Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **434,** 3423–3444 (2013) doi:10.1093/mnras/stt1264
Advance Access publication 2013 August 1

A comparison of period finding algorithms

Matthew J. Graham,⋆ Andrew J. Drake, S. G. Djorgovski, Ashish A. Mahabal, Ciro Donalek, Victor Duan and Allison Maker

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

Accepted 2013 July 8. Received 2013 July 8; in original form 2013 May 7

ABSTRACT

This paper presents a comparison of popular period finding algorithms applied to the light curves of variable stars from the Catalina Real-Time Transient Survey, MACHO and ASAS data sets. We analyse the accuracy of the methods against magnitude, sampling rates, quoted period, quality measures (signal-to-noise and number of observations), variability and object classes. We find that measure of dispersion-based techniques – analysis of variance with harmonics and conditional entropy – consistently give the best results but there are clear dependences on object class and light-curve quality. Period aliasing and identifying a period harmonic also remain significant issues. We consider the performance of the algorithms and show that a new conditional entropy-based algorithm is the most optimal in terms of completeness and speed. We also consider a simple ensemble approach and find that it performs no better than individual algorithms.

Key words: methods: data analysis – techniques: photometric – astronomical data bases: miscellaneous – virtual observatory tools.

1 INTRODUCTION

The last decade has seen the emergence of large collections of time series data from searches for microlensing, e.g. MACHO (Alcock et al. 2003), OGLE (Udalski et al. 1993) and exoplanets, e.g. CoRoT (Auvergne et al. 2009), Kepler (Koch et al. 2010), as well as legacy variability collections, e.g. ASAS (Pojmanski 2002). For the first time, these are amenable to statistical and machine learning-based analyses, particularly for classification and outlier detection, e.g. Debosscher et al. (2007), Shin, Sekora & Byun (2009), Dubath et al. (2011), Richards et al. (2011), with an eye to the new generation of synoptic sky surveys, e.g. Catalina Real-time Transient Survey (CRTS; Drake et al. 2009), PTF (Rau et al. 2009), Pan-STaRRs (Kaiser et al. 2002) and LSST (Ivezić et al. 2011), which will increase the amount of available data by several orders of magnitude. These surveys are also not unique to optical wavelengths with efforts underway across the electromagnetic spectrum – LOFAR and SKA and its pathfinder precursor projects in the radio, IR, X-ray – as well as in the more exotic regimes of particle astrophysics (neutrino) and gravitational waves (LIGO). Although many different approaches have been attempted, they all follow the same basic pattern: characterization, categorization and classification.

Time series vary widely in their temporal coverage, sampling rates and regularity, number of points and error bars, making a very disparate data set. Comparing raw light curves[1] is therefore difficult; rather a representation of each light curve in a given data collection in terms of a feature set is required for any analysis. There is no standardized set – somewhere around 100 different features[2] have been used or suggested in the literature for characterizing time series, e.g. moments, flux and shape ratios, variability indices, etc. – but many of them rely on a derived period for an object, even when it does not necessarily display any periodic behaviour. Dubath et al. (2011) show a ∼11 per cent misclassification error rate for non-eclipsing variable stars with an incorrect period. Richards et al. (2011) also estimate that periodic feature routines account for 75 per cent of the computing time used in their time series characterization.

This irregularity means that astronomical time series data does not lend itself to the standard Fourier-based analysis techniques that are found in general statistics literature. Consequently, there is a long history in period finding algorithms with common ones based upon discrete Fourier transform (Deeming 1975), or a least-squares approximation to it (LS; Lomb 1976; Scargle 1982), string length (STR) (Dworetsky 1983), phase dispersion minimization (PDM; Jurkevic 1975; Stellingwerf 1978) and analysis of variance (AOV)

[1] Note that we use the terms 'time series' and 'light curve' interchangeably in this paper.

[2] A feature is defined as an individual measurable heuristic property of an object that can be used to characterize it.

⋆ E-mail: mjg@caltech.edu

methods (Schwarzenberg-Czerny 1989, 1996), as well as a host of others (e.g. Huijse et al. 2011; Kato & Uemura 2012).

Obviously with so many different methods, the question arises as to which one is the best, if any. Heck, Manfroid & Mersch (1985) used numerical simulations to compare discrete Fourier transforms, STR and PDMs and found that none of them was superior to the others. Schwarzenberg-Czerny (1999) compared model function and phase binning methods using hypothesis-testing theory to evaluate their relative sensitivity to different kinds of signals. He found that the methods using smooth model functions, such as Lomb–Scargle (LS), are more sensitive than those using the step function, i.e. phase binning, and that sensitivity increases for models that more closely fit features in the signal with the orthogonal multiharmonic AOV method (AOVMHW; Schwarzenberg-Czerny 1996) being optimal.

He also found that a number of methods relying on phase binning are equivalent for the same number of bins. Similarly, Swingler (1989) argues that PDM methods should be regarded simply as approximations to the Fourier method (LS) and that the latter should be viewed as the periodogram technique of choice. Distefano et al. (2012) have compared discrete Fourier, LS and PDM techniques for recovering the rotation periods of solar-like stars from irregular time sampling of *Gaia* using synthetic time series. They find that LS is the most efficient method with at best a recovery rate of ∼60 per cent.

Dubath et al. (2011) report that a single method can lead to a recovery fraction of around 80 per cent but do not specify to what degree of accuracy. They also suggest that an ideal combination of all methods could potentially raise that value to close to 100 per cent but cannot identify the automated strategy for predicting which method leads to the correct period for a specific light curve. However, they propose, as a first step, a combination of unweighted and weighted LS, depending on the skewness of a source's magnitude distribution.

In this work, we present a detailed comparison of the most commonly used period finding algorithms and their efficiencies against observable parameters. This is the first survey using real rather than simulated data (so with noise, gaps, etc.) to consider both a wide range of variable stellar classes and light curves generated by different sampling strategies. It is hoped that we can identify the most effective algorithm with a particular view to the next generation of survey projects which require automated and efficient period finding methods.

The paper is structured as follows: in Section 2, we describe the data sets that form the basis of our analysis whilst in Section 3, we present the algorithms that we are considering here. We analyse and discuss our results in Sections 4 and 5 and present our conclusions in Section 6. We also have provided implementation details about Open Computing Language (OpenCL)-based versions of the LS and generalized Lomb–Scargle (GLS) algorithms in an Appendix.

2 DATA SETS

In this analysis, we consider three sets of light curves, drawn from different surveys – CRTS, ASAS, MACHO, which we take to be representative of the bulk of ground-based light-curve data sets currently available and characteristic of future large samples such as LSST. Together these span a magnitude range of $\sim 4 \leq V \leq 21$ (see Fig. 1) and a sampling of up to ∼1800 observations (see Fig. 2) over a baseline of up to ∼8 yr. Details of the three are summarized in Table 1. All observation times are converted to Heliocentric Julian Date from MJD.

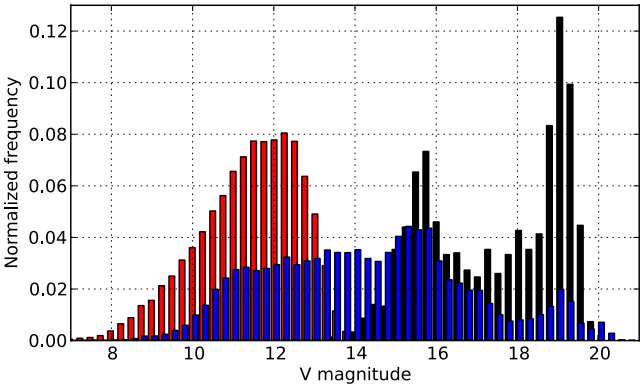

Figure 1. This shows the *V*-band magnitude distribution of the three data sets considered in this paper: ASAS (red), MACHO (black) and CRTS (blue).

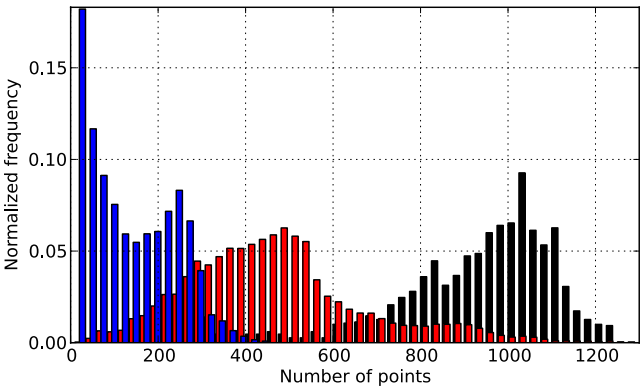

Figure 2. This shows the distribution of observations per light curve for the three data sets considered in this paper: ASAS (red), MACHO (black) and CRTS (blue).

Table 1. This summarizes the three data sets used in this analysis.

Data set	No. of sources	Median values			
		Magnitude	Observations	Baseline (d)	Period (d)
ASAS	50 124	11.59	456	2645	44.51
CRTS	15 522	14.35	105	2182	0.59
MACHO	1500	17.82	966	2721	1.74

2.1 Catalina real-time transient survey

The CRTS (Drake et al. 2009) is the largest open time domain survey currently operating, covering ∼33 000 deg^2 between $-75^\circ <$ Dec. $< 75^\circ$ (except for within ∼10–15° of the Galactic plane). It leverages the data streams from three telescopes used in a search for near-Earth objects, operated by the Lunar and Planetary Laboratory at University of Arizona, with four exposures per visit, separated by 10 min, reaching to $V \sim 19$ to 21.5 mag (depending on telescope), over 21 nights per lunation. All data are automatically processed in real time, and optical transients are immediately distributed using a variety of electronic mechanisms.[3] Light curves of several hundred million objects are available[4] with an average of ∼250 observations over an 8 yr baseline.

[3] http://www.skyalert.org
[4] http://crts.caltech.edu

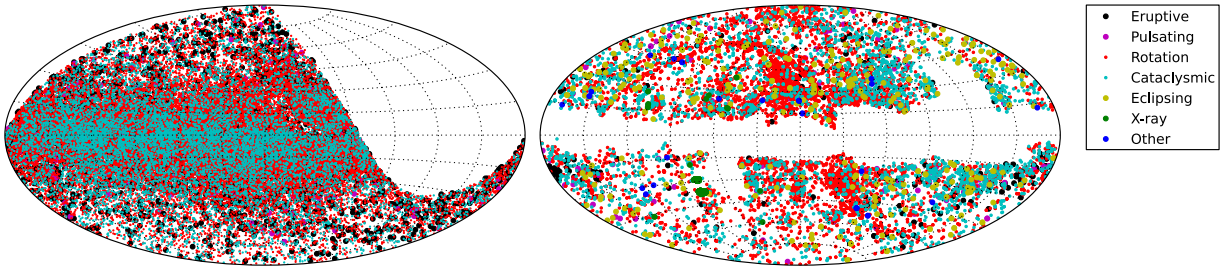

Figure 3. This shows the sky distribution in galactic coordinates of ASAS (left) and CRTS (right) sources. MACHO sources are localized to the LMC and not shown. The sources are colour coded according to their broad variable class: eruptive (black), pulsating (red), rotation (maroon), cataclysmic (yellow), eclipsing (cyan), X-ray (blue) and other (green). Larger symbols are employed for the less populous classes.

To get a sample of as wide a range of variable sources as possible, all objects in SIMBAD and the AAVSO Variable Star Index (VSX; Watson 2006) with a recorded period were selected, giving 146 655 sources (71 per cent of the total combined number). Light curves were then extracted for those which had been observed by CRTS. Since very many of the initial data set lie in the galactic plane and CRTS explicitly avoids the galactic plane, this brought the number of sources covered to a manageable 15 522 (see Fig. 3 for the distribution of the sources on the sky). The poorer sampling near the plane also explains why the median number of observations for this data set is 105.

All the light curves were inspected by three of us to verify (via phased light curves) the quoted period against fit periods from two other methods: AOV and conditional entropy (CE, Graham et al. 2013). When there was no consensus opinion, the quoted period was used.

2.2 ASAS

The ASAS Catalog of Variable Stars (ACVS; Pojmanski, Pilecki & Szczygiel 2005) represents one of the largest collections of light curves of variable stars available, covering a range of 11 science classes (albeit predominantly MISC) and with good consistent data. The sky distribution with the limit $\delta < +28$ is shown in Fig. 3. Richards et al. (2012) (MACC) have applied probabilistic classifiers to ACVS light curves to create a 28-class machine-learned catalogue of 50 124 sources. We have followed a similar prescription to MACC to construct our data set of ACVS light curves: the data for individual objects are retrieved from the ACVS website (ACVS 1.1[5]) and those epochs with a quality `GRADE=D` or quality `GRADE=C` when `MAG=29.999` excluded, corresponding to a non-detection. This gives a median of 456 usable epochs of *V*-band observations covering 2644.92 d.

ASAS provides five aperture measurements using diameters ranging from 2 pixels (30 arcsec) to 6 pixels (90 arcsec) and describes a basic algorithm for choosing which aperture to use for an object given its average magnitude (Pojmanski et al. 2005). MACC constructed a simple classifier to determine the optimal aperture to use for each object which we have followed: using 2 pixels for $V > 12.25$, 3 pixels for $11.675 < V < 12.25$, 4 pixels for $10.675 < V < 11.675$ and 6 pixels for $V < 10.675$.

ACVS periods were determined using an AOV algorithm and confirmed visually. MACC has also determined a period for each object using a GLS-based algorithm (Zechmeister & Kürster 2009) with corrections for eclipsing and aliased periods (see Section 3.2). The quoted agreement between the two is 77.2 per cent (exactly matching) for the 12 008 objects which ACVS confidently classified into a single periodic class (see Section 4.1 for a discussion of this). We have inspected a representative sample of the light curves and consider the ACVS period to be the true value.

2.3 MACHO

The MACHO survey (Alcock et al. 2003) was designed to search for gravitational microlensing events in the Magellanic Clouds and the Galactic Bulge and more than 20 million stars were observed, making it an important resource for variable star studies. A 'gold standard' data set of light curves has been produced from the MACHO survey by the Harvard Time Series Center,[6] consisting of approximately 500 each of RR Lyrae, eclipsing binaries and Cepheids, respectively covering the Large Magellanic Cloud (LMC) ($75^\circ < \mathrm{RA} < 85^\circ$, $-71^\circ < \mathrm{Dec.} < -67^\circ$). Although MACHO data normally consists of blue and red channel data for each stellar object, only the blue channel (*V*-band equivalent) have been used here. The median time span of the data is 2720.88 d. This data set has also been used in two correntropy-based (generalized correlation) approaches for estimating periods in non-uniformly sampled time series: Mishra et al. (2011) employs slots (intervals) (Mishra et al. 2011) to determine the statistic of interest whilst Huijse et al. (2012) uses a kernel.

2.4 Variable classes

Objects in the CRTS data set have been labelled with classes drawn from VSX[7] which is itself based on the General Catalogue of Variable Stars (GCVS; Samus et al. 2009) classification scheme (a maximum cross-match distance of 3 arcsec was used). This is a relatively detailed system that covers most types of variable stellar phenomena. Objects fall into one of seven broad classes reflecting both extrinsic and intrinsic phenomena: eruptive, pulsating, rotating, cataclysmic, eclipsing, X-ray and other. For convenience, we have converted the VSX codes into a hierarchical coding scheme: for example, the eclipsing class is P.5, an eclipsing binary is P.5.1 and a β Lyrae-type eclipsing binary is P.5.1.2.

