Mon. Not. R. Astron. Soc. **369,** 677–696 (2006) doi:10.1111/j.1365-2966.2006.10327.x

Finding outlier light curves in catalogues of periodic variable stars

P. Protopapas,[1⋆] J. M. Giammarco,[2] L. Faccioli,[2] M. F. Struble,[2] R. Dave[2] and C. Alcock[1]

[1]*Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA*
[2]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*

Accepted 2006 March 14. Received 2006 March 14; in original form 2006 February 8

ABSTRACT

We present a methodology to discover outliers in catalogues of periodic light curves. We use a cross-correlation as the measure of 'similarity' between two individual light curves, and then classify light curves with lowest average 'similarity' as outliers. We performed the analysis on catalogues of periodic variable stars of known type from the MACHO and OGLE projects. This analysis was carried out in Fourier space and we established that our method correctly identifies light curves that do not belong to those catalogues as outliers. We show how an approximation to this method, carried out in real space, can scale to large data sets that will be available in the near future such as those anticipated from the Panoramic Survey Telescope & Rapid Response System (Pan-STARRS) and Large Synoptic Survey Telescope (LSST).

Key words: methods: data analysis – astronomical data bases: miscellaneous – catalogues – binaries: eclipsing – Cepheids – stars: variables: other.

1 INTRODUCTION

One major byproduct of the completed MACHO and ongoing OGLE, EROS, and MOA microlensing surveys are catalogues of $\sim 10^5$ variable stars generated from long temporal photometric monitoring of stars in selected fields of the Magellanic Clouds and the Galactic bulge (Ferlet, Maillard & Raban 1997; Paczyński 2001). The fraction of periodic optical variables amongst these variable stars depends on the magnitude limit of surveys, the criteria for periodicity, and, on the techniques used. These periods were estimated via various statistical techniques of differing fidelity, from periodogram analysis of Lomb (1976) to the supersmoother developed by Riemann (1994), which was used in MACHO surveys. For example, the MACHO survey of a combined sample of $\sim 2 \times 10^5$ galactic and Magellanic Cloud variables (Faccioli et al. 2006) found ∼32 per cent are periodic. The General Catalog of Variable Stars (GCVS) indicates the majority of catalogued variables in the Galaxy are periodic (see Allen 1973); the updated version with $\sim 4 \times 10^4$ variables from the Sternberg Astronomical Institute[1] indicates ∼60 per cent are classified as periodic. In these surveys, classification of periodic variables was performed by eye, based primarily on the visual appearance of their light curves folded with an estimated period, and their locations in the colour–magnitude and period–luminosity diagrams. Automatic classification procedures are also employed using direct parametric analysis (Udalski et al. 1999a), fully automated neural networks (Pojmanski 2000, 2002), neural network method (Belokurov, Evans & Du 2003; Belokurov, Evans & Le Du 2004), and Bayesian classifiers (Eyer & Blake 2005); other techniques are under development (Wożniak et al. 2002). The reliability of the type classification of light curves with these automated techniques is estimated to be ∼90 per cent (Wożniak et al. 2002). Interestingly, the fraction of variables classified as periodic by automated techniques is lower than those classified by eye in the GCVS and MACHO data: Pojmanski (2000) and Żebruń et al. (2001a) found ∼10 per cent and Mizerski & Bejger (2002) found ∼20 per cent for their comparatively smaller samples.

A natural question that arises concerns the detection of outliers in variable star catalogues, that is, members whose light curves deviate at a prescribed statistical level from the rest. There could be several reasons for this: a poor or incorrect period caused by noisy photometric data, outright misclassification, or, perhaps rarely and more interestingly, an intrinsic physical difference such as a slowly changing period or brightness amplitude which introduces noise in the folded light curve, analogous to the longer-term variability of the

⋆E-mail: pprotopapas@cfa.harvard.edu
[1] http://www.sai.msu.su/groups/cluster/gcvs/gcvs/iii/vartype.txt.

