and one where the GPU-based period finding is run three times; while two outliers remain, this is enough to clean up the distributions at the cost of computational efficiency. We expect that if there are algorithmic speed-ups enabling this to be performed feasibly over the entire data set, it will be useful to mitigate this issue.

5 CONCLUSION

In this paper, we introduce the variability metrics and the periodicity algorithms that are used to derive ZTF’s variable star catalogue based on ZTF’s DR2. In addition to the variability metrics suggested by Pashchenko et al. (2017), we included robust measurements of the light curve amplitude and a Fourier decomposition to characterize the shape of the folded light curves. We designed a hierarchical period-finding scheme that utilized multiple period-finding algorithms in order to be as sensitive as possible across the period parameter space of amplitude and period. This analysis provides the input statistical features that will provide training sets for the machine learning-based catalogues, presented in a partner publication (van Roestel et al. 2021). This publication will present classifiers and associated metrics that evaluate the variability, periodicity, completeness, purity, and classifications for these objects. To create the catalogue, *ztf-periodic* has been applied to a variety of large-scale computing systems and NVIDIA-based GPUs. Setting up and executing the proposed analysis required access to GPU arrays of this size, and the computational needs will only grow in future releases from ZTF and larger data sets from future surveys.

Going forward, the priority will be to analyse the data from the ZTF’s Third Data Release (2020 June 24).⁸ Many of the choices made in this paper should be revisited for this process. These include:

- (i) Restricting to one point within 30 min (to remove high cadence data)
- (ii) Restricting to periods greater than 30 min (to make computing tractable)
- (iii) Restricting to a minimum of 50 points per light curve (for metric efficacy)
- (iv) Period range exclusions (to remove effects from cadence aliasing)
- (v) Period finding choices, including the choice of the oversampling factor

Outside of more epochs yielding more light curves passing the 50 point cut, we expect to revisit the choices made in the period