For ASAS data, we use the MACC classifications from ASAS_CATALOG_CLASS_V3.0.[8] MACC employs a 28-term scheme taken from Debosscher et al. (2007) with the addition of SX Phe and splitting T Tauri into two subclasses: classic T Tauri and

[5] http://www.astrouw.edu.pl/asas/

[6] http://timemachine.iic.harvard.edu

[7] http://www.aavso.org/vsx/help/VariableStarTypeDesignationsInVSX.pdf

[8] http://www.bigmacc.info/

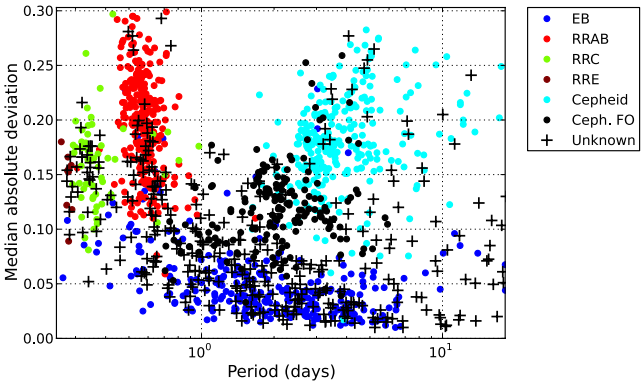

Figure 4. This shows the distribution of MACHO light curves in the MAD (from the median) – period plane. The different colours denote different MACHO classes of object: blue (EB), red (RRAB), cyan (Cepheid fundamental), green (RRC), purple (RRE), black (Cepheid first overtone). The crosses indicate objects for which there is no MACHO classification in the literature and one must be imputed by a nearest neighbour classifier.

weak-line T Tauri. Six of these classes do not have an equivalent in the VSX/GCVS scheme and so we add them to ours – they are: long secondary period red giants, small amplitude red giants split according to Wood et al. (1999) (SARG_A, SARG_B), red supergiants (RSG), chemically peculiar (ChemPec), and Herbig AE/BE (HAEBE). Note that only 8572 out of the 50 124 ASAS sources have a MACC classification with a probability of 90 per cent or higher.

The MACHO data initially consisted of just three classes: RR Lyrae, eclipsing binaries and Cepheids. However, additional data for these objects (Alcock et al. 2003) gave finer-grained classifications for 1139 stars based on an automated statistical analysis of its photometry over time. A plot of the median absolute deviation (MAD) from the median of the light curves of these objects versus their quoted period can be useful for discriminating between different classes (see Fig. 4). A nearest-neighbour classifier in the MAD-period plane was then used to impute classes to the remaining 361 objects without finer-grained classifications. These were checked with SIMBAD: of the 305 objects with a SIMBAD classification but not a fine-grained one, ∼91 per cent agreed with their imputed class. This is the same level of accuracy as the MACC classifications.

Table 2 gives the relative numbers of objects per class in the three data sets considered in this paper. The distribution of periods over the three data sets is shown in Fig. 5. As mentioned above, these have all been visually confirmed (either by us or other authors) and so this is the true distribution with no contamination from aliased periods. The peaks at log(period) ∼−0.4 in the three distributions are from RR Lyrae and eclipsing binary objects. Similarly, the peak at log(period) ∼1.5 in the ASAS data is from small amplitude red giants, the peak at log(period) ∼0.5 in the MACHO data from Cepheids, and the peak at log(period) ∼2.5 in the CRTS data from Mira variables, respectively.

3 ALGORITHMS

Period finding algorithms can be divided into a number of types. The most popular seek to model a light curve via a least-squares fit to some set of (orthogonal) basis functions, most commonly trigonometric, such as LS (Lomb 1976; Scargle 1982) and its derivatives/extensions (e.g. Zechmeister & Kürster 2009), though more complicated function sets, such as wavelets (Foster 1996), have also been tried. Another approach is to minimize some measure of the dispersion of time series data in phase space, such as binned means (Stellingwerf 1978), variance (Schwarzenberg-Czerny 1989), total distance between points (Dworetsky 1983) or entropy (Cincotta, Mendez & Nunez 1995), which can often be regarded as an expansion in terms of periodic orthogonal step functions. Bayesian methods (Gregory & Loredo 1992, Wang, Khardon & Protopapas 2012) are also becoming common and there have even been attempts to search for periodicity using neural networks (Baluev 2012).

The basis of an algorithm also often determines how well it copes with the real world aspects of time series data, such as irregular sampling, gaps and errors, e.g. standard Fourier analysis is impossible for any data diverging from regular sampling. de Jager, Raubenheimer & Swanepoel (1989) argue that in the case of weak signals, most period finding methods only work well with certain kinds of periodic shapes and that this causes a selection effect for the general identification of weak periodic signals. Similar shape dependences are found in Schwarzenberg-Czerny (1999).

For this analysis, we have selected a representative set of the most common algorithms used which claim to be fast and accurate (see Table 3 for operational details). This is a necessary condition for automated large-scale analyses of time series – we consider an algorithm to be fast if it returns an answer in less than 5 s (assuming ∼250 points for a light curve and an ∼2 GHz CPU – see Section 5.8. This is a fairly conservative definition but roughly equates to analysing 1000 light curves in just under 1.5 h on a single processor. There are methods which can attain very high degrees of accuracy but do so at the expense of taking up to several minutes to work on a single time series, e.g. by using a very fine grain resolution in searching frequency space or involving a multistage process, such as SuperSmoother (Reimann 1994), Jetsu & Pelt (1999) or Shin & Byun (2004). These are well suited for small-scale detailed analyses but not for the bulk processing that the new synoptic sky surveys warrant. However, for the sake of comparison, we have included SuperSmoother results for the MACHO data set, since it is largely regarded as the most accurate period finding technique.

Intuitively the fastest period finding algorithm will involve a single pass through a data set per trial period and integer counting operations, e.g. histogram binning. Any higher order function calls, particularly per data point in a time series, will extend the average calculation time per trial period and, consequently, the overall time taken by the algorithm to determine a correct period. Of the algorithms considered, CE, AOV and PDM come closest to this ideal with a basic implementation employing integer arithmetic. AOV then requires two passes through a data set per trial period – one to compute the mean/variance of each bin, $\overline{x}_i$, and one to subtract the appropriate mean value from each data point, $x_{ij} - \overline{x}_i$. PDM is similar but CE only requires one pass. Note that a single pass version of AOV has been defined (Schwarzenberg-Czerny & Beaulieu 2006)

One particular issue for automated period finders (particularly LS) is that they misidentify a multiple (or submultiple) of the period as the 'true' period, i.e. the identified period, $p_i = mp_0$, where m is an integer n or its reciprocal, $1/n$, and p_0 is the correct period. This is a common problem for binary systems where the half period is frequently the most significant peak in a periodogram. For example, Richards et al. (2012) initially find 70 per cent of their periods for eclipsing binaries (EBs; ∼49 per cent of all objects) in the ACVS (Pojmanski et al. 2005) to be half periods. As discussed in Wang et al. (2012), this is attributable to two aspects: for symmetric EBs, the true period and half its value are not clearly distinguishable quantitatively. Meanwhile, methods that are successful for EBs tend

Table 2. The relative numbers of each class of variable stellar object used in this analysis. Only those classes which have instances are included. The codes in parentheses are the GCVS or VSX code for this type of variable or, if marked with an asterisk, the code used in MACC. The method given is the most reliable method for finding periods for this class (see Section 5.6 for details) with an asterisk indicating that less than 10 periods were recovered by it. A dash denotes that no method recovered an accurate period for the class. When two methods are given, they both recovered the same number of periods accurately.

Class	Label	CRTS	MACC	MACHO	Method
Eruptive					
Be variable	P.1.2.1 (BE)	–	337	–	GLS
Poorly studied irreg. var. of inter. to late spectral type	P.1.3.2 (IB)	43	–	–	AOVMHW*
Orion variable	P.1.3.3 (IN)	5	–	–	LS*
Orion variable - early spectral type	P.1.3.3.1 (INA)	85	–	–	GLS*
T Tauri	P.1.3.3.3 (INT, IT)	28	5	–	AOVMHW*
Weak-line T Tauri	P.1.3.3.3.2 (WTTS)	–	2239	–	GLS
Rapid Orion variable of intemediate to late spectral type	P.1.3.3.6 (INSB)	2	–	–	–
Rapid T Tauri	P.1.3.3.7 (INST)	5	–	–	AOV*
Rapid irregular variable without nebula	P.1.3.4 (IS)	3	–	–	AOVMHW*
R Cor Bor type variable	P.1.4 (RCB)	–	53	–	GLS*
RS Can Ven type variable	P.1.5 (RS)	178	263	–	AOVMHW
S Dor type variable	P.1.6 (SDOR)	–	1	–	–
UV Ceti type variable	P.1.7 (UV)	4	–	–	–
Flaring Orion variable of spectral type Ke - Me	P.1.7.1 (UVN)	1	–	–	–
Young Stellar Object	P.1.9 (YSO)	4	–	–	–
Herbig AE/BE	P.1.10 (HAEBE*)	–	111	–	CE*
Red supergiant	P.1.11 (RSG*)	–	827	–	GLS
Pulsating					
General pulsating variable	P.2 (PULS)	58	–	–	GLS
Beta Cephei type variable	P.2.2 (BCEP)	3	259	–	AOVMHW
Cepheid	P.2.3 (CEP)	25	568	287	CE
Multimode Cepheid	P.2.3.1 (CEP(B))		202	230	CE
W Vir type variable	P.2.4 (CW)	2	–	–	AOV
W Vir type variable with period longer than 8 d	P.2.4.1 (CWA)	26	–	–	CE
W Vir type variable with period shorter than 8 d	P.2.4.2 (CWB)	30	–	–	AOVMHW
Classical Cepheid (Delta Cep)	P.2.5 (DCEP)	53	–	–	AOVMHW*
D Cep type variable with light amp. and symmetrical LC	P.2.5.1 (DCEPS)	2	–	–	CE*
Delta Scuti type variable	P.2.6 (DSCT)	92	1527	–	FC
Low amplitude Delta Scuti type variable	P.2.6.1 (DSCTC)	5	–	–	AOVMHW*
High amplitude Delta Scuti type variable	P.2.6.2 (HADS)	95	–	–	LS
Slow irregular variable of late spectral type	P.2.7.1 (LB)	7	–	–	AOVMHW/FC*
Mira type variable	P.2.8 (M)	979	3086	–	GLS
RR Lyrae type variable	P.2.10 (RR)	1957	–	11	CE
RR Lyrae type variable with fundamental mode	P.2.10.1 (RRAB)	4518	1460	343	CE
RR Lyrae type variable with fundamental overtone	P.2.10.3 (RRC)	692	476	91	AOVMHW
RR Lyrae type variable with double mode	P.2.10.4 (RRD)	137	130	15	GLS
RV Tauri type variable	P.2.11 (RV)	7	452	–	CE
RV Tauri type variable that does not vary in mean mag.	P.2.11.1 (RVA)	18	–	–	AOVMHW*
RV Tauri type variable that varies periodically in mean mag.	P.2.11.2 (RVB)	1	–	–	STR*
Semiregular variable	P.2.12 (SR)	436	9982	–	GLS
Semiregular late-type giant with persistent periodicity	P.2.12.1 (SRA)	142	–	–	AOVMHW
Semiregular late-type giant with poorly defined periodicity	P.2.12.2 (SRB)	168	–	–	AOVMHW
Semiregular late-type supergiant	P.2.12.3 (SRC)	1	–	–	–
Semiregular variable giant/supergiant	P.2.12.4 (SRD)	16	–	–	CE*
Semiregular variable	P.2.12.5 (SRS)	2	–	–	–
SX Phe type variable	P.2.13 (SPXHE)	35	12	–	FC
ZZ Ceti type variable	P.2.14 (ZZ)	2	–	–	–
Anomalous Cepheid type variable	P.2.17 (BLBOO)	13	–	–	AOVMHW*
G Dor type variable	P.2.18 (GDOR)	2	–	–	CE*
Small amplitude red giants - type A	P.2.19.1 (SARG_A*)	–	3974	–	GLS
Small amplitude red giants - type B	P.2.19.2 (SARG_B*)	–	7820	–	GLS
Long secondary period red giants	P.2.20 (LSP*)	–	5096	–	GLS
Rotating					
Alpha2 Can Ven type variable	P.3.1 (ACV)	1	–	–	AOVMHW*
BY Draconis type variable	P.3.2 (BY)	89	–	–	AOVMHW
Ellipsoidal variable	P.3.3 (ELL)	18	2	–	AOVMHW*
Chemically peculiar	P.3.7 (ChemPec*)	-	345	-	GLS

Table 2 – *continued*

Class	Label	CRTS	MACC	MACHO	Method
Cataclysmic					
Cataclysmic variable	P.4 (CV)	17	–	–	CE*
Fast novae	P.4.1.1 (NA)	4	–	–	–
Slow novae	P.4.1.2 (NB)	6	–	–	PDM*
Novalike variable	P.4.1.4 (NL)	57	–	–	AOVMHW*
Recurrent novae	P.4.1.5 (NR)	5	–	–	AOVMHW*
U Gem type variable	P.4.3 (UG)	88	–	–	PDM2*
SS Cyg type variable	P.4.3.1 (UGSS)	21	–	–	PDM/CE*
SU U Ma type variable	P.4.3.2 (UGSU)	140	–	–	AOVMHW
Z Cam type variable	P.4.3.3 (UGZ)	22	–	–	AOVMHW*
WZ Sag type variable	P.4.3.4 (UGWZ)	19	–	–	–
Z Andr type variable	P.4.4 (ZAND)	1	–	–	–
	P.4.5 ()	6			
Eclipsing					
Eclipsing binary system	P.5.1 (E)	109	–	522	CE
Beta Persei (Algol) type system	P.5.1.1 (EA)	1025	2855	–	STR
Beta Lyrae type system	P.5.1.2 (EB)	866	1963	–	CE
W U Ma type system	P.5.1.3 (EW)	1479	6025	–	AOVMHW
Contact system	P.5.8 (K)	20	–	–	CE
Semidetached system	P.5.9 (SD)	1	–	–	STR*
AM Her type system	P.5.10 (AM)	76	–	–	CE*
Close binary system with strong reflection	P.5.11 (R)	9	–	–	CE*
Planet eclipsing system	P.5.12 (EP)	2	–	–	–
X-ray					
DQ Her variable type / low-mass X-ray binary	P.6.1.8 (DQ, LMXB)	30	–	–	AOVMHW*
Close-binary super-soft source	P.6.2 (CBSS)	1	–	–	LS*
Other					
Variable	P.7	1433	–	–	LS

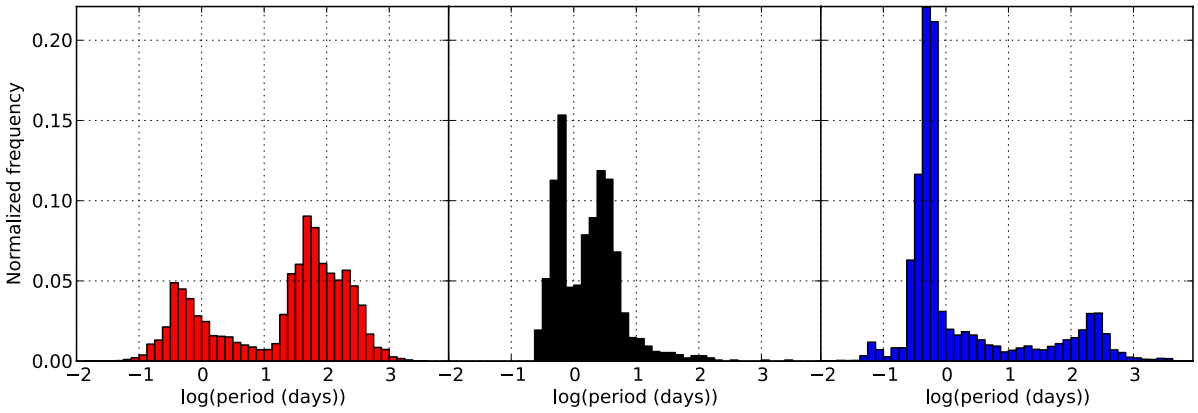

Figure 5. This shows the distribution of quoted periods in days for the three data sets considered in this paper: ASAS (red), MACHO (black) and CRTS (blue).

to find integer multiple periods of 'single bump' stellar types, such as RR Lyrae and Cepheids, and vice versa. EBs also have two minima per cycle, while only one is expected by methods looking for sinusoidal-like variations. Clearly using a period (sub)harmonic instead of the true value can be a problem for period-based statistics, such as Fourier decomposition where particular components would be assigned the wrong weights (amplitudes). We will consider the issue of period harmonics further in Section 5.

3.1 Frequency sampling

The frequency sampling strategy used with a period finding algorithm is important. Vio, Diaz-Trigo & Andreani (2013) show that irregular sampling reduces the width of the peak at the correct frequency in the LS periodogram of a light curve if its temporal baseline is large. This means that there is a concrete risk of missing the peak if the periodogram is not computed for a sufficient large number of test frequencies. However, it also means that the error on the computed period will be less since this is also dependent on the width of the associated peak in the periodogram.