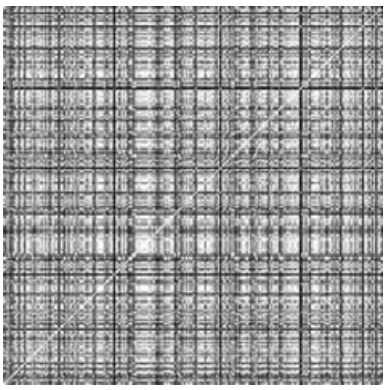

Figure 1. Similarity matrix for 200 light curves of randomly chosen and Eclipsing Binaries from the MACHO survey. Each point represents the square of the correlation between two light curves. The bright points and light areas correspond to strongly correlated/anti-correlated pairs of light curves. Dark points correspond to weakly correlated pairs of light curves. Note that the indexing starts at the lower left-hand corner with (1,1).

Cepheid variable Polaris (Evans, Sasselov & Short 2002; Engle, Guinan & Koch 2004), or apsidal motion in eccentric Eclipsing Binaries (Wolf, Diethelm & Hornoch 2001; Wolf et al. 2004). While catalogue membership is nearly complete for variable stars derived from the MACHO and OGLE projects, the growth of massive data bases of variable stars at fainter magnitudes is anticipated (Paczyński 2001), largely using automated procedures in tandem with data mining (Belokurov et al. 2003). This circumstance recommends the development of a fast, reliable procedure to eliminate contaminating outliers, so they may be subject to later review, analysis, and reclassification. Developing such a procedure to find outliers in large data sets of variable stars provides the motivation for the methodology described in this line work. The method described in this paper is applicable in general to surveys where unclassified groupings exist. However, it makes sense to first try and apply the methodology in surveys where classified groups exist; this provides an independent check on the methodology for the previously classified group members, as well as the possibility to improve the robustness of the grouping. Thus, in this work, we have chosen to apply our methodology only to surveys containing previously classified sets of light curves.

This paper is organized in the following way. Section 2 is devoted to the methodology. In Section 3, we show how our method can be extended to a large number of light curves. In Section 4, we present the results from runs on MACHO and OGLE catalogues. Future work is presented in Section 5 and conclusions are in Section 6.

2 METHODOLOGY

Our main objective is to identify outliers in a data set of periodic variable stars.

We assume that the periods are known and the observational times are transformed to the phase as:[2]

$$t\prime \equiv \left\{ \frac{t - t_0}{T} \right\}, \tag{1}$$

where T is the period, t_0 is an arbitrary starting point and the symbol $\{\}$ represents the fractional (non-integer) part of the fraction $\frac{t-t_0}{T}$. These type of light curves are usually called folded light curves or folded-curves. Since all the light curves used in this paper are folded light curves, we adapt the term light curve to mean folded light curve.

The basic procedure is conceptually straightforward; compare the light curve(s) in the data set with that of every other light curve in the data set, and see which light curve(s) is least like all others. Closer scrutiny reveals some of the difficulties of this process. First, given the size of the data sets ($\sim 10^5$ for existing data sets, growing to $\sim 10^8$ in the near future), the comparison method(s) must be fast and scale favourably. Secondly, the size of the data sets also prohibits human supervision, so the methods must be automated and very robust.

Finding an outlier requires two separate comparisons. The first comparison is between two individual light curves to determine how similar, or dissimilar, they are to each other. This comparison will be described in Section 2.2. Once this comparison is done for every pair of light curves in the data set, we form a similarity matrix (see Fig. 1). Each row of the similarity matrix represents the similarity of a given light curve to all other light curves in the data set. To determine which light curve in the data set is least like all others, we compare the rows of the similarity matrix and determine which row has on average the smallest similarity with every other light curve. This second comparison is described in more detail in Section 2.3.

We begin by describing the preprocessing of the light curves, and then the actual comparison tests.