⁸<https://www.ztf.caltech.edu/page/dr3>

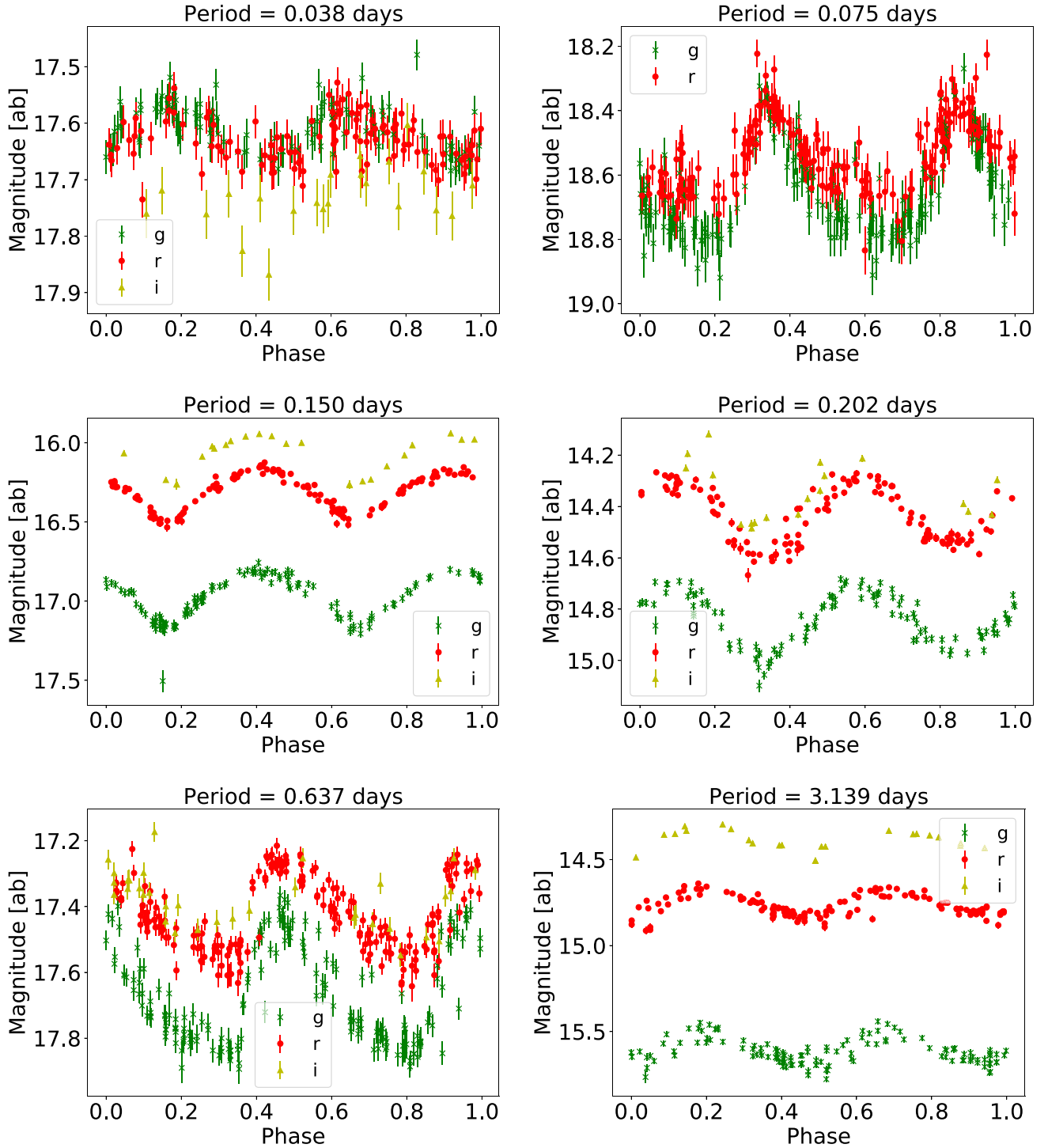


Figure 10. Example periodic variables (their computed periods are given in the titles) folded at twice their computed period, identified in the various bins we use for scanning. We fold at twice the computed period to evaluate the consistency of the reconstructed light curves across two orbits, useful for assessing the correctness of the period. These bins are chosen to simplify comparison of periodicity significances estimated from source variable on very different time scales: top left, 30–60 min (delta Scuti), top right, 1–2 h (delta Scuti), middle left 2–4 h (W UMa), middle right, 4–12 h (W UMa), bottom left, 12–72 h (RRab star), and bottom right, 3–10 d (Cepheid). For W UMa's, the true period is two times longer than the computed period.