For a regularly sampled time series with time spacing, Δt, the Nyquist frequency, $\nu_{\rm N} = 1/2\Delta t$, constitutes an upper limit to the frequency range over which a periodogram can be uniquely calculated. For irregularly sampled time series, however, this value can be much higher – Koen (2006) gives an upper limit of 0.5Δ for this frequency, where Δ is the best accuracy with which time is recorded.

Table 3. Details of the various period finding algorithms used in this analysis. Where possible, we have used provided code, e.g. AOV/AOVMHW, PDM2, FastChi, and default parameter settings. The asterisk denotes those algorithms which were only applied to the MACHO data set.

Algorithm	Implementation	Behaviour	Reference
Lomb–Scargle (LS)	OpenCL[a]	$\mathcal{O}(n^2)$	Lomb (1976); Scargle (1982); Townsend (2010)
Generalized Lomb–Scargle (GLS)	OpenCL[a]	$\mathcal{O}(n^2)$	Zechmeister & Kürster (2009)
Binned analysis of variance (AOV)[b, c]	F95	$\mathcal{O}(nN)$	Schwarzenberg-Czerny (1989)
Multiharmonic analysis of variance (AOVMHW)[b, d]	F95	$\mathcal{O}(nN)$	Schwarzenberg-Czerny (1996)
Phase dispersion minimization (PDM)[e]	F90	$\mathcal{O}(nN)$	Stellingwerf (1978)
Phase dispersion minimization (PDM2)[f, g]	C	$\mathcal{O}(nN)$	Stellingwerf (2011)
FastChi (FC)[h, i]	C	$\mathcal{O}(N \log N)$	Palmer (2009)
String length (STR)	F90	$\mathcal{O}(nN)$	Dworetsky (1983)
Conditional entropy (CE)[j]	F90	$\mathcal{O}(nN)$	Graham et al. (2013)
Supersmoother (SS)*[k]	C	$\mathcal{O}(nN)$	Reimann (1994)
Correntropy kernel periodogram (CKP)*[l]	C	$\mathcal{O}(n^2)$	Huijse et al. (2012)

[a]see Appendix ; [b]http://users.camk.edu.pl/alex/soft/aovdist.tgz; [c]overlapping phase bins; [d]5 harmonics; [e]10 phase bins; [f]http://www.stellingwerf.com/rfs-bin/index.cgi?action=PageView&id=34; [g]Stellingwerf's improved algorithm; [h]http://public.lanl.gov/palmer/fastchi.html; [i]harmonics = 3, oversampling = 4; [j]10 overlapping phase bins, 5 magnitude bins; [k]code obtained from Andy Becker; [l]Code from Pablo Huijse.

For example, in CRTS time is recorded to five decimal places giving $\nu_{\rm N} = 5 \times 10^{-4}{\rm d}^{-1}$; in practice, though, a more manageable lower value would be used.

Two common frequency gridding strategies are applied in the literature when working with large collections of time series. The first (Debosscher et al. 2007; Richards et al. 2012) uses for all light curves: $\nu_{\rm min} = 0$, $\nu_{\rm max} = 10$ and $\delta\nu = 0.1/\Delta\tau$, where $\Delta\tau$ is the data timespan. The second (Schwarzenberg-Czerny 1996) estimates optimal values for each light curve: $\nu_{\rm min} = 0$ or $\delta\nu$, $\nu_{\rm max} = 1/2\tau_{med}$ and $\delta\nu = 1/(A \times \Delta\tau)$, where $\tau_{\rm med}$ is the median difference between successive ordered times, A is a factor, typically 10–15, taking into account oversampling and binning or the number of harmonics used in a Fourier fit, and $\Delta\tau$ is the data timespan.

We have applied both the optimal strategy and fixed $\delta\nu$ values of $\delta\nu = 0.0001$, 0.001 and 0.01 over a frequency range with $\nu_{\rm min} = 0$ and $\nu_{\rm max} = 20$. The median timespan for all the data is $\sim$2618 d which gives a median $\delta\nu = 2.5 \times 10^{-5}{\rm d}^{-1}$, assuming $A = 15$ in the optimal case. We also note that many algorithms use a finer resolution grid to get a more accurate period estimate once a primary peak has been found with the coarse grid.

4 METRICS

For the purposes of this analysis, we define three metrics: one to evaluate the accuracy of the recovered period of a light curve compared to the value of its true period and two to measure the quality of a light curve.

4.1 Accuracy metric

Oluseyi et al. (2012) define a matching criterion for period recovery using the quality of the period-folded data as a metric:

$$\frac{|P_{\rm al} - P_{\rm in}|}{P_{\rm in}} \leq \frac{\delta\phi_{\rm max} P_{\rm in}}{\Delta\tau},$$

where $P_{\rm in}$ is the known input period, $P_{\rm al}$ is the period according the algorithm under investigation, $\Delta\tau$ is the duration of the time series and $\delta\phi_{\rm max}$ is the maximum allowed phase offset after period-folding N cycles. For simulated RRab light curves in LSST, this translates to

$$\frac{|P_{\rm al} - P_{\rm in}|}{P_{\rm in}^2} \leq 10^{-5}{\rm d}^{-1}$$

for a maximum period-folded phase offset of 1/27th of a cycle or a period within $\sim$0.22 s of the true value for a 0.5d period star and a 10 yr survey. Given the variation in the baselines of the light curves in this analysis, particularly within the CRTS data set where there is a dependence on both galactic latitude and which telescope was used to observe the object, a fixed survey length makes little sense. Instead for each object, we consider its temporal coverage but keep $\delta\phi_{\rm max} = 1/27$ so that a 10 yr baseline will give equivalent accuracy to LSST. We will use the equivalent accuracy level of 10^{-5} d^{-1} as a fiducial value. For the median baselines of the three surveys, the corresponding accuracy values are 1.4×10^{-5} d^{-1} for ASAS and MACHO and 1.7×10^{-5} d^{-1} for CRTS.

Dubath et al. (2011) consider a period as good if the difference between the calculated period (using LS/GLS) and the quoted value leads to a cumulative shift in phase of less than 20 per cent over the full timespan of the light curve. This equates to an accuracy level of $\sim 10^{-4}$ d^{-1} for a 10 yr baseline. Meanwhile, Richards et al. (2012) claimed 77.2 per cent exact agreement between the periods they found for ASAS objects (using a GLS-based algorithm) and those given by ACVS. However, in terms of the matching criterion used here, only 20.2 per cent of the periods actually agree between the two sets at the 10^{-5} d^{-1} accuracy level. The quoted agreement of Richards et al. is found at an accuracy level of $\sim 10^{-3}$ d^{-1}.

We have therefore considered equivalent accuracy cutoffs for a 10-yr baseline of 10^{-3} ($\delta\phi_{\rm max} = 100/27$), 10^{-4} ($\delta\phi_{\rm max} = 10/27$) and 10^{-5} d^{-1} ($\delta\phi_{\rm max} = 1/27$), respectively for our comparisons to reflect the range used in the literature and, for the value of 10^{-5}, with a view to future surveys. We note that the error in the determined period could be larger than a particular accuracy cutoff. However, as already noted, most of the algorithms in this analysis use a finer grain resolution to get a more accurate estimate once an initial value has been found with a coarser grain resolution. This is typically a factor of a hundred smaller than the coarse grain resolution step and in this analysis would be a maximum of $\Delta\nu = 10^{-5}$. The effect

of this should therefore be minimal with reference to a particular cutoff value.

We also want to have an accuracy metric relevant for period harmonics since periodicity in an object can still be detected, even if only a harmonic of the true period is found (Huijse et al. 2012). We modify the criteria used by Huijse et al. (2012) so that an accurate harmonic is identified if

$$\left|\frac{P_{\rm al}}{P_{\rm in}} - \left\|\frac{P_{\rm al}}{P_{\rm in}}\right\|\right| < \frac{\delta\phi_{\rm max} P_{\rm in}}{\Delta\tau} \text{ for } P_{\rm al} > P_{\rm in}$$

and

$$\left|\frac{P_{\rm in}}{P_{\rm al}} - \left\|\frac{P_{\rm in}}{P_{\rm al}}\right\|\right| < \frac{\delta\phi_{\rm max} P_{\rm in}}{\Delta\tau} \text{ for } P_{\rm al} < P_{\rm in},$$

where $\|x\|$ is the nearest integer to x relative to the same accuracy cutoff used for periods.

4.2 Quality measure

There are several sets of light curves of the same class of variable object that we would like to compare on a common quality basis but they have been produced by different telescopes and so span different magnitude ranges; for example, an RR Lyrae light curve in the ASAS data set has the same subjective quality, i.e. visually appears the same in terms of error size and scatter, at 12th magnitude as a 20th magnitude light curve in the MACHO data set.

The signal-to-noise ratio (SNR; mean of signal/standard deviation of noise) provides a general matching criterion. Rimoldini (2013) gives an expression for the SNR of a light curve and we employ a slightly modified form here:

$$SNR = \left[\frac{\sum_{i=1}^{n} w_i (x_i - x_m)^2 + \sum_{i=1}^{n} w_i^2 \epsilon_i^2 / W}{\sum_{i=1}^{n} w_i \epsilon_i^2}\right]^{1/2},$$

where x_m is the median magnitude (instead of the mean), ϵ_i the photometric error of the $i_{\rm th}$ data value and $W = \sum_i w_i$. We employ $w_i = 1$ in this analysis. Fig 6 shows the distribution of SNR values for the three surveys considered here.

We note, though, that this measure is based on mean quantities and that changes in the overall shape of a light curve for different object types will have an impact: for example, there is a strong correlation between the SNR of a light curve and the amplitude of its variability and this is also survey dependent (see Fig. 7). Since SNR is essentially a measure of the intrinsic scatter within a data

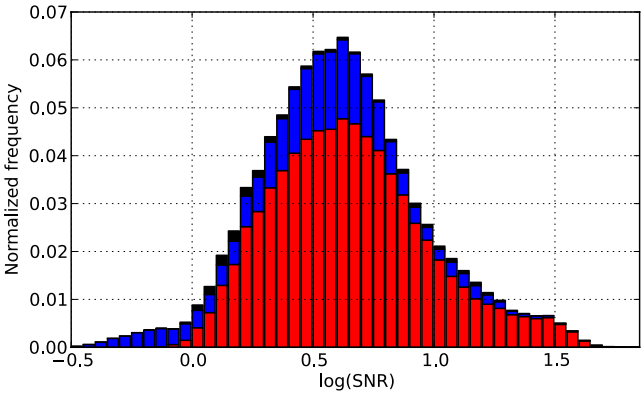

Figure 6. This shows the overall distribution of the SNR for all the light curves and the stacked relative contributions of each of the individual data sets: ASAS (red), CRTS (blue) and MACHO (black).

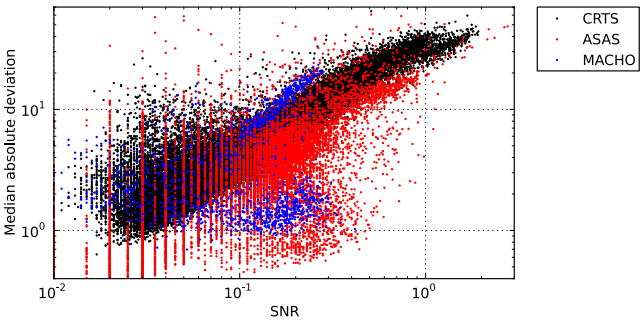

Figure 7. This shows the distributions of the SNR versus the MAD (from the median) for the three surveys: ASAS (red), CRTS (blue) and MACHO (black). There is clearly a strong correlation between SNR and the amplitude of variability and this is a survey-dependent effect – the same SNR value equates to a different range of variability for each survey.

set, it is conceptually similar to the entropy. Standard estimators for entropy, though, are optimized for signal detection and do not take into account the contributions of noise. Cincotta (1999) defines a modified estimator for the Shannon entropy of a data set which takes observational errors into account and we will use this as class-based quality comparisons. The Shannon entropy, H_0, for a distribution on the unit square partitioned into k partitions is:

$$H_0 = -\sum_{i=1}^{k} \mu_i \ln(\mu_i); \forall \mu_i \neq 0,$$

where μ_i is the occupation probability for the $i^{\rm th}$ partition. For a data set where measurement v_i has an error ϵ_i, the occupation probability is given by

$$\widetilde{\mu} = \frac{R}{2} \sum_{i=1}^{n_l} [erf(w_{\rm im} + \Delta w_i) - erf(w_{\rm im})],$$

where erf(x) is the error function, $n_l \simeq N/L$ is the number of points in the $l^{\rm th}$ partition of the x-axis (i.e. with phase in the $[l/L, (l+1)/L]$ interval), and

$$w_{\rm im} = \frac{m - Mv_i}{\sqrt{2}M\epsilon_i}, \Delta w_i = \frac{1}{\sqrt{2}M\epsilon_i},$$

$$\frac{2}{R} = \sum_{i=1}^{N} [erf(w_{iM}) + erf(|w_{i0}|)]$$

and L and M are the number of partitions along the x-axis and y-axis, respectively.

Fig 8 shows the distribution of the modified entropy for all the light curves, phased at their quoted periods, in the CRTS, ASAS and MACHO data sets against their median photometric errors. This shows that there is a general relationship between the two quantities and therefore light curves with the same modified entropy can be considered to be qualitatively similar (in broad SNR terms) and therefore compared on an equal footing. Note that the computational cost of the estimator – here involving error function calculations – is too high for consideration as a viable period finding algorithm in this paper.

5 RESULTS AND ANALYSIS

The efficacy of the different period finding algorithms is clearly dependent on a number of factors. We evaluated each of the algorithms

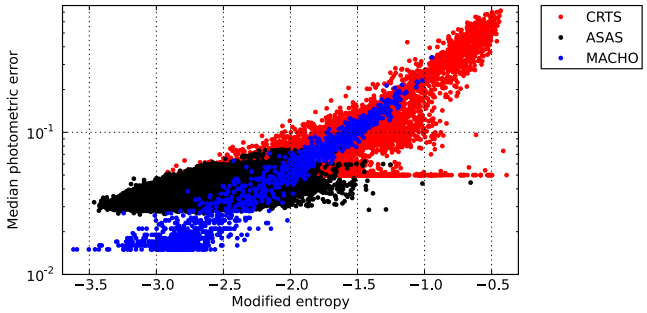

Figure 8. This shows the distribution of the median photometric error and the modified entropy for all the light curves phased at their quoted period considered in this paper. The three surveys are denoted by red (CRTS), blue (MACHO) and black (ASAS). The artefact at mean error = 0.05 in the CRTS data set results from a lower limit to error size in this data set. The small median errors at the highest entropy level may indicate non-monoperiodic (multiperiodic, irregular, etc.) sources.

in terms of completeness, i.e. the fraction of true periods recovered within a defined accuracy limit, as a function of various quantities. The two most obvious variables to consider are magnitude and the number of observations in a light curve. The resolution used to scan the range of trial periods (frequencies) will also have an effect. The variability of a light curve, both naturally due to the actual variability of the source and acquired as measurement scatter (noise), as well as the actual class of variable object (and the shape of the light curve) and its period are also factors to evaluate. Finally, the actual time taken to determine an accurate period can be an important aspect in determining the *usability* of particular algorithms in addition to their accuracy.

5.1 Magnitude

Fig. 9 shows the completeness fraction as a function of magnitude for the three data sets and accuracy cutoffs, respectively. With the MACHO and CRTS data, there is a general decline in accuracy with increasing magnitude, particularly past the 90th percentile in the magnitude distribution, as the photometric errors become more significant and the light curves noisier. There are also dips in both data sets around the 60th percentile which is most likely connected to the relative magnitude distributions of different classes of object; for example, in the MACHO data, ∼90 per cent of the objects between 14th and 16th magnitude are Cepheids, whereas between 16th and 18th magnitude, there are twice as many eclipsing binaries as Cepheids. The ASAS data seems fairly flat except at its very faintest end. With this data, magnitude is weakly correlated with SNR, i.e. fainter objects at fainter magnitudes tend to have slightly better quality light curves and the methods are more likely to recover the true period for an object with a higher SNR (see Section 5.3). The combination of the two in this case gives a fairly constant relationship between completeness and magnitude. The low levels of completeness relative to the other two data sets are a consequence of the large number of semiregular variables and similar pulsating objects in this data set (∼50 per cent) for which accurate periods could not be established (see below).