2.1 Preprocessing

There is no one-shot approach to preprocessing a data set of light curves. A smoothing technique will remove undesired noise, but could also remove true features of the light curve. An interpolation may generate a more natural looking light curve, but can also insert features

[2] In our pipeline, we have incorporated a period finder developed by Riemann (1994).

that are not physical. Sophisticated signal processing methods can be used to determine the best smoothing/interpolation/designaling method; however, this will only be true for a single light curve. Since we are dealing with a large collection of light curves and essentially we are looking for a few different light curves, using a universal preprocessing algorithm is not a sensible strategy. For these reasons, we have chosen a minimal preprocessing scheme; one that preserves the main light curve features, but does not allow obvious spikes to dominate the statistics.[3]

The steps that are described below in this section are the steps used for the analysis done in Section 4 on the MACHO and OGLE catalogues. We have, however, experimented with a number of different schemes and the resulting modules developed will be released as part of the software suite. We have concluded that while the comparisons between pairs of light curves do depend on the choice of preprocessing scheme, the measure of overall outlier does not closely depend on the choice of parameters used in the preprocessing or the preprocessing method (assuming we stay within reasonable limits).

For any measure of similarity to be meaningful, the light curves must be preprocessed to retain the true features of the data, while minimizing the effects of noise and spurious measurements. Currently, our comparison methods require the values of the light curves at predetermined, uniformly spaced, times.[4] Since we need the values of the light curves at uniformly spaced intervals, we need to interpolate the light curves. All light curves have spurious data due to noise and other effects, and many have spikes.

Any interpolation method may be adversely affected by these spikes and high-frequency noise. For this reason, we have built into our methodology a three-step spike-removal/interpolation/data-smoothing process. We first perform a running average on the light curve data (spike-removal), we then perform an interpolation to obtain the values of the light curve at prescribed times. We then perform a smoothing process on the interpolated data. This smoothing process is a generalized Savitzky–Golay (SG) smoothing (Gorry 1990).

Running average. Our running average scheme replaces the value of each data point by the average of the data points contained within a box centred on the data point. Since our data are not evenly spaced, we weigh the influence any value can have on the running average by its distance to the box centre. We use a Gaussian weight that depends on the distance from the 'current point' and has a standard deviation half the window size. The results of a running average are somewhat dependent on the width of the running window size. Since we wish to remove spikes, but not features, we determined that a width of 1 per cent of the light curve phase worked well. An extension to this method is to additionally weigh the values by the observational error using Gaussian weights. This modification turned out to be extremely useful in very large data sets where observational errors cannot be accounted for in the measure of correlation. This point will become clearer in the following sections.

Interpolation. We use simple linear interpolation in order to produce uniformly spaced light curve points. We have found that a linear interpolation, in combination with the spike-removal and the smoothing, described next, works well in practice.

Smoothing. The post-interpolation smoothing method uses a generalized SG method. SG is a well-known and widely used smoothing method (Press et al. 1992). The method we employ is generalized because it does not truncate the endpoints of the data set in the smoothing process. It does this by employing the Gram polynomials. A typical implementation of the SG smoothing algorithm is, in a sense, a running least-squares fit to the data and requires solution of a matrix equation as we march through the data. Using the recursive properties of the Gram polynomials, as in Gorry (1990), SG smoothing can be accomplished without the need to solve matrix equations.

There are two adjustable parameters in our SG smoothing: the order of the polynomials, and the width of the smoothing window. Since smoothing of the data is the principle objective of this procedure, we typically use third-order polynomials, attempting to smooth out the higher-order oscillations. The width of the smoothing window determines the range of influence a given point has over neighbouring points (the larger the window, the more neighbouring points affect the smoothed value of the current point). Since we use folded light curves, we also use periodic boundary conditions. Not wanting to 'smooth-out' any features, we determined that a width of 4 per cent of the period worked well. A review of the properties of SG filters can be found in Luo et al. (2005).

Fig. 2 shows the modifications in a given to a folded light curve as it is passed through the preprocessing steps described above. The points in the top panel shows the original light curve. The solid line in the same panel shows the light curve after the spike-removal is performed. The solid line in the second panel shows the final result after interpolation and smoothing. In the same panel, the results after spike-removal are shown for comparison. Upon inspection of Fig. 2, one will note that the differences between the initial, pre-spike-removal light curve and final smoothed light curve are perhaps not as dramatic as could be achieved, or that more smoothing could have been accomplished in the spike-removal stage. While this is true we preferred to err on the side of caution, resisting the temptation to produce very smooth light curves while being certain to preserve features within the light curve.

Note that at each preprocessing step we have estimated the errors using typical error propagation techniques (see Appendix B for details). Hence the final light curve contains observational errors that are necessary for the next stage.