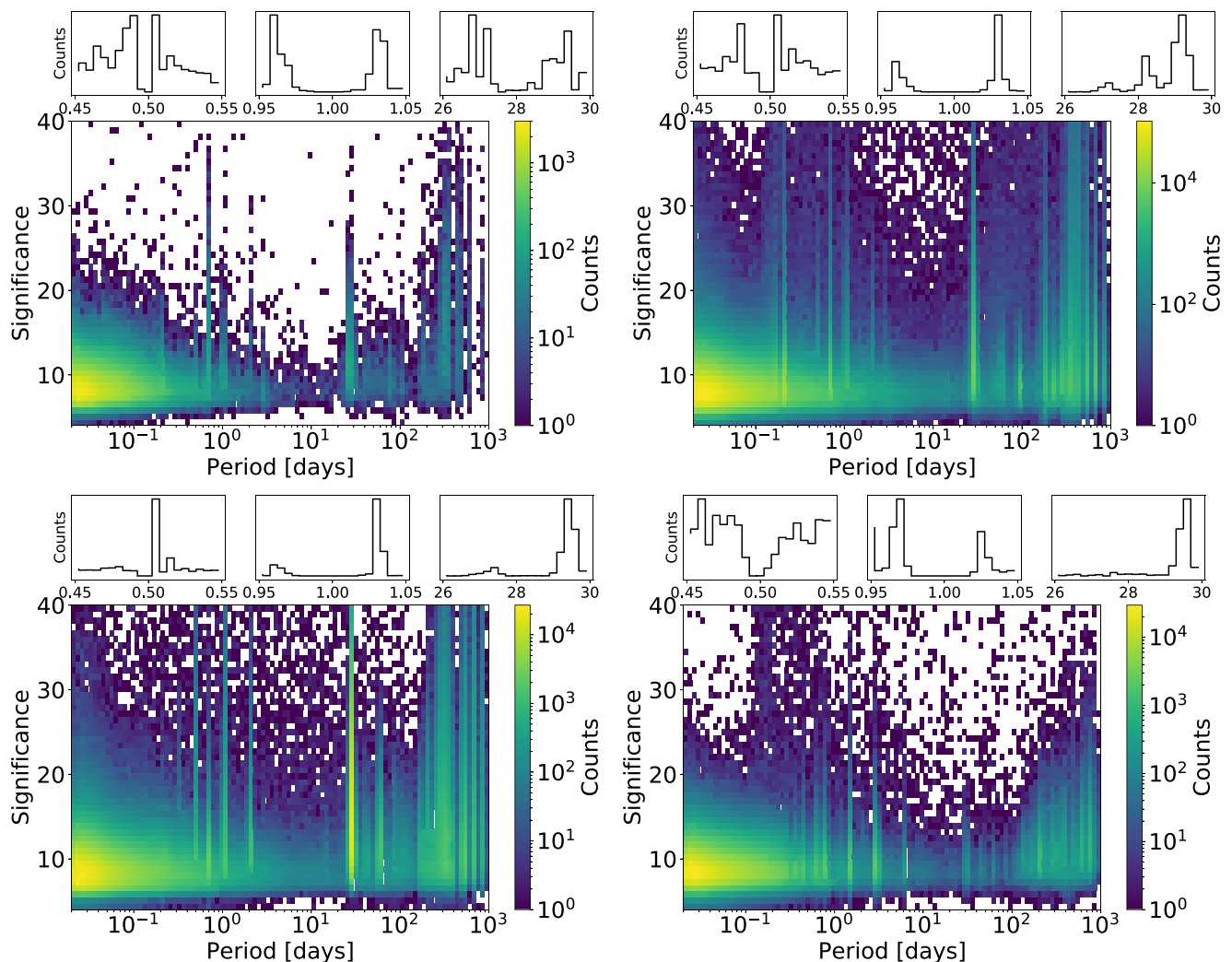


Figure 11. Two-dimensional histograms of the period significance based on the analysis of variance computation versus highest significance period identified for field 296 (top left), 487 (top right), 682 (bottom left), and field 853 (bottom right). The one-dimensional histograms focus on periods of 0.5 d and 1 d to evaluate the effect of diurnal sampling on aliasing, and 28 d to evaluate the same for the lunar cycle.

finding. Ongoing work includes devising faster algorithms to improve the scaling of the processing. Other options include combining photometric points from other surveys.

We also are exploring the benefits of ‘clipping’ light curves for outlier removal, either using a comparison with a given σ difference or percentile based cuts. In general, we continue to improve the algorithms looking ahead, given that computational burdens are growing significant. For LSST in particular, the number of variable sources expected is significantly higher, because of both its increased depth and ~ 5 mmag precision, given that the fraction of variables increases as a power law with increasing photometric precision (Huber, Everett & Howell 2006). Translation of the work described here to LSST is obviously of great interest, as LSST will overlap with the faint end of ZTF, and have a limiting magnitude far beyond

what ZTF can reach, meaning that it can access a far larger volume of sources. However, translating the work described here to LSST will be accompanied by several challenges. Obviously the cadence will be substantially lower due to LSST’s smaller field of view relative to ZTF, which will impact primarily the ability to recover periods at the short end. Perhaps the biggest challenge is that LSST will not acquire sufficient samples to detect periodic behaviour in many sources until several years into the survey, which will increase the baseline of the sampling relative to that of ZTF; this in turn means that a larger frequency grid must be searched to recover equivalent objects. Additionally, because LSST will contain far more sources than ZTF due to its depth, the computational cost will be significantly higher, as it is proportional to the total number of sources. This further motivates continued speed-ups of the algorithms presented here.