The comparatively better performance of AOVMHW and CE at brighter magnitudes with the CRTS data set indicates that these algorithms work well with data which may contain saturated values. The nominal saturation limit for CRTS is $V \sim 12$ and the magnitude used in this analysis is the mean magnitude of the light curve so there may well be observations of bright objects near maxima which are saturated. These algorithms are not attempting to model the phased light curve as a sum of sinusoidal functions and so are less susceptible to non-sinusoidal or truncated sinusoidal-shaped light curves that may occur near a survey's saturation limit. The poor performance of PDM2 with this data set indicates an issue with the irregular sampling strategy of CRTS light curves relative to those in the other two data sets.

5.2 Observations

Fig. 10 shows the completeness fraction as a function of number of observations, n, for the three data sets and accuracy cutoffs. The MACHO and ASAS data sets show no strong dependence on the number of observations, except possibly at smaller values ($n < 200$). However, the CRTS data show a definite dependence. For $n < 200$, there is generally insufficient coverage or sampling of phase for the algorithms to detect the true period effectively. This is compounded by the observing strategy of the CRTS survey: a set of four observations, each separated by 10 min, repeated once or twice per lunation.

Fig. 11 shows the distribution of the time difference, dt, in d between successive observations for the three data sets considered here. From these distributions, we can estimate the number of observations that would be required to ensure a particular minimum phase coverage density for an object of a given periodicity. For each data set, we generate a random observing schedule (set of successive observations) drawn from the appropriate distribution and then determine what the corresponding phased light curve coverage would be for a particular test period in terms of the minimum bin occupancy assuming bin widths of $\Delta\phi = 0.1$. Table 4 gives the median number of observations per time series from 5000 simulations of each data set over a frequency (1./ period) range of 0–4 and for minimum bin occupancies of 5, 10 and 20 observations, respectively.

The bimodal observing distribution of dt of the CRTS data set (see Fig. 11) means that a larger number of individual observations are required for the same phase coverage relative to the other distributions. However, this requires fewer actual nights since each night provides four individual observations. This observing strategy also provides greater sensitivity to short time-scale phenomena - Vio et al. (2013) show that irregular sampling permits one to retrieve information about frequencies much greater than the Nyquist frequency. We note that the proposed core LSST observing strategy is very similar with two back-to-back 15 s exposures and a return to the same pointing within 15–60 min, giving four observations within an hour (Oluseyi et al. 2012).

It may still be the case, however, that there is not enough baseline in any of the surveys to accurately establish the periods of objects with very long periods. If we assume a minimum bin occupancy in phase space of b per bin of width $\Delta\phi$ then *regular* sampling of an object with period P would require an observation every $\Delta t = P\Delta\phi/b$ days. The total number of observations in a light curve, n, for a survey with baseline τ would then be given by $n = \tau/\Delta t$. Rearranging this gives the minimum baseline for a survey to adequately sample a light curve as $\tau \geq Pn\Delta\phi/b$. For an object with a period of 2000 d, say, which is observed regularly to ensure a minimum bin occupancy of 10 with $\Delta\phi = 0.1$ bin widths, the minimum baseline with 150 observations would be 3000 d.

Fig. 12 shows the results as a function of the quoted period, p, for the three data sets and accuracy cutoffs. All three surveys have baselines >2000 d and this is clearly sufficient to recover periods with an accuracy cutoff of 10^{-3} for long-period objects but less

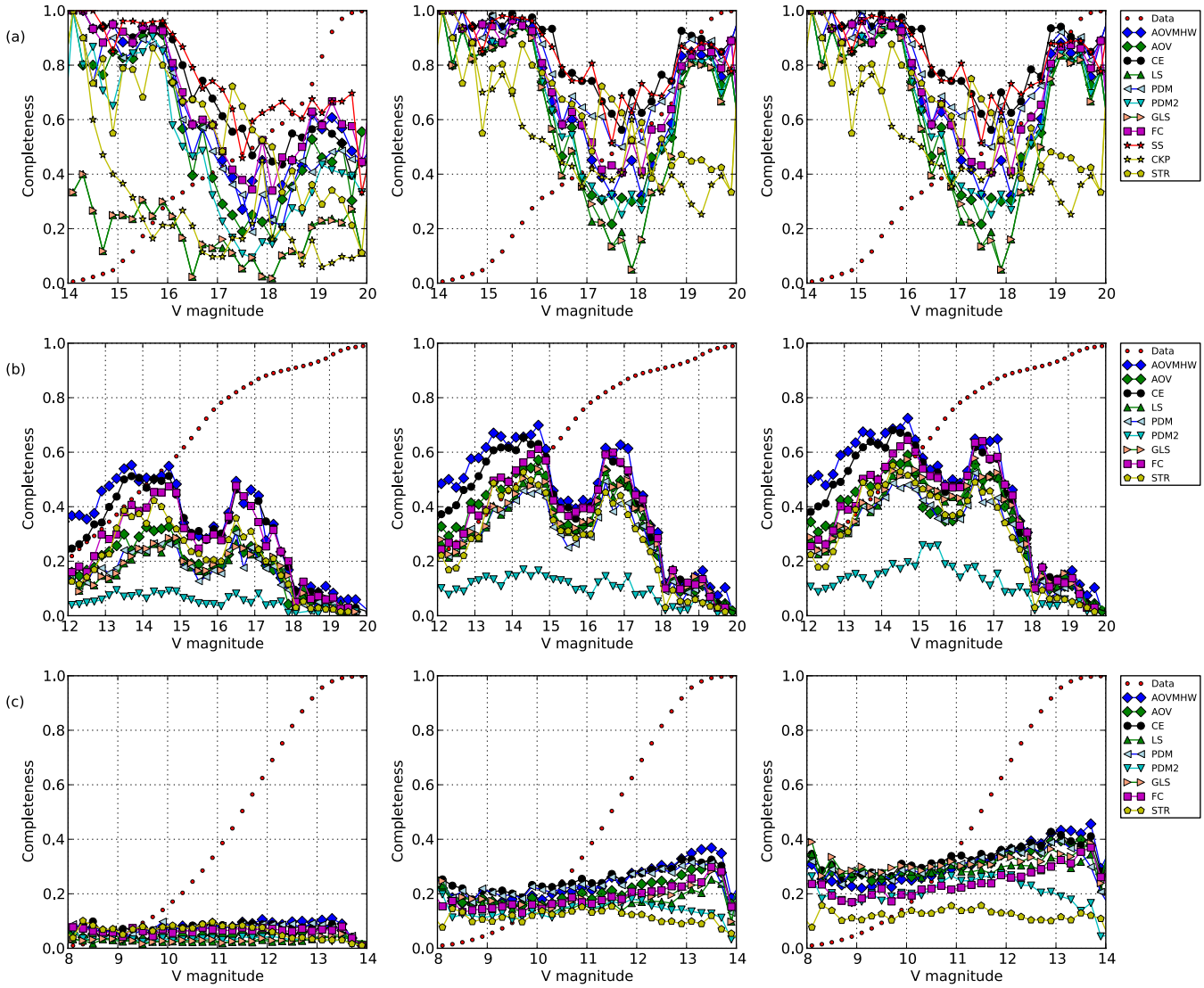

Figure 9. This shows the completeness fraction for the different period finding algorithms as a function of magnitude for each data set: (a) MACHO, (b) CRTS and (c) ASAS. The three plots in each row are for different accuracy cutoffs, equivalent from the left to 10^{-5}, 10^{-4} and 10^{-3} d^{-1} over a 10 yr baseline, respectively (see the text). The different algorithms are denoted by: AOVMHW (blue diamonds), AOV (green diamonds), CE (black circles), LS (green triangles), PDM (left-facing blue triangles), PDM2 (cyan inverted triangles), GLS (right-facing orange triangles), FC (magenta squares), SS (red stars), CKP (yellow stars) and STR (yellow pentagons). The optimal frequency sampling was used where relevant. The small red dots indicate the cumulative magnitude distribution of the relevant data set.

so if higher degrees of accuracy are required. The overall lack of performance for objects with period between roughly 10 and 100 d is due to the (in)efficiencies of the methods with the particular classes of object with those period lengths (see Section 4.3.4).

In terms of both the number of observations and the quoted period, the relative performances of the period finding algorithms seen as a function of magnitude in the previous section are also repeated here with AOVMHW and CE the most successful. As well PDM2 again shows the same issues with CRTS data. We also infer that all the period finding algorithms are stable with a minimum bin occupancy of ~ 10, assuming bin widths of $\Delta\phi = 0.1$.

5.3 Resolution and quality

We have combined the three data sets in terms of our quality measure (modified entropy). Fig. 13 shows the results for this data for each algorithm that allowed the frequency resolution to be set – AOVMHW, AOV, CE, STR, LS, GLS and PDM – and the different frequency resolutions employed. At the accuracy cutoff used (10^{-3} – note that any possible effects of errors in the derived period will be two magnitudes smaller than the cutoff value), there is very little difference between the performance of $\delta\nu = 0.0001$ and the optimal $\delta\nu$ (as noted in Section 3.1, the median optimal $\delta\nu$ is 2.5×10^{-5}) for all the algorithms considered. This suggests that computation time can be saved in future by using a standard frequency resolution of $\delta\nu = 0.0001$ in (initial) frequency range scans and then a finer/optimal resolution for higher accuracy if required. If a lower resolution is preferred then the CE algorithm gives the best performance relative to the others, even for $\delta\nu = 0.01$.

The overall performance of the algorithms as the quality of the light curves varies is broadly consistent. None of the algorithms work with the noisiest of light curves, i.e. those showing the most

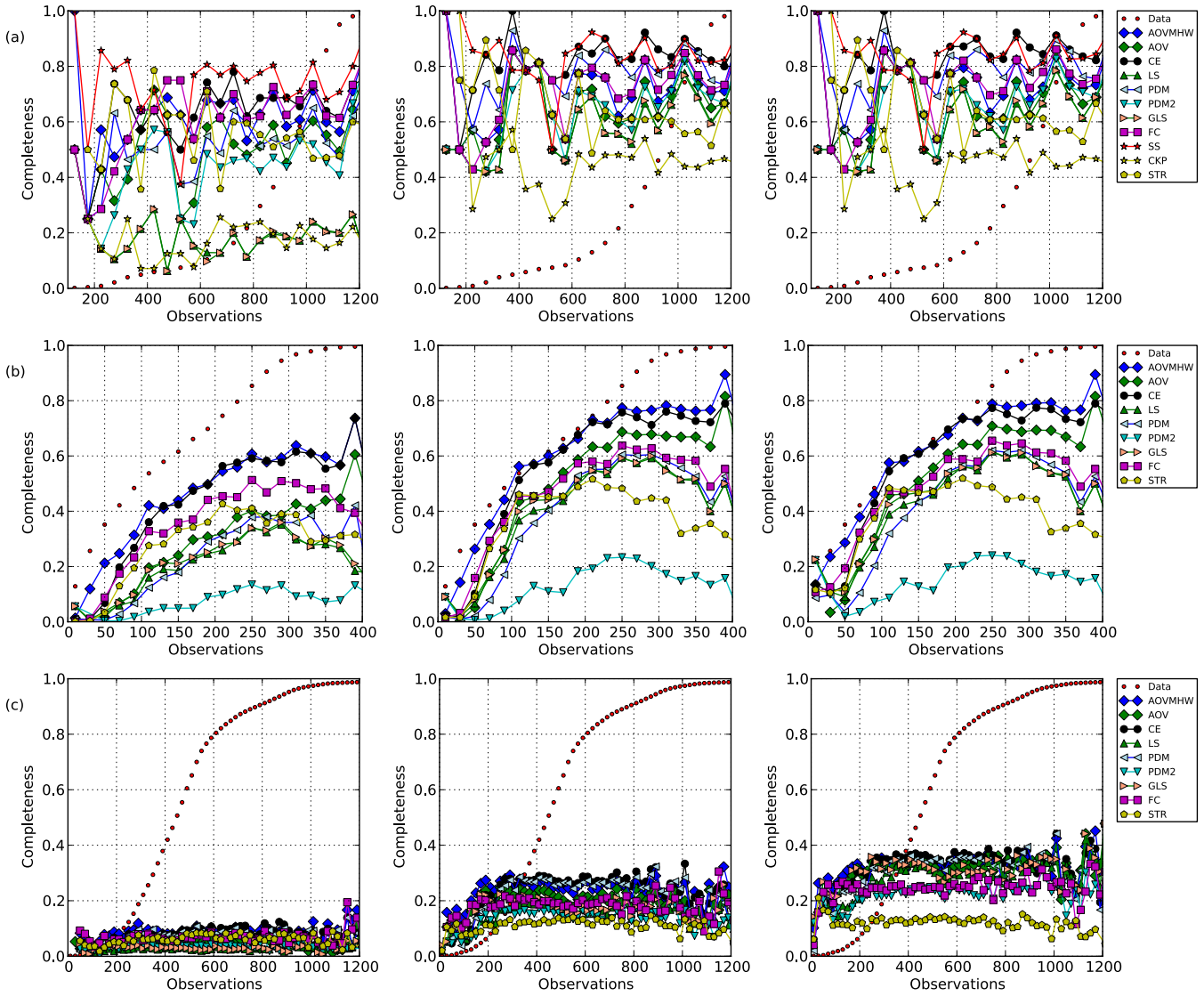

Figure 10. This shows the completeness fraction for the different period finding algorithms as a function of the number of observations per time series for each data set: (a) MACHO, (b) CRTS and (c) ASAS. The three plots in each row again correspond to the different accuracy cutoffs: 10^{-5}, 10^{-4} and 10^{-3} d^{-1} over a 10 yr baseline. The same symbols are used for each algorithm as in Fig. 9 with the optimal frequency sampling used where relevant.

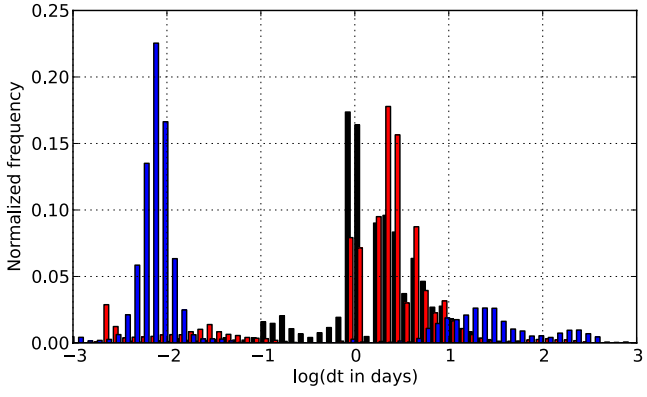

Figure 11. This shows the distribution of time differences in days between successive observations for the three data sets: ASAS (red), MACHO (black) and CRTS (blue). The bin widths are 0.1 dex in log(t).