2.2 Comparison of light curves

For most tests, a comparison of two light curves is a point-by-point comparison of two time-series. In this work, we have concentrated on the use of the correlation between two light curves as the measure of their 'similarity'. There are many choices of the measure of similarity and

[3] Here statistics refers to the overall outlier measure which is described in Section 2.3.

[4] Our current FFT method requires measurements uniformly spaced in time. Additionally, any time-domain comparison method would require knowing the measurements at predetermined times.

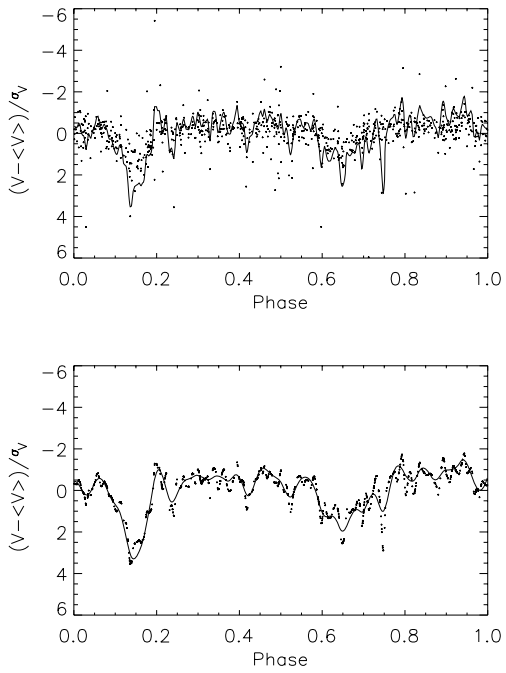

Figure 2. A light curve as it is passed through the preprocessing steps. The points in the top panel shows the original light curve. The solid line in the same panel shows the light curve after the spike-removal is performed. The solid line in the second panel shows the final result after interpolation and smoothing. In the same panel, the results after spike-removal (the solid line of the first panel) are shown for comparison as points.

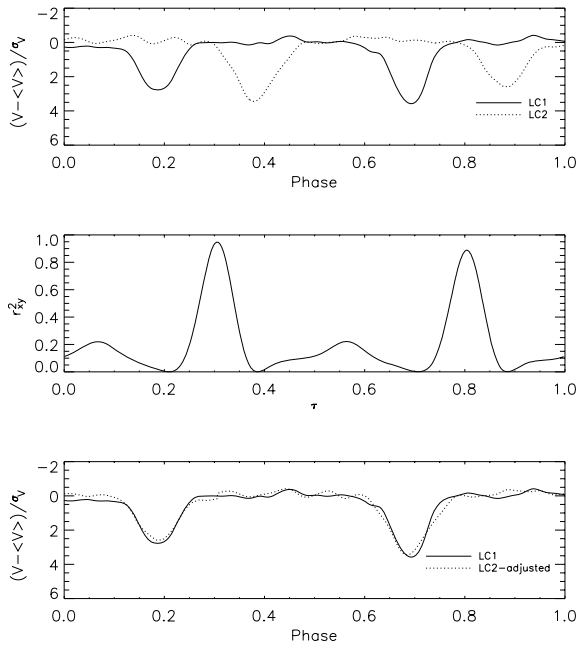

Figure 3. The top panel shows two similar light curves with arbitrary epochs after being normalized and shifted to set the mode to be 1. The middle panel shows the square of the correlation plotted as a function of the epoch. The maximum occurs at ∼0.3. Finally, the bottom panel shows the two light curves after one of the two light curves are time-shifted by ∼0.3.

depending on the 'features' of the light curves, some work better than others. Cross-correlation and chi-square tests are the simplest choices. One can show though, that the order of outliers remains the same none the less. Future work will investigate different measures of similarity.

2.2.1 Correlation coefficient of two time-series with measurement errors

The uncertainties in the flux measurements of a typical light curve can vary significantly. For this reason, any analysis based on the flux must account for the variations in the uncertainties in the flux measurements.

The goal is to derive a modified correlation coefficient r of two light curves that incorporates the errors of the measurements.