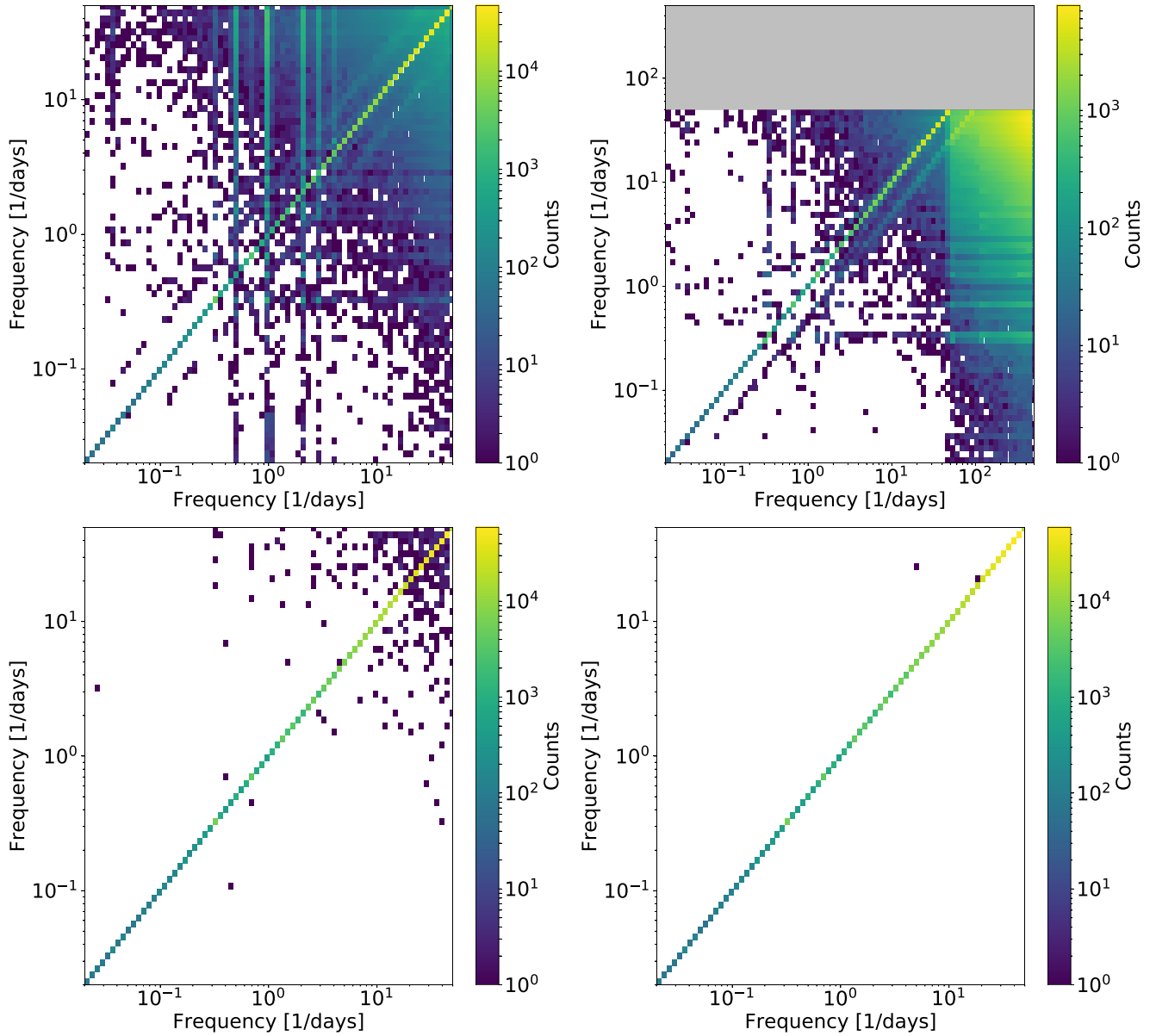


Figure 12. On the top left is a comparison between two period finding analyses of field 700 comparing one without the aliasing-dominated frequency bands removed (x-axis) and the typical setup (y-axis). On the top right is the same where the x-axis version has a period range searched down to 3 min. The grey region corresponds to a period region not available to that run. On the bottom row, we compare runs where we use the typical setup (left) and one where the GPU-based period finding is run three times (right).