Table 4. The median number of observations required to ensure the specified minimum coverage in the binned phased light curve of an object for each data set, assuming bin widths of $\Delta\phi = 0.1$

Sampling distribution	Minimum bin occupancy		
	5	10	20
ASAS	90	155	276
MACHO	87	150	270
CRTS	138	214	350

acquired scatter as opposed to variability. All methods show a peak at $\widetilde{\mu} \sim -2$ in the best resolution curves, corresponding to RR Lyrae stars, and a slight hump at $\widetilde{\mu} \sim 2.5$ from eclipsing variables. The best quality light curves ($\widetilde{\mu} < -3$) are dominated by semiregular and pulsating red giant variables. The slightly better relative

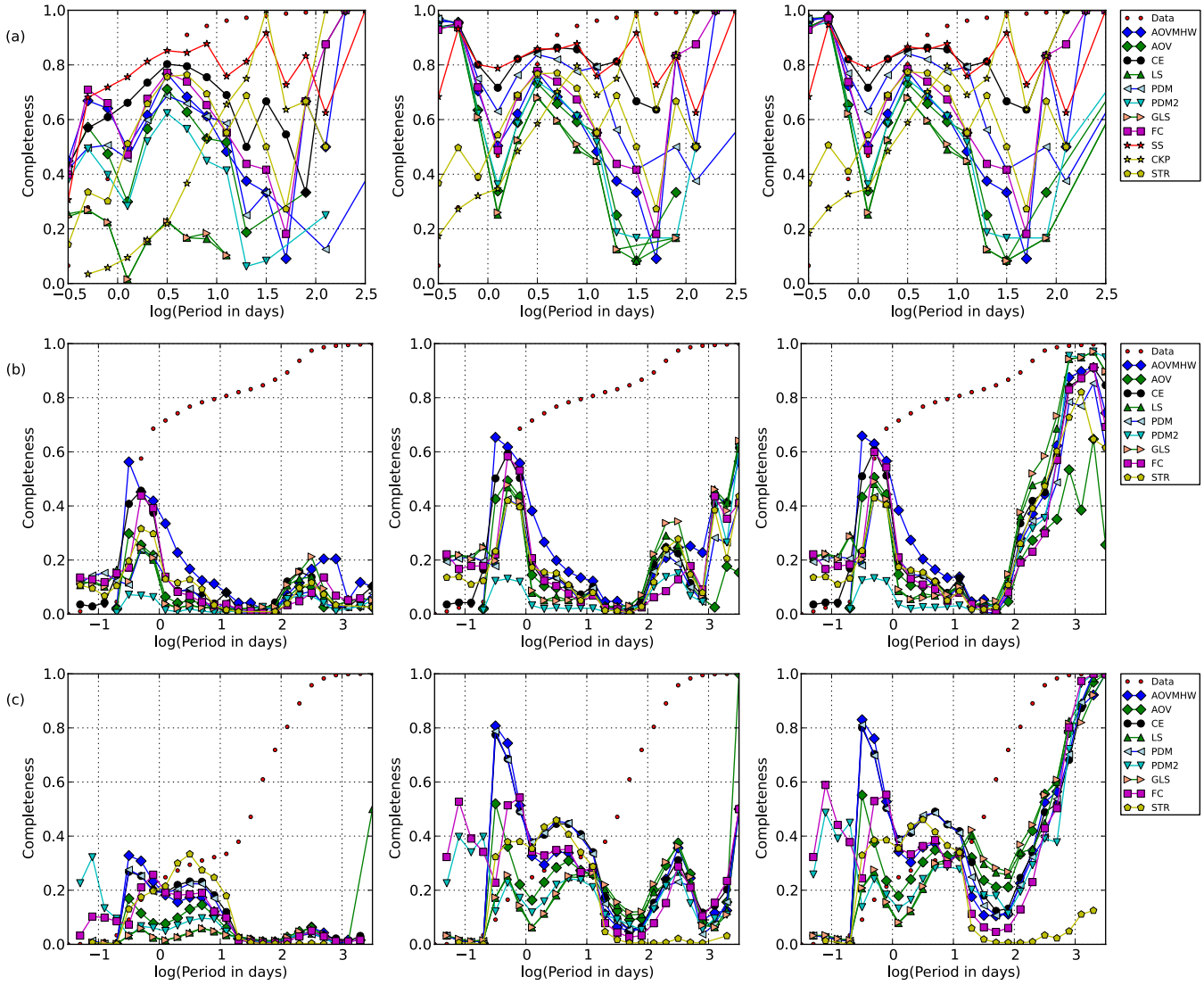

Figure 12. This shows the completeness fraction for the different period finding algorithms as a function of the quoted period in days for each data set: (a) MACHO, (b) CRTS and (c) ASAS. The three plots in each row again correspond to the different accuracy cutoffs: 10^{-5}, 10^{-4} and 10^{-3} d^{-1} over a 10 yr baseline. The same symbols are used for each algorithm as in Fig. 9 with the optimal frequency sampling used where relevant.

performance of the best frequency resolution LS and GLS algorithms with these classes is most likely related to the quoted periods for these objects also having been determined with these algorithms. We will discuss this more in the next subsection.

5.4 Class

The combined data set can also be considered in terms of the various classes of object represented in the data. Fig. 14 shows the results as a function of modified entropy for each of the broadest class designations used: eruptive (P.1, 4194 objects), pulsating (P.2, 45599 objects), rotating (P.3, 455 objects), cataclysmic (P.4, 386 objects), eclipsing (P.5, 14952 objects), X-ray (P.6, 31 objects) and other (P.7, 1434 objects). Unsurprisingly the best results are obtained for the pulsating and eclipsing variable classes as these contain the best defined periodic objects; however, the periods of rotating objects can also be recovered to a reasonable degree. The poor performance for the other classes is most likely caused by a general lack of any clear periodic signal in the light curves for these types of object, for example, LPVs do not seem to oscillate in a clean fashion and so their periods are intrinsically not very well defined.

The shapes of these curves can be attributed to the relative contributions made by different subclasses of object within each of the broadest classes. For example, within the pulsating class, classical Cepheids (Delta Cep) have a mean $\widetilde{\mu} = -1.22$, RR Lyrae have $\widetilde{\mu} = -1.96$, Mira have $\widetilde{\mu} = -2.37$ and semiregular variables have $\widetilde{\mu} = -2.74$. Similarly, within the eclipsing class, there is a sequence from AM Her variables to Algol types to Beta Lyrae types to W UMa types, although the performance for the three eclipsing binary classes is fairly constant. The peak at $\widetilde{\mu} \sim 3$ in the eruptive class results corresponds to weak-line T Tauri stars.

Relative to the other algorithms, PDM2 shows poorer performance with RR Lyrae and eclipsing binaries. The (STR) algorithm also seems to fare worse with semiregular variables than with other pulsating types. However, the clearest differentiation between the algorithms comes with eclipsing variables. AOVMHW and CE are again the most successful and LS and GLS the least. If, however, we relax our accuracy criterion and also include (sub)harmonics

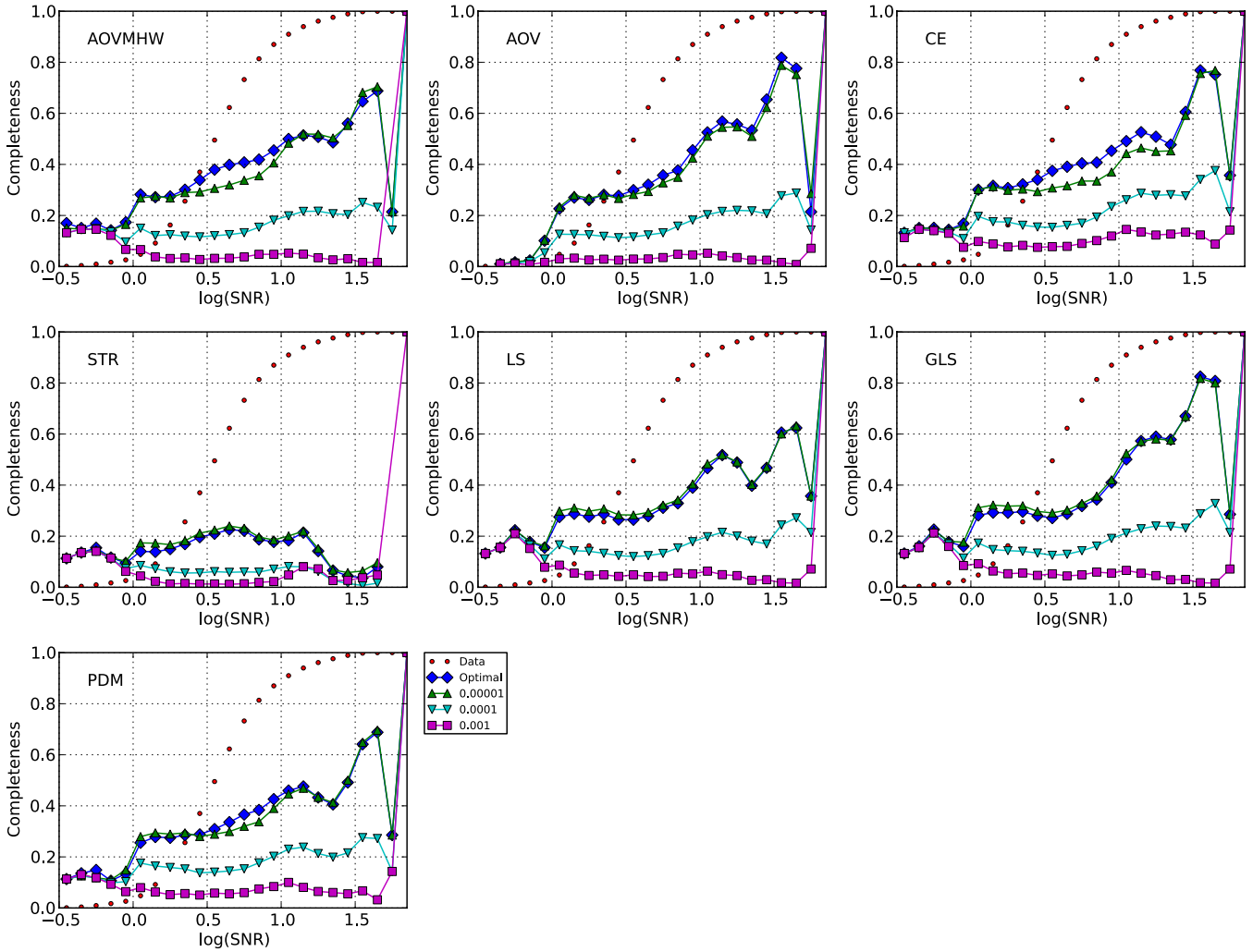

Figure 13. This shows the completeness fraction for the combined data set for the seven algorithms where multiple frequency sampling strategies were applied: AOVMHW, AOV, CE, STR, GLS, LS and PDM. The four curves per plot are: optimal $\delta\nu$ (blue diamonds), $\delta\nu = 0.0001$ (green triangles), $\delta\nu = 0.001$ (inverted cyan triangles) and $\delta\nu = 0.01$ (magenta squares), respectively. An accuracy cutoff of 10^{-3} was used for greatest contrast. The quality of the light curves improves from left to right, i.e. there is less acquired scatter in a light curve with increasing SNR. The red dots indicate the cumulative SNR distribution of the combined data set.

of the true period then we find a significant improvement in LS and GLS relative to the other algorithms (see Fig. 15). Clearly, LS and GLS are the most susceptible of all the algorithms considered here to misidentifying a multiple of the period as the true value and this seems to be particularly the case with W UMa-type eclipsing binaries (hence the decline in performance at $\widetilde{\mu} \sim 3$).

Dubath et al. (2011) report a correct period recovery rate of 91 per cent for LS/GLS depending on skewness value for non-eclipsing variable periods. The same approach finds a half period for 82 per cent of eclipsing variables. They note that better results are obtained with PDM with 38 per cent of the correct periods found and 38 per cent with half periods. However, for eclipsing variables, they adopt a strategy of only assigning a period once the object type has been assigned – doubling an LS-derived period for eclipsing binaries and ellipsoidal variables. Our results for LS/GLS certainly support this approach.

Finally, we note that Drake et al. (2012) find that VSX periods for RR Lyrae have an intrinsic error of $\sim$0.004 per cent which equates to an accuracy cutoff of $\sim 10^{-4}$. This may contribute to the difference in recovery completeness seen in Sections 4.3.1 and 4.3.2 between cutoffs of 10^{-5} and 10^{-4}. However, we have used a limit of 10^{-4} in this section for comparison and so any effect would be reduced.

5.5 Variability

The results in the previous section show the dependences of the various algorithms on the specific object classes but it is also interesting to see whether there is any more general dependence on the variability of an object, i.e. it is easier to find the period of an object with strong variability than with weak. Note that the source of the variability here is in the physical nature of the star rather than in measurement errors which is covered by the dependence on the quality of the light curves (see Section 5.3). Fig. 16 shows the results for the algorithms as a function of MAD (from the median) – a more robust measure of the amplitude of variation than just the extrema values in a light curve.

At the most conservative accuracy cutoff (10^{-5} d^{-1}), there is essentially no dependence on object variability; however, with the

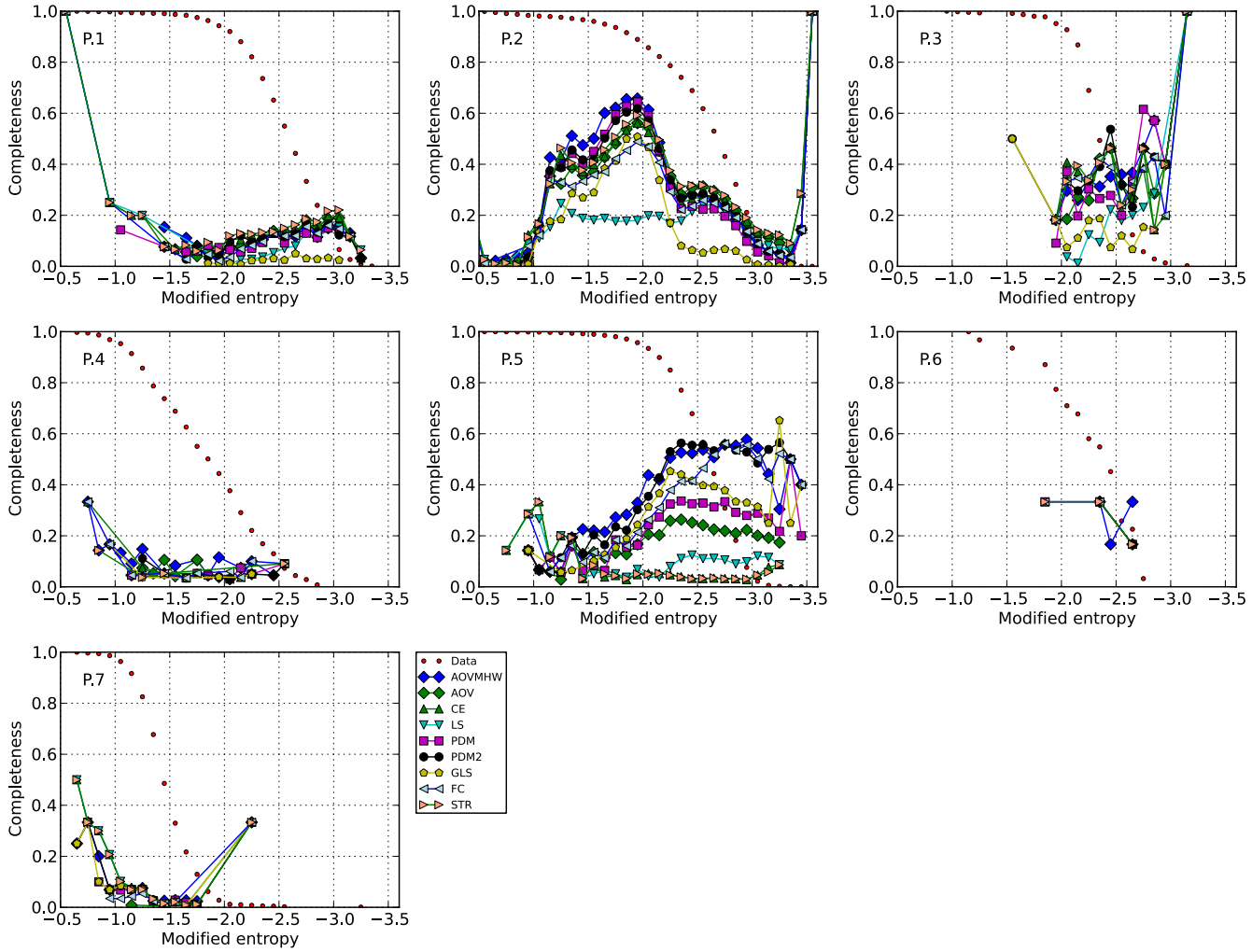

Figure 14. This shows the completeness fraction for the different period finding algorithms on the full combined data set in terms of the seven different broadest classes of variable object represented: P.1 (eruptive), P.2 (pulsating), P.3 (rotating), P.4 (cataclysmic), P.5 (eclipsing), P.6 (X-ray) and P.7 (other). The same symbols are used for each algorithm as in Fig. 9 with the optimal frequency sampling used where relevant. An accuracy cutoff of 10^{-4} was used.