We begin by considering the 'standard' correlation coefficient (without observational errors) of two time-series $y(n)$ and $x(n)$ where n is the discrete time. For each measurement $y(n)$ and $x(n)$, there are the associated measurement errors $\sigma_y(n)$ and $\sigma_x(n)$. For the moment, we assume the means of $y(n)$ and $x(n)$ to be zero. The unweighted mean is defined as:

$$\bar{x} \equiv \frac{1}{N}\sum_n x(n), \tag{2}$$

where N is the number of observations.

We examine how well the data fit the line $y = \alpha x$. Using a least-squares fit

$$\chi^2 = \sum_n [y(n) - \alpha x(n)]^2 , \tag{3}$$

and then by taking the derivative with respect to α, we can show that the χ^2 is a minimum when

$$\alpha = \frac{\sum_n y(n)\, x(n)}{\sum_n x^2(n)}. \tag{4}$$

Performing a least-squares fit on the inverse equation $x = \beta y$, we can similarly show that

$$\beta = \frac{\sum_n y(n)\, x(n)}{\sum_n y^2(n)}. \tag{5}$$

The correlation coefficient is defined as (Weisstein 1999):

$$r_{xy} \equiv \sqrt{\alpha\beta} = \frac{\sum_n y(n)x(n)}{\sqrt{\sum_n y^2(n) \sum_n x^2(n)}}. \tag{6}$$

This is the correlation coefficient without observational errors. In the case of observational errors, fitting the linear equations $y = \alpha x$ and $x = \beta y$ using a χ^2 yields

$$\chi^2 = \sum_n \frac{[y(n) - \alpha x(n)]^2}{\sigma_y^2(n)}. \tag{7}$$

Setting the derivative with respect to α equal to zero, we can show that

$$\alpha = \frac{\sum_n y(n)x(n)/\sigma_y^2(n)}{\sum_n x^2(n)/\sigma_y^2(n)}, \tag{8}$$

and equivalently

$$\beta = \frac{\sum_n y(n)x(n)/\sigma_x^2(n)}{\sum_n y^2(n)/\sigma_x^2(n)}. \tag{9}$$

Using the above definition of the correlation coefficient, we can show

$$r_{xy} = \sqrt{\alpha\beta} = \sqrt{\frac{\sum_n y(n)x(n)/\sigma_y^2(n) \sum_n y(n)x(n)/\sigma_x^2(n)}{\sum_n x^2(n)/\sigma_y^2(n) \sum_n y^2(n)/\sigma_x^2(n)}}. \tag{10}$$

If the mean values of x and y are not zero, we can extend the above analysis by using the following transformations,

$$x'(n) \rightarrow x(n) - \bar{x}$$

$$y'(n) \rightarrow y(n) - \bar{y}$$

Substituting for the new variables in equation (10) we can show that

$$r_{xy}^2 = \frac{\sum_n \left[(y(n) - \bar{y})(x(n) - \bar{x})/\sigma_y^2(n)\right] \sum_n \left[(y(n) - \bar{y})(x(n) - \bar{x})/\sigma_x^2(n)\right]}{\sum_n \left[(x(n) - \bar{x})^2/\sigma_y^2(n)\right] \sum_n \left[(y(n) - \bar{y})^2/\sigma_x^2(n)\right]} \tag{11}$$

2.2.2 Cross-correlation in Fourier space

The comparison of two light curves using the correlation coefficient described above hinges on the choice of epoch. Since the phase of the first signal can be arbitrarily chosen, a comparison could yield a small r^2 even if two light curves are alike. Therefore, this arbitrary epoch has to be adjusted for all light curves prior to any comparison.

An obvious approach is to move the epoch of one of the two light curves until a maximum r^2 is achieved. Though conceptually simple, this approach could be quite computationally costly as it would need to be calculated for every pair of light curves. Fortunately, this can be performed quite economically in Fourier space using the convolution theorem.