ACKNOWLEDGEMENTS

The authors would like to thank our referee, Dr. Aren Heinze, for very useful reports, improving the content of the paper. MWC acknowledges support from the National Science Foundation with grant number PHY-2010970. MLK acknowledges support from the National Science Foundation under grant DGE-0948017 and the Chateaubriand Fellowship from the Office for Science & Technology of the Embassy of France in the United States. MR has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement n759194 - USNAC). DK is supported by NSF grant AST-1816492.

The authors acknowledge the Minnesota Supercomputing Institute⁹ (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper under project ‘Identification of Variable Objects in the Zwicky Transient Facility.’ This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231 under project ‘Towards a complete catalog of variable sources to support efficient searches for compact binary mergers and their products.’ This work

⁹<http://www.msi.umn.edu>

used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) COMET at SDSU through allocation AST200016. MLK also acknowledges the computational resources and staff contributions provided for the Quest/Graill high performance computing facility at Northwestern University.

Based on observations obtained with the Samuel Oschin Telescope 48-inch and the 60-inch Telescope at the Palomar Observatory as part of the Zwicky Transient Facility project. ZTF is supported by the National Science Foundation under Grant No. AST-1440341 and a collaboration including Caltech, IPAC, the Weizmann Institute for Science, the Oskar Klein Center at Stockholm University, the University of Maryland, the University of Washington (UW), Deutsches Elektronen-Synchrotron and Humboldt University, Los Alamos National Laboratories, the TANGO Consortium of Taiwan, the University of Wisconsin at Milwaukee, and Lawrence Berkeley National Laboratories. Operations are conducted by Caltech Optical Observatories, IPAC, and UW.

DATA AVAILABILITY

The data underlying this article are derived from public code found here: <https://github.com/mcoughlin/ztfperiodic>. DR2, including how to access the data, can be found at <https://www.ztf.caltech.edu/page/dr2>.