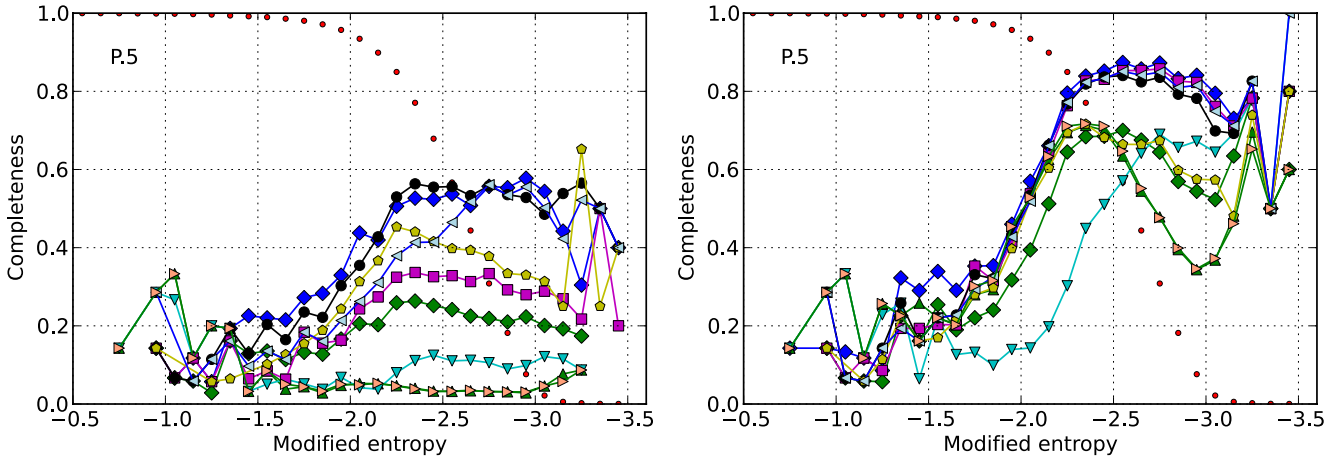

Figure 15. This shows the completeness fraction for the different period finding algorithms for eclipsing variables (P.5) in the full combined data set using just strict period matching (left-hand plot) and allowing period (sub)harmonics as well (right-hand plot). The same symbols are used for each algorithm as in Fig. 9 with the optimal frequency sampling used where relevant. An accuracy cutoff of 10^{-4} was used.

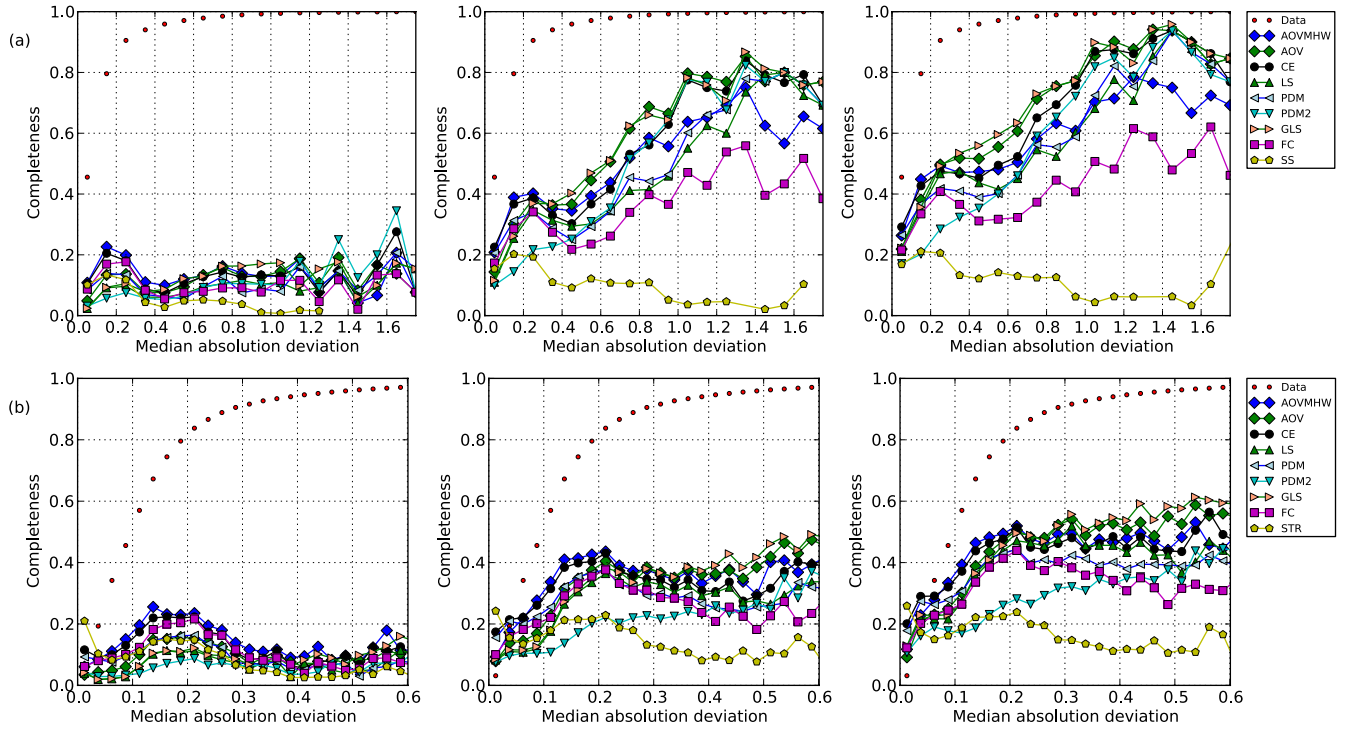

Figure 16. This shows the completeness fraction for the different period finding algorithms on the full combined data set in terms of the MAD from the median of the light curve of the variable object. (a) gives almost the full range of MAD covered by the data set whilst (b) focuses on the smaller range covered by 97 per cent of the data. The three plots in each row again correspond to the different accuracy cutoffs: 10^{-5}, 10^{-4} and 10^{-3} d^{-1} over a 10 yr baseline. The same symbols are used for each algorithm as in Fig. 9 with the optimal frequency sampling used where relevant.

other cutoffs, all the algorithms except STR show better performance with more variable objects. The objects with a correct period at the strictest cutoff tend to have more observations in their light curves so the periodic signal is already better sampled and the increased variability has no real effect. At the other cutoffs, those objects with a poorly sampled light curve due to fewer observations get a boost from larger amplitude variability which makes the periodic signal easier to detect by the algorithms.

This behaviour is modulated by the noise characteristics of the light curves: objects with the same amplitude variability but different noise levels will have different period recovery accuracies. We can use the ratio MAD/log(SNR) as a proxy for the noise in the light curve and Fig. 17 shows the recovery accuracy in terms of this quantity for the different algorithms on the combined data set. The structure in this plot is due to the individual contributions from the data sets with the initial peak at MAD/log(SNR) ~ -2.5 from ASAS and that at ~ -1.2 from CRTS.

One difficulty with the low amplitude variability sources is that the phase errors could be substantial and yet we would not be able to visually recognize this, i.e. a correctly phased light curve and an incorrectly phased one are indistinguishable in the limit of vanishing variability – they both appear constant within observational error tolerance. This should particularly affect those object classes associated with low scale variability, such as small amplitude red giants or weak-line T Tauri objects. If we assume a photometric error of 0.05 mag then $\sim$19 per cent of objects have a MAD value less than this and could potentially have a misassigned period.

5.6 Reliability

For each class in Table 2, we have determined the most reliable method, i.e. method which gives the most number of periods within

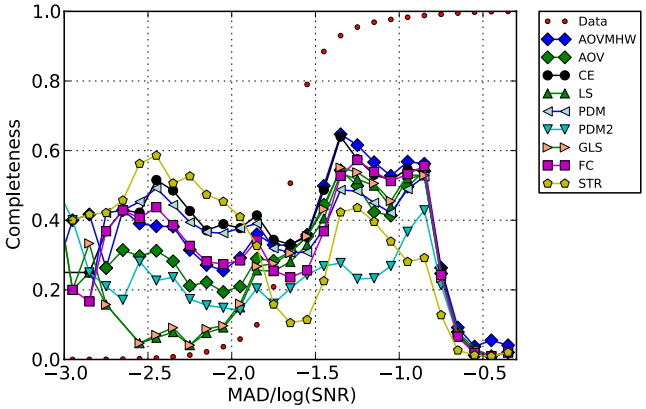

Figure 17. This shows the completeness fraction for the different period finding algorithms on the full combined data set in terms of the ratio of the MAD to log(SNR). An accuracy cutoff of 10^{-3} d^{-1} was used. The same symbols are used for each algorithm as in Fig. 9 with the optimal frequency sampling used where relevant.

an accuracy of 10^{-4} d^{-1}. This is a somewhat subjective measure since a method which finds a few highly accurate periods may be considered more reliable than one which gives a larger number of less accurate ones: when a correct answer is given, it will be very accurate but a larger number of workable periods might be more useful for a particular study. The accuracy limit of 10^{-4} reflects a suitable tradeoff between the two. Fig. 18 shows the distribution of period accuracies for δ Scuti stars (DSCT, P.2.6) for AOVMHW and CE. Although the overall distributions are similar, CE provides slightly more accurate periods (below 10^{-5}) whereas AOVMHW

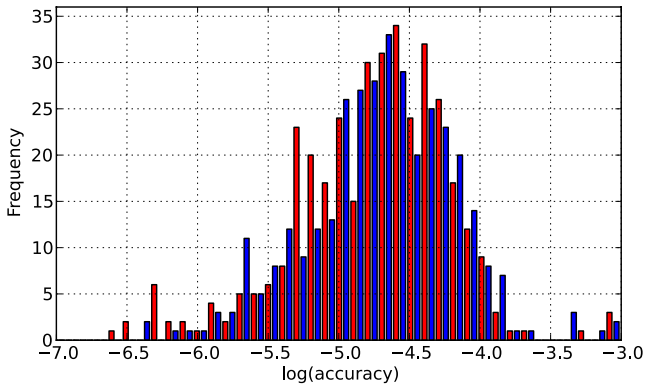

Figure 18. This shows the distribution of the accuracies of AOVMHW (blue) and CE (red) periods for δ Scuti stars (P.2.6).

gives ∼10 per cent more periods overall less than the 10^{-4} d^{-1} limit and so is the more 'reliable' of the two.

Fig. 19 shows the distribution of accuracies for all methods with the combined data set. This shows that CE and PDM both perform well below 10^{-4} along with AOVMHW. The poor reliability of PDM2 is also very clear. The peak at log(accuracy) ∼ 0 is largely due to methods finding half-periods for objects with periods around 1d. AOVMHW shows less susceptibility to this since it involves fitting higher harmonic orders.

5.7 Ensemble method

The accuracy of each of the individual algorithms is clearly dependent on observational factors such as the number of epochs in and the overall quality of a light curve as well as aspects natural to the source itself, such as the amplitude of variability and the actual object type. An ensemble approach, however, might serve to mitigate the effects of these dependences and give a more robust and consistent result. While we reserve a full comparison of ensemble techniques to a forthcoming paper (Graham et al., in preparation), we will consider a simple approach here involving majority opinion.

Each light curve is associated with a set of period estimates, one for each algorithm considered. Within a set, we identify the largest subset of similar values, i.e. those which are within a specified tolerance of each other, and take the median value of this subset as the ensemble period estimate. Table 5 gives the relative performance of AOVMHW, CE and GLS against the ensemble estimate for the three different accuracy cutoff levels used here. We have used the accuracy cutoff as the tolerance value for the three cases, although for a specific accuracy cutoff there is only ∼10 per cent variation in performance if a fixed tolerance of 10^{-5} is used. The results show that a simple ensemble approach does no better than the two strongest single algorithms. If we reduce the set of algorithms considered to just one of each type (AOVMHW, GLS, PDM, STR, CE, FC), we get similar results and just using the mean of the set performs poorly in comparison.

Table 5. The relative performance of some of the algorithms compared against the ensemble majority opinion estimate in terms of total numbers of objects accurately measured at the various accuracy cutoffs.

Algorithm	10^{-5}	10^{-4}	10^{-3}
AOVMHW	10 804	20 983	25 402
CE	9980	20 818	25 746
GLS	4318	15 230	22 468
Ensemble	8452	18 249	24 516
Mean	1678	3075	8188

The relative insensitivity of the ensemble result to the specific tolerance level used (10 per cent over three orders of magnitude) suggests that care should be taken when selecting values that are similar from multiple algorithms.

5.8 Performance

The time taken by an algorithm to determine the period of a light curve is another important factor for large-scale automated analyses. Binning algorithms should show $\mathcal{O}(nN)$ behaviour, where n is the number of measurements and N is the number of frequencies tested whereas FFT-based algorithms should exhibit an $\mathcal{O}(N \log N)$ dependence (Palmer 2009). Graphics processing unit (GPU)-based algorithms, at least for LS and GLS methods, should show $\mathcal{O}(nN)$ scaling tending to $\mathcal{O}(n^2)$ in the limit of large n (Townsend 2010). Of course, the constant in front of the dependent terms in all cases will vary between algorithms and this can be the deciding factor too.

We have measured the computational time required by each algorithm to process each light curve in the MACHO data set with different frequency resolutions (spanning 0.1 to 10^{-6} d^{-1}) on the same machine (an Apple iMac with a 2.8 GHz Intel Core i7 CPU and 8 GB 1333 MHz DDR3 memory running Mac OS X 10.7.4 and AMD Radeon HD 6770M with 512MB for the GPU algorithms) and then performed a linear regression fit to the time taken in seconds as a function of the expected behaviour. Note that the

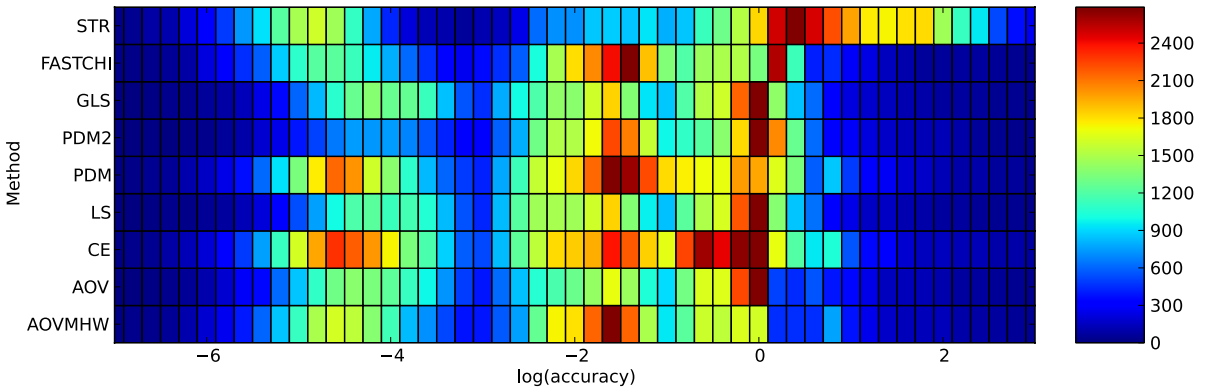

Figure 19. This shows the relative distribution of the accuracies of all the methods with the combined data set.

Table 6. The parameters of the regression fits to the timings of the various algorithms in seconds as a function of number of observations in the light curve, n, and the number of trial frequencies tested, N: $t = An^xN^y + c$ or $t = AN\log N + c$, respectively. An asterisk indicates a GPU-based algorithm. Note that the GLS and LS fits are only valid for $N > 10^5$, otherwise a constant value of 1.5s should be assumed.

	Algorithm	$\log A$	x	y	c
$\mathcal{O}(n^x N^y)$	AOV	−7.939	0.987	0.989	−0.010
	AOVMHW	−6.754	0.997	0.998	0.480
	PDM	−9.446	0.686	0.990	0.156
	PDM2	−5.067	0.948	0.376	0.010
	STR	−9.846	1.073	0.995	0.289
	CE	−8.921	0.600	0.955	0.053
	SS	−1.293	1.007	0.0	0.436
	CKP	−3.166	2.009	0.0	−16.3
	LS*	−2.732	−0.007	0.513	0.078
	GLS*	−2.793	−0.007	0.523	0.088
$\mathcal{O}(N \log N)$	FC	−7.085	–	–	1.472

frequency resolution cannot be set as an argument for SS and CKP and so we just estimate N for these algorithms based on their documentation. We find that the binned algorithms are better described by an $\mathcal{O}(n^x N^y)$ relationship than a strict $\mathcal{O}(nN)$ one but the FFT-based algorithms agree well with the expected $N \log N$ dependence. The GPU-based algorithms (LS, GLS) show an essentially constant timing behaviour to $N = 10^5$ and then transition to $\mathcal{O}(nN)$. Unfortunately we do not have sufficiently large values of n relative to N to show the asymptotic scaling. The constant term is an implementation artefact, attributable to memory overheads in transferring data between the CPU and GPU, and only for $N > 10^5$ does the GPU computation begin to take a discernible amount of time.