REFERENCES

- Alcock C. et al., 2000, *ApJ*, 542, 281
- Amaro-Seoane P. et al., 2017, Laser Interferometer Space Antenna. <https://arxiv.org/abs/1702.00786>
- Antonini F., Toonen S., Hamers A. S., 2017, *ApJ*, 841, 77
- Baluev R. V., 2008, *MNRAS*, 385, 1279
- Banerjee S., 2018, *MNRAS*, 473, 909
- Behnel S., Bradshaw R., Citro C., Dalcin L., Seljebotn D. S., Smith K., 2011, *Comput. Sci. Eng.*, 13, 31
- Bellm E. C. et al., 2018, *PASP*, 131, 018002
- Bellm E. C. et al., 2019, *PASP*, 131, 068003
- Bhatti W., Bouma L., Joshua J., Price-Whelan A., 2020, doi: 10.5281/zenodo.3723832
- Breivik K., Mingarelli C. M. F., Larson S. L., 2020, *ApJ*, 901, 4
- Burdge K. B. et al., 2019a, *Nature*, 571, 528
- Burdge K. B. et al., 2019b, *ApJ*, 886, L12
- Cannizzo J. K., Nelemans G., 2015, *ApJ*, 803, 19
- Catelan M., 2009, *Ap&SS*, 320, 261
- Chen T., Guestrin C., 2016, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York
- Coughlin M. W. et al., 2019a, *PASP*, 131, 048001
- Coughlin M. W. et al., 2019b, *ApJ*, 885, L19
- Coughlin M. W. et al., 2020, *MNRAS*, 494, L91
- Dekany R. et al., 2020, *PASP*, 132, 038001
- Drake A. J. et al., 2009, *ApJ*, 696, 870
- Drake A. J. et al., 2014, *ApJS*, 213, 9
- Duev D. A. et al., 2019, *MNRAS*, 489, 3582
- Freedman W. L. et al., 2001, *ApJ*, 553, 47
- Fuller J., Lai D., 2011, *MNRAS*, 412, 1331
- Gaia Collaboration, 2018, *A&A*, 616, A1
- Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., 2013a, *MNRAS*, 434, 2629
- Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., Duan V., Maker A., 2013b, *MNRAS*, 434, 3423
- Graham M. J. et al., 2019, *PASP*, 131, 078001
- Hoffleit D., Jaschek C., 1991, in Hoffleit D., Jaschek C., eds, The Bright Star Catalogue, 5th rev. edn.. Yale University Observatory
- Huber M. E., Everett M. E., Howell S. B., 2006, *AJ*, 132, 633
- Ivezic Z., Tyson J. A., Allsman R., Andrew J., Angel R., 2019, *ApJ*, 873, 111
- Katz M. L., Cooper O. R., Coughlin M. W., Burdge K. B., Breivik K., Larson S. L., 2021, *MNRAS*, 503, 2665
- Kochanek C. S. et al., 2017, *PASP*, 129, 104502
- Kremer K., Chatterjee S., Breivik K., Rodriguez C. L., Larson S. L., Rasio F. A., 2018, *Phys. Rev. Lett.*, 120, 191103
- Kupfer T. et al., 2018, *MNRAS*, 480, 302
- Kupfer T. et al., 2019, *ApJ*, 878, L35
- Kupfer T. et al., 2020, *ApJ*, 898, L25
- Lomb N. R., 1976, *Ap&SS*, 39, 447
- Lopes C. E. F. et al., 2020, *MNRAS*, 496, 1730
- Masci F. J. et al., 2018, *PASP*, 131, 018003
- McGibbon R., Zhao Y., 2019, npcuda-example. <https://github.com/rmcgibbon/npcuda-example>
- Morgan J. S., Kaiser N., Moreau V., Anderson D., Burgett W., 2012, *Proc. SPIE - Int. Soc. Opt. Eng.*, 8444, 0H
- Mortier A., Collier Cameron A., 2017, *A&A*, 601, A110
- Mortier A., Faria J. P., Correia C. M., Santerne A., Santos N. C., 2015, *A&A*, 573, A101
- Naul B., van der Walt S., Crellin-Quick A., Bloom J. S., Pérez F., 2016, Cesium: Open-Source Platform for Time-Series Inference. <https://arxiv.org/abs/1609.04504>
- Nelemans G., Tout C. A., 2005, *MNRAS*, 356, 753
- Ngeow C. C., Lee C. D., Yu P. C., Masci F., Laher R., Kupfer T., Golkhou V. Z., ZTF Collaboration, 2019, in Malasan H. L. et al. *Journal of Physics Conference Series*, Vol. 1231, *Journal of Physics Conference Series*. Institute of Physics Publishing, p.012010
- Nickolls J., Buck I., Garland M., Skadron K., 2008, *Queue*, 6, 40
- Nun I., Protopapas P., Sim B., Zhu M., Dave R., Castro N., Pichara K., 2015, *preprint (arXiv:1506.00010)*
- Pashchenko I. N., Sokolovsky K. V., Gavras P., 2017, *MNRAS*, 475, 2326
- Saha A., 1984, *ApJ*, 283, 580
- Saha A., 1985, *ApJ*, 289, 310
- Scargle J. D., 1982, *ApJ*, 263, 835
- Schwarzenberg-Czerny A., 1998, *Balt. Astron.*, 7, 43
- Shapiro S. S., Wilk M. B., 1965, *Biometrika*, 52, 591
- Shappee B. J. et al., 2014, *ApJ*, 788, 48
- Stephens M. A., 1974, *J. Am. Stat. Assoc.*, 69, 730
- Stetson P. B., 1996, *PASP*, 108, 851
- Szkody P. et al., 2020, *AJ*, 159, 198
- Tiwari D. et al., 2015, 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA). IEEE, Burlingame, CA, p. 331
- Tonry J. L. et al., 2018, *PASP*, 130, 064505
- Towns J. et al., 2014, *Comput. Sci. Eng.*, 16, 62
- Udalski A., 2003, *AcA*, 53, 291
- Udalski A., Szymański M. K., Szymański G., 2015, *AcA*, 65, 1
- van Roestel J. et al. 2021, *AJ*, 161, 267
- Zechmeister M., Kürster M., 2009, *A&A*, 496, 577

This paper has been typeset from a \LaTeX file prepared by the author.