Table 6 gives the details of the regression fits to the respective behaviours. Whilst the absolute performance of the algorithms will depend on the hardware used (CPU speed, memory configuration, etc.), the constant values, (A), can give a reasonable indication of their relative speeds. For example, there is a clear factor of ∼15 in performance time between the two versions of AOV – the faster just relying on binning and the slower on model fitting. The two GPU-based algorithms also show a slight difference with GLS being slightly slower as it involves slightly more trigonometric function calls. Note that the intercept (c) for these indicates the memory overhead time and so for small values of N these are not particularly performant.

In fact, we can estimate the minimum approximate time taken by the relative algorithms to determine the period for any light curve and frequency sampling. For a light curve consisting of n observations covering a timespan T and with a fastest time-scale of interest, δt, the maximum frequency $\nu_{\max} \gtrsim 1/2\delta t$ and frequency sampling $\delta\nu \lesssim 1/T$. The minimum number of frequencies to test is then $N_{\min} \simeq T/2\delta t$.

The MACHO data set consists of RR Lyrae, Cepheids and eclipsing binaries and so a realistic fastest time-scale of interest is $\delta t = 0.2$ d. Taking the median timespan ($T = 2720.881$d) and number of observations ($n = 966$), we can calculate a representative minimum time for each algorithm as an indicator of performance. Fig. 20 shows the accuracies versus performance for the

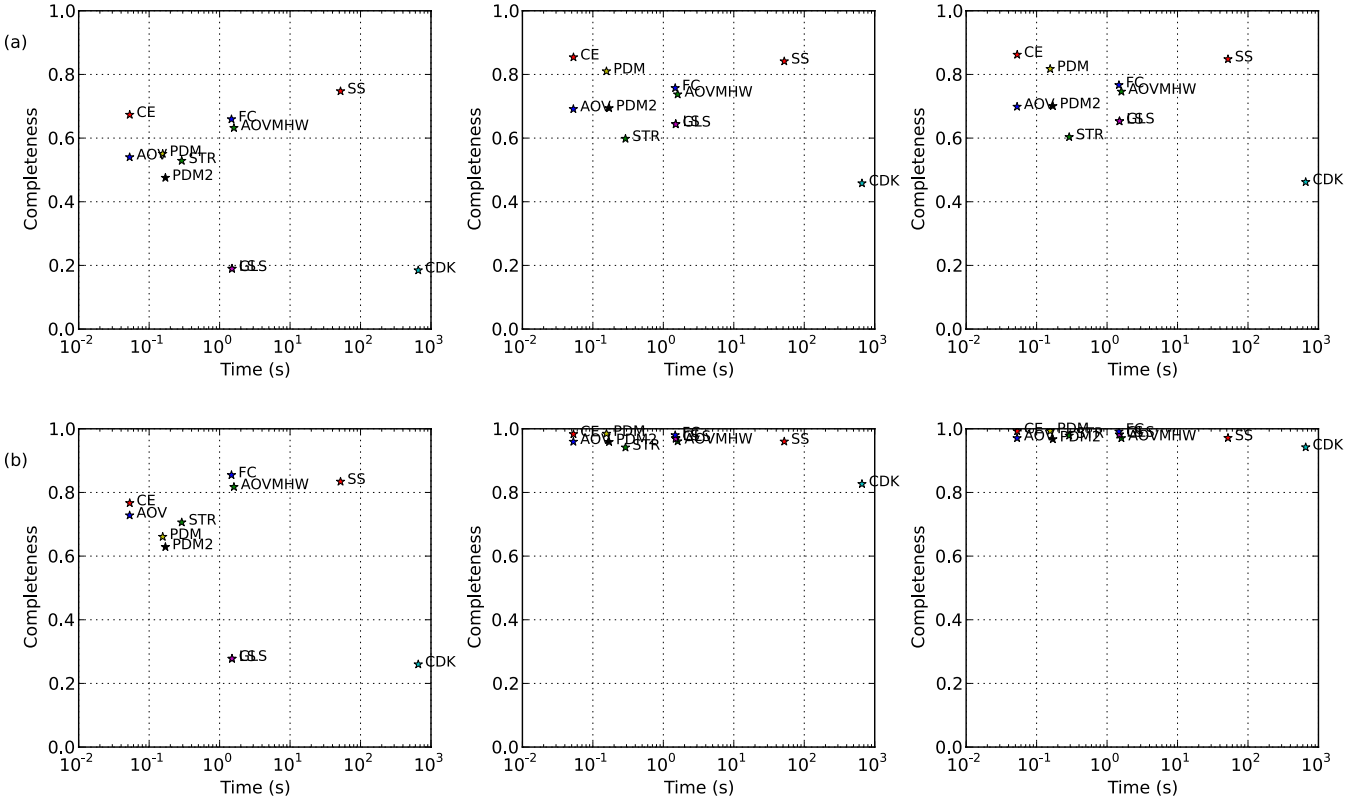

Figure 20. This shows the distribution of the accuracies and timings in seconds for the algorithms considered in this analysis applied to the MACHO data set with $n = 966$ and $N_{\min} = 6803$ for (a) exact periods and (b) including period harmonics. The three plots in each row again correspond to the different accuracy cutoffs: 10^{-5}, 10^{-4} and 10^{-3} d^{-1} over a 10 yr baseline.

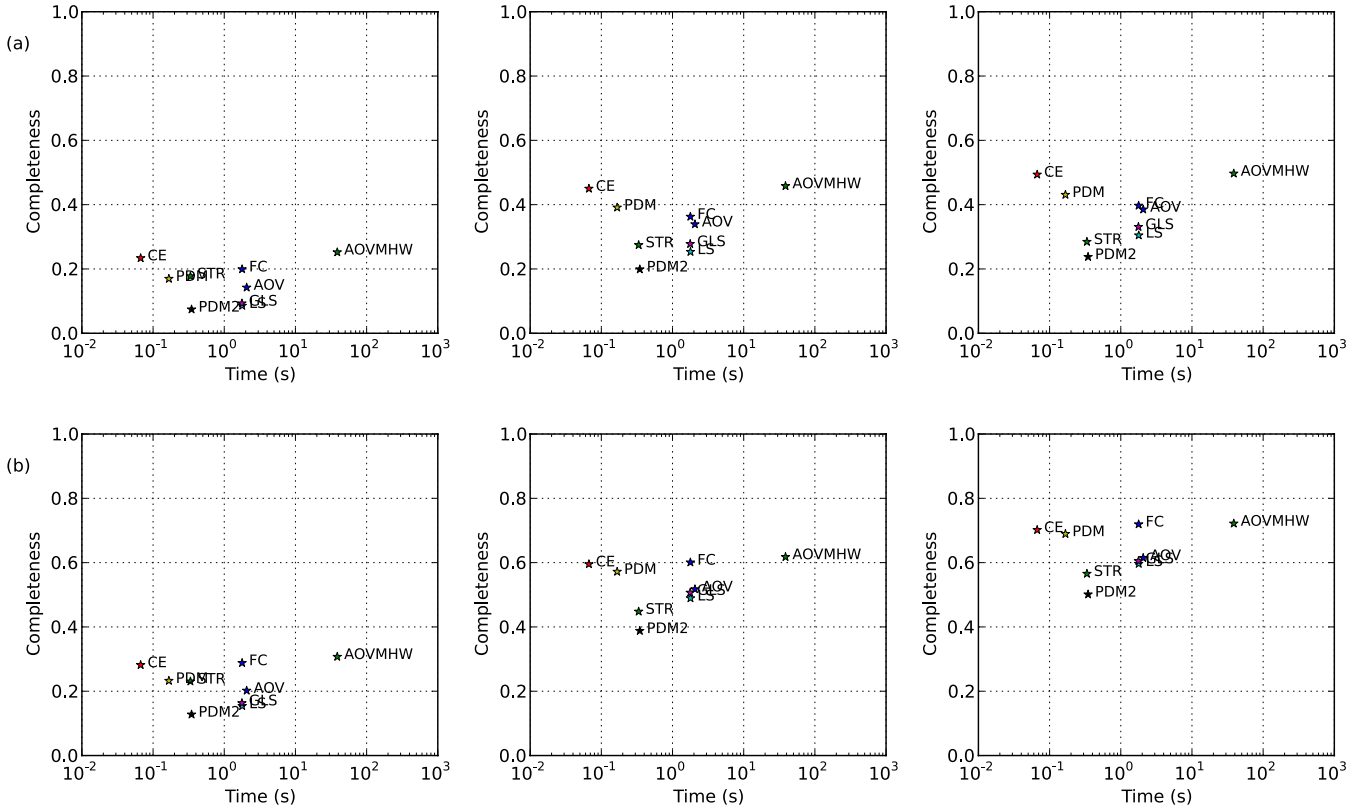

Figure 21. This shows the distribution of the accuracies and timings in seconds for the algorithms applied to the regular periodic variables data set with $n = 347$ and $N_{\min} = 648\,425$ for (a) exact periods and (b) including period harmonics. The three plots in each row again correspond to the different accuracy cutoffs: 10^{-5}, 10^{-4} and 10^{-3} d^{-1} over a 10 yr baseline.

algorithms applied to the MACHO data set, for both exact periods and harmonics.

We have also considered a set of regular periodic variables drawn from all three surveys, consisting of all objects in the following classes or class families: T Tauri (P.1.3.3.3), RS Can Ven (P.1.5), Beta Cepheid (P.2.2), Cepheid (P.2.3), W Vir (P.2.4), Delta Cep (P.2.5), Delta Scuti (P.2.6), Mira (P.2.8), RR Lyrae (P.2.10), rotating (P.3) and eclipsing binary (P.5.1). This has a total of 40 550 members with a median timespan of 2593.97 d and a median number of observations of 347. Fig. 21 shows the accuracies versus timings for the algorithms when applied to this data set, assuming a fastest timescale of interest of $\delta t = 0.02$ (for Delta Scuti objects) and a frequency oversampling of 10. In both sets of plots, the ideal algorithm will be closest to the top left of the plot, i.e. high completeness and low timing slope (fast). We identify this as the CE method.

6 DISCUSSION

The results in the previous section show that at best period finding algorithms can recover the period of a regularly periodic object with a reasonable degree of accuracy (an equivalent phase offset between 10^{-3} and 10^{-4} d^{-1}, say, over a 10 yr baseline) in only about 50 per cent of cases. If one is only interested in detecting periodic behaviour, i.e. the period or a (sub)harmonic, then rates of $\sim$70 per cent are achievable. For objects which do not show simple periodicity, i.e. they are semiperiodic, quasi-periodic or multiperiodic, the situation is broadly much worse, typically only around 10–20 per cent of cases. Of course, the fundamental assumption underlying this analysis is that the quoted period is correct. We have been careful, however, to use data sets where all the light curves have been inspected and the periods confirmed visually.

It should be noted that many of the algorithms score very highly when tested on simulated periodic signals, typically sinusoids with Gaussian noise; the problem seems to come with real data. For many objects, quoted periods would have originally been determined from a small number of observations over a short time baseline. The advent of large-scale synoptic sky surveys means that hundreds of observations over baselines of 5 to 10 yr are now readily available and future projects such as LSST will extend this to baselines of a couple of decades. The digitization of the Harvard plate library (DASCH[9]) offers multidecade baselines for many objects as do other similar historical collections. At the other end of the scale, exoplanet searches and space astroseismology projects, such as SuperWASP, Kepler and CoRoT, are providing (very) high-resolution samplings of a few periodic cycles over periods of days and months.

This wealth of new information allows the long-term stability of periods to be examined as well as intra/intercycle variations and is now suggesting that, even for astrophysical objects exhibiting periodicity, a single value is not capable of characterizing their temporal behaviour. Kepler results have shown that 60 per cent of dwarf stars are more variable than the Sun and probably pulsating variables (Mcquillan, Aigrain & Roberts 2011). RR Lyrae are one of the most populous of pulsating variables and employed as standard candles in studies of galactic structure. However, it has long been known that $\sim$10 per cent exhibit a long-term, generally quasi-periodic modulation of widely varying strength known as the

[9] http://hea-www.harvard.edu/DASCH

Blazhko effect. Studies of variable stars in M3 now show that about a third of RR Lyrae display Blazhko behaviour and the discovery of small amplitude cycle-to-cycle modulations of RRabs (Szabo et al. 2010), in addition to Blazhko effects, cautions that large surveys may have seriously underestimated the number of modulated RR Lyrae stars.

For other populous classes, the situation is equally as complicated with many types of variables showing cyclic period changes over multidecade baselines, such as close binary systems (Zavala et al. 2002) and long-period variables (Lebzelter 2011). Sterken & Jaschek (1996) note that a subgroup of semiregular variables show very clear double periods. In some cases, the longer period may be due to orbital effects indicating that the star is in a binary system. Other semiregular variables apparently show multiperiodicity (e.g. Kerschbaum, Lebzelter & Lazaro 2001), but in general it is not clear whether these stars are truly multiperiodic, chaotic or both, although the actual existence of irregular, i.e. non-periodic, variables among red giants is in dispute (Lebzelter & Obbrugger 2009).

The traditional approach to characterizing periodicity variation is the O−C diagram (e.g. Sterken 2005) which tracks the evolution of the time of appearance of a feature (say the light curve maximum) relative to the corresponding multiples of the period. The functional form of the period change (dp/dt) determined from it can be used in principle to infer the physics of the situation, e.g. a steadily increasing pulsation period implies an expanding star; however, stochastic evolution, e.g. the mean period follows a random walk, can produce equivalent effects in the O−C diagram and distinguishing between the two is an area of active research (Koen 2007). However, the method cannot be applied to long-period pulsating variables where the intrinsic scatter of the period is usually comparable to the experimental error in the period determination (Lombard & Koen 1992). It also has issues with multiperiodic light curves and those with strong modulation.

Alternate approaches rely on techniques from communication and signal processing theory, e.g. wavelets (Foster 1996; Blackman 2011), carrier signals (Pelt et al. 2011) and other time frequency analysis methods. Though these can be very powerful, they are complicated and it is difficult to distil the results down to a single useful characterizing feature, akin to the period. It is possible that the first derivative of the period as a measure of periodicity variation or the (largest) Lyapunov exponent to describe the degree of chaos in a time series (Wolf et al. 1985) may be suitable; however, further discussion of these is outside the scope of this paper.

Another issue potentially affecting the results in this paper is that of object misclassification or class uncertainty. Dubath et al. (2011) only assign a period to eclipsing binaries and ellipsoidal variables once the object type has been determined (to mitigate the half-period issue with these objects). Our results support this as a viable strategy for the LS/GLS algorithms: the improvement seen on our combined data set is ∼4 per cent recovery to ∼50 per cent whereas other algorithms show a significant drop. The biggest source of error, however, will be those objects that have been misidentified as eclipsing variables, although this could be mitigated by a high classification accuracy for eclipsing binaries. Whilst a detailed discussion on object classification is beyond the scope of this paper, we will note a few points.

The MACC classes that we employ for the ASAS data use a probabilistically determined 28-term scheme whilst the original ACVS classifications for the same objects used 439 different categories (different combinations and permutations of a set of about 20 terms), although 60 per cent of objects were classified as 'MISC'. One of the hardest classes to distinguish between is RR Lyrae with fundamental overtones (RRC) and W UMas (EC) and the effect of misidentification would be that a half-period (for a W UMa) is reported as the true period (for a RRC). 12 per cent of MACC W UMas are considered to be RRCs by ACVS and about 10 per cent vice versa, although in only a handful of cases are the probabilities of both classes within 5 per cent. The MACHO data set shows a similar level of misclassification (∼10 per cent) between the provided object type (RR) and the MACHO assigned class (eclipsing binary). We therefore estimate that there may be ∼10 per cent error in the class-based results arising from misassigned object types. Note, however, that if data from more than one band is available then these types can be better distinguished with PCA (Süveges et al. 2012).

It is also possible to use additional data to check the classifications, e.g. T Tauri/HAEBE stars and any massive star classes should be near the plane. Weak-line T Tauri (WTTS) objects in the ASAS data set do not correspond very well with the plane or nearby star-forming regions casting doubt on the reliability of these classifications (Feigelson, private communication). We have also compared the reported periods for objects against the expected period ranges for their class drawn from Debosscher et al. (2007) and object definitions in VSX. We find that of the 41 299 objects in the combined data set for which we have ranges, 40 052 have periods which lie within the expected class ranges. This is certainly well within the ∼10 per cent misclassification error we have estimated and suggests that class uncertainty is not a significant issue in this analysis. We note, however, that coupling classification and period finding may produce more accurate results, e.g. Rimoldini (2013) finds improved classifications using a weighting scheme based on the period-folded light curve.

7 CONCLUSIONS

In this paper, we have analysed the performance and dependences of the most popular period finding algorithms against a comprehensive set of light curves. We find that

(i) all methods are dependent on the quality of the light curve and show a decline in period recovery with lower quality light curves as a consequence of fewer observations, fainter magnitudes and/or noisier data and an increase in period recovery with higher object variability;

(ii) all algorithms are stable with a minimum bin occupancy of 10 (assuming $\Delta\phi = 0.1$);

(iii) a bimodal observing strategy consisting of pairs (or more) of short δt observations per night and normal repeat visits is better than single observations with normal repeats;

(iv) a minimum frequency step of $\delta\nu = 0.0001$ is sufficient;

(v) the algorithms work best with pulsating and eclipsing variable classes;

(vi) straightforward ensemble methods show no improvement over single algorithms.

We also confirm that LS/GLS are strongly effected by the half-period issue for eclipsing binaries and find that PDM2 has issues with irregular sampling of light curves and that AOVMHW and CE work well at bright magnitudes (containing saturated values). Finally, in terms of overall performance factors considered here – greatest period recovery and time – CE is the best algorithm with AOV and PDM viable alternatives.

New and better techniques may be proposed that change the findings of this analysis. To keep track of these, we intend to maintain

an online version of this work, updating it as appropriate. If anyone has an algorithm that they would like to see included then they should get in touch with us.

ACKNOWLEDGEMENTS

We thank the referee, Lorenzo Rimoldini, for their meticulous reading of the paper and useful comments. We thank Eric Feigelson for useful discussions on this work and the various providers of software.

This work was supported in part by the NSF grants AST-0909182 and IIS-1118041, by the W. M. Keck Institute for Space Studies, and by the US Virtual Astronomical Observatory, itself supported by the NSF grant AST-0834235.

This research has made use of data obtained from or software provided by the US Virtual Astronomical Observatory, which is sponsored by the National Science Foundation and the National Aeronautics and Space Administration.

This research has made use of the SIMBAD data base, operated at CDS, Strasbourg, France and the International VSX data base, operated at AAVSO, Cambridge, Massachusetts, USA.

REFERENCES

Alcock C. et al., 2003, VizieR On-line Data Catalog: II/247
Auvergne M. et al., 2009, A&A, 506, 411
Baluev R., 2012, in Mickaelian A. M., Malkov O. Yu., Samus N. N., eds. Fifty years of Cosmic Era: Real and Virtual Studies of the Sky. NAS RA, Yerevan, p. 230
Barsdell B. R., Barnes D. G., Fluke C. J., 2010, MNRAS, 408, 1936
Blackman C., 2010, ApJS, 191, 185
Cincotta P. M., 1999, MNRAS, 307, 941
Cincotta P. M., Mendez M., Nunez J. A., 1995, ApJ, 449, 231
de Jager O. C., Raubenheimer B. C., Swanepoel J. W. H., 1989, A&A, 221, 180
Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159
Deeming T. J., 1975, Ap&SS, 36, 137
Distefano E., Lanzafame A. C., Lanza A. F., Messina S., Korn A. J., Eriksson K., Cuypers J., 2012, MNRAS, 421, 2774
Drake A. J. et al., 2009, ApJ, 696, 870
Drake A. J. et al., 2012, ApJ, 763, 32
Dubath. P. et al., 2011, MNRAS, 414, 2602
Dworetsky M. M., 1983, MNRAS, 203, 917
Foster G., 1996, AJ, 112, 1709
Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., 2013, MNRAS, preprint (arXiv:1306.6664)
Gregory P. C., Loredo T. J., 1992, ApJ, 398, 146
Heck A., Manfroid J., Mersch G., 1985, A&AS, 59, 63
Huijse P., Estevez P. A., Zegers P., Principe J. C., Protopapas P., 2011, IEEE Signal Process. Lett., 18, 371
Huijse P., Estevez P. A., Protopapas P., Zegers P., Principe J. C., 2012, IEEE Trans. Signal Process., 60, 5135
Ivezić Ž. et al., 2011, preprint (arXiv:0805.2366)
Jetsu L., Pelt J., 1999, A&AS, 139, 629
Jurkevich I., 1971, Ap&SS, 13, 154
Kaiser N. et al., 2002, in Tyson J. A., ed., Proc. SPIE Vol. 4836, Survey and Other Telescope Technologies and Discoveries. SPIE, Bellingham, p. 154
Kato T., Uemura M., 2012, PASJ, 64, 122
Kerschbaum F., Lebzelter T., Lazaro C., 2001, A&A, 375, 527
Koch D. G. et al., 2010, ApJ, 713, L79
Koen C., 2006, MNRAS, 371, 1390
Koen C., 2007, in Babu G. J., Feigelson E. D., eds, ASP Conf. Ser. Vol. 371, Statistical Challenges in Modern Astronomy IV, Astron. Soc. Pac., San Francisco, p. 324
Lebzelter T., 2011, A&A, 530, 35
Lebzelter T., Obbrugger M., 2009, Astron. Nachr., 330, 390
Lomb N. R., 1976, Ap&SS, 39, 447
Lombard F., Koen C., 1992, MNRAS, 263, 309
Mcquilan A., Aigrain S., Roberts S., 2011, A&A, 539, 137
Mishra B. P., Principe J. C., Estévez P. A., Protopapas P., 2011, IEEE International MLSP Workshop, Estimation of Periodicity in Non-Uniformly Sampled Astronomical Data Using a 2D Kernel in Correntropy. IEEE, Los Alamos, p. 1
Oluseyi H. M. et al., 2012, AJ, 144, 9
Palmer D. M., 2009, ApJ, 695, 496
Pelt J., 2011, Olspert N., Mantere M., Tuominen I., 2011, A&A, 535, 23
Perryman M. A. C.,& ESA, 1997, ESA SP-1200: The HIPPARCOS and TYCHO Catalogues, Astrometric and Photometric Star Catalogues Derived from the ESA Space Astrometry Mission. ESA, Noordwijk
Pojmanski G., 2002, Acta Astron., 52, 397
Pojmanski G., Pilecki B., Szczygiel D., 2005, Acta Astron., 55, 275
Rau A. et al., 2009, PASP, 121, 1334
Reimann J. D., 1994, PhD thesis, Univ. California, Berkeley
Richards J. W. et al., 2011, ApJ, 733, 10
Richards J. W., Starr D. L., Miller A. A., Bloom J. S., Butler N. R., Brink H., Crellin-Quick A., 2012, ApJ, ApJS, 203, 32
Rimoldini L., 2013, MNRAS, preprint (arXiv:1304.6616)
Samus N. N. et al., 2009, General Catalog of Variable Stars (version 2012-04-15). Centre de Données astronomiques de Strasbourg, Strasbourg
Scargle J. D., 1982, ApJ, 263, 835
Schwarzenberg-Czerny A., 1989, MNRAS, 241, 153
Schwarzenberg-Czerny A., 1996, ApJ, 460, L107
Schwarzenberg-Czerny A., 1999, ApJ, 516, 315
Schwarzenberg-Czerny A., Beaulieu J.-Ph., 2006, MNRAS, 365, 165
Shin M. S., Byun Y. I., 2004, J. Korean Astron. Soc., 37, 79
Shin M. S., Sekora M., Byun Y. I., 2009, MNRAS, 400, 1897
Stellingwerf R. F., 1978, ApJ, 224, 953
Stellingwerf R. F., 2011, in McWilliam A., ed., Carnegie Observatories Astrophysics Series Vol. 5, RR Lyrae Stars, Metal-Poor Stars, and the Galaxy. Observatories of the Carnegie Institution of Washington, Pasadena, p. 47
Sterken C., 2005, in Sterken, C., ed., ASP Conf. Ser. Vol. 335, The Light-Time Effect in Astrophysics. Astron. Soc. Pac., San Francisco, p. 3.
Sterken C., Jaschek C., eds, 1996, Light Curves of Variable Stars: A Pictorial Atlas. Cambridge Univ. Press, Cambridge
Süveges M. et al., 2012, MNRAS, 424, 2528
Swingler D. N., 1989, AJ, 97, 280
Szabo R. et al., 2010, MNRAS, 409, 1244
Townsend R. H. D., 2010, ApJS, 191, 247
Udalski A., Szymanski M., Kaluzny J., Kubiak M., Krzeminski W., Mateo M., Preston G. W., Paczynski B., 1993, Acta Astron., 43, 289
Vio R., Diaz-Trigo M., Andreani P., 2013, Astron. Comput., 1, 5
Wang Y., Khardon R., Protopapas P., 2012, ApJ, 756, 67
Watson C. L., 2006, SASS, 25, 47
Wolf A., Swift J. B., Swinney H. L., Vastano J. A., 1985, Physica D, 16, 285
Wood P. R. et al., 1999, in Le Bertre T., Lebre A., Waelkens C., eds, IAU Symp. 191, Asymptotic Giant Branch Stars. Astron. Soc. Pac., San Francisco, p. 151
Zavala R. T. et al., 2002, AJ, 123, 450
Zechmeister M., Kürster M., 2009, A&A, 496, 577

APPENDIX A: GPU VERSIONS OF LOMB–SCARGLE ALGORITHMS

GPUs offer a significant performance improvement for parallelizable algorithms (see Barsdell, Barnes & Fluke 2010 for a review of their potential for astronomy). Townsend (2010) provides a LS periodogram code implemented within NVIDIA's CUDA framework. We have ported this to OpenCL which provides a platform-neutral

manner to program devices such as multicore CPUs and GPUs at a slight performance expense. This allows us to run the code on non-NVIDIA devices, such as AMD Radeon GPUs. We have also implemented an OpenCL version of the GLS periodogram (Zechmeister & Kürster 2009). Details of both are given below.

A1 Porting CUDA Lomb–Scargle to OpenCL

Porting the CULSP computation kernel essentially consists of just three steps.

A1.1 Rewriting the kernel signature

The kernel signature under CUDA is

```
__global__ void
__launch_bounds__(BLOCK_SIZE)
culsp_kernel(float *d_t, float *d_X, float *d_P,
  float df, int N_t) {
```

Under OpenCL, this becomes

```
__kernel void culsp_kernel(__global float *d_t,
  __global float *d_X, __global float *d_P,
  float df, int N_t) {
```

A1.2 Thread management

OpenCL has global commands for addressing threads so `blockIDx.x` is given by `get_group_id(0)` and `threadIdx.x` by `get_local_id(0)`. Synching threads within a block, `_syncthreads`, is replaced with `barrier(CLK_LOCAL_MEM_FENCE)`. Shared memory is also allocated with a `__local` keyword instead of `__shared__`.

A1.3 Intrinsic function calls

The OpenCL library has slightly different versions of certain functions to CUDA. `rintf` is `rint` under OpenCL and the CUDA function call `__sincosf(TWOPI*ft, &s, &c)` becomes a variable assignment: `s = sincos(TWOPI*ft, &c)`.

A2 An OpenCL generalized Lomb–Scargle kernel

As noted in Townsend (2010), the expressions derived in Zechmeister & Kürster (2009) to calculate the GLS periodogram are very similar in form to those used in the CUDA LS kernel. It is therefore straightforward to construct a GPU kernel for the GLS (see Fig. A1).

```
__kernel void k_clglsp(__global float *d_t, __global float *d_Y,
                       __global float *d_w, __global float *d_P,
                       float df, int N_t)
{

__local float s_t[BLOCK_SIZE];
__local float s_Y[BLOCK_SIZE];
__local float s_w[BLOCK_SIZE];

int bx = get_group_id(0);
int tx = get_local_id(0);

// Calculate the frequency

float f = (bx * BLOCK_SIZE + tx + 1) * df;

// Calculate the various sums

float YC = 0.f;
float YS = 0.f;
float CC = 0.f;
float CS = 0.f;
float YY = 0.f;
float Y = 0.f;
float C = 0.f;
float S = 0.f;

float YC_chunk = 0.f;
float YS_chunk = 0.f;
float CC_chunk = 0.f;
float CS_chunk = 0.f;
float YY_chunk = 0.f;
float Y_chunk = 0.f;
float C_chunk = 0.f;
float S_chunk = 0.f;

int j;

for (j = 0; j < N_t - BLOCK_SIZE; j += BLOCK_SIZE) {

  // Load the chunk into shared memory

  barrier(CLK_LOCAL_MEM_FENCE);

  s_t[tx] = d_t[j + tx];
  s_Y[tx] = d_Y[j + tx];
  s_w[tx] = d_w[j + tx];

  barrier(CLK_LOCAL_MEM_FENCE);

  // Update the sums

  #pragma unroll
  for(int k = 0; k < %(BLOCK_SIZE)s; k++) {

    // Range reduction

    float ft = f*s_t[k];
    ft -= rint(ft);

    float c;
    float s;

    s = sincos(TWOPI * ft, &c);

    YC_chunk += s_w[k] * s_Y[k] * c;
    YS_chunk += s_w[k] * s_Y[k] * s;
    CC_chunk += s_w[k] * c * c;
    CS_chunk += s_w[k] * c * s;
    YY_chunk += s_w[k] * s_Y[k] * s_Y[k];
    Y_chunk += s_w[k] * s_Y[k];
    C_chunk += s_w[k] * c;
    S_chunk += s_w[k] * s;


  }

  YC += YC_chunk;
  YS += YS_chunk;
  CC += CC_chunk;
  CS += CS_chunk;
  YY += YY_chunk;
  Y += Y_chunk;
  C += C_chunk;
  S += S_chunk;


  YC_chunk = 0.f;
  YS_chunk = 0.f;
  CC_chunk = 0.f;
  CS_chunk = 0.f;
  YY_chunk = 0.f;
  Y_chunk = 0.f;
  C_chunk = 0.f;
  S_chunk = 0.f;

}

// Handle the final chunk

barrier(CLK_LOCAL_MEM_FENCE);

if(j+tx < N_t) {

  s_t[tx] = d_t[j + tx];
  s_Y[tx] = d_Y[j + tx];
  s_w[tx] = d_w[j + tx];

}

barrier(CLK_LOCAL_MEM_FENCE);

for(int k = 0; k < N_t-j; k++) {

  // Range reduction

  float ft = f * s_t[k];
  ft -= rint(ft);

  float c;
  float s;

  s = sincos(TWOPI * ft, &c);

  YC_chunk += s_w[k] * s_Y[k] * c;
  YS_chunk += s_w[k] * s_Y[k] * s;
  CC_chunk += s_w[k] * c * c;
  CS_chunk += s_w[k] * c * s;
  YY_chunk += s_w[k] * s_Y[k] * s_Y[k];
  Y_chunk += s_w[k] * s_Y[k];
  C_chunk += s_w[k] * c;
  S_chunk += s_w[k] * s;

}

YC += YC_chunk;
YS += YS_chunk;
CC += CC_chunk;
CS += CS_chunk;
YY += YY_chunk;
Y += Y_chunk;
C += C_chunk;
S += S_chunk;

float SS = 1.f - CC;

// Calculate the tau terms

float ct;
float st;
float CCW = CC - C * C;
float CSW = CS - C * S;
float SSW = SS - S * S;

st = sincos(0.5f * atan2(2.f * CSW, CCW - SSW), &ct);

// Calculate P

float YCW = YC - Y * C;
float YSW = YS - Y * S;

d_P[bx*%(BLOCK_SIZE)s+tx] =
         ((ct*YCW + st*YSW)*(ct*YCW + st*YSW)/
          (ct*ct*CCW + 2*ct*st*CSW + st*st*SSW) +
          (ct*YSW - st*YCW)*(ct*YSW - st*YCW)/
          (ct*ct*SSW - 2*ct*st*CSW + st*st*CCW)) / (YY - Y * Y);

// Finish
}
```

Figure A1. This gives an OpenCL computation kernel for the GLS periodogram.

This paper has been typeset from a TEX/LATEX file prepared by the author.