The correlation between the light curves x and y with time-lag τ is given by

$$r^2_{xy}(\tau) = \sum_{n=0}^{N-1} x(n)\, y(n-\tau), \tag{12}$$

where n is the discrete time. According to the convolution theorem (see Appendix A), the correlation can be written as

$$r^2_{xy}(\tau) = \mathcal{F}^{-1}[\mathcal{X}(\nu)\,\bar{\mathcal{Y}}(\nu)](\tau) \tag{13}$$

where $\mathcal{X}(\nu)$ is the Fourier transform of $x(n)$ and $\bar{\mathcal{Y}}(\nu)$ is the complex conjugate of the Fourier transform of $y(n)$. Therefore, one can find the maximum correlation by finding the maximum of the inverse Fourier transform of the product of the Fourier transforms of the two light curves. For fast Fourier transforms (FFTs), each Fourier transform requires $2N\,\log(N)$ operations, where N is the number of observations. Thus, for each pair of light curves, a total of $6N\,\log(N)$ operations are required. This is to be compared to N^2 operations required doing the analysis in regular space.

The above equations can be extended to include measurement errors (equation 10):[5]

$$r^2_{xy}(\tau) = \frac{\boldsymbol{\mathcal{F}^{-1}}\left[\boldsymbol{\mathcal{F}}(y(n)/\sigma_y^2(n))\bar{\boldsymbol{\mathcal{F}}}(x(n))\right](\tau)\ \ \boldsymbol{\mathcal{F}^{-1}}\left[\boldsymbol{\mathcal{F}}(y(n))\bar{\boldsymbol{\mathcal{F}}}(x(n)/\sigma_x^2(n))\right](\tau)}{\boldsymbol{\mathcal{F}^{-1}}\left[\boldsymbol{\mathcal{F}}(1/\sigma_y^2(n))\bar{\boldsymbol{\mathcal{F}}}(x^2(n))\right](\tau)\ \ \boldsymbol{\mathcal{F}^{-1}}\left[\boldsymbol{\mathcal{F}}(1/\sigma_x^2(n))\bar{\boldsymbol{\mathcal{F}}}(y^2(n))\right](\tau)} \tag{14}$$

The top panel in Fig. 3 shows two light curves with arbitrary epochs. The middle panel shows the square of the correlation as a function of the time-lag, $r^2_{xy}(\tau)$. The maximum occurs at $\tau \approx 0.3$, calculated using equation (14). The bottom panel shows the same light curves after one of the light curves is time-shifted by 0.3.

2.3 Outlier measure

Once we have completed the comparisons of each pair of light curves, thus populating the similarity matrix, we compare the rows of the similarity matrix to determine the outliers. For each line in the similarity matrix, we compute the *outlier measure* as the average of the correlations:

$$\mathcal{R}^2_x = \frac{1}{N_{\rm LC}-1}\sum_{y\neq x} r^2_{xy}, \tag{15}$$

where y runs over all light curves in the set except for x and $N_{\rm LC}$ is the number of light curves.

For each light curve, we calculate the average of the correlations as above and then we rank this measure. The light curves with the lowest correlation are classified as probable outliers and are further inspected.

How many light curves should be inspected? A natural choice is to set a threshold based on the actual value of the average correlation ($\mathcal{R}$). For example, we could set the threshold at $\mathcal{R} = 0.3$; thus, any light curve below that value should be examined. Yet, this is not exactly what we are looking for. Consider the following scenario: a catalogue consists of light curves which are all alike (e.g. a collection of well-separated Eclipsing Binary stars (EBs) with circular orbits and components that are both O stars). The light curves of this collection will be naturally strongly correlated. If one of the objects in the catalogue is a binary system with one of the stars being a B star, then the correlation to the rest of the light curves will be slightly lower, but not low in absolute terms. Nevertheless, that light curve will have the lowest $\mathcal{R}$ in the set, and thus should be flagged for additional inspection. The same may apply to a collection of light curves that are classified together, but their light curves show weak correlation (this is an indication that the band in which the observations were made is not the primary manifestation of the physical classification) and, therefore, a low correlation does not necessarily mean that the particular light curve is an outlier. Hence, what really matters is the average correlation, $\mathcal{R}$, compared to the rest of the values of $\mathcal{R}$ in the set.

One could calculate the expectation value and variance of the distribution of the values of $\mathcal{R}$ and determine which light curves are at least 2σ away from the mean. This would have been a reasonable approach assuming that the underling distribution was Gaussian. Unfortunately, this is not true in general. First, consider the case that all pairs of light curves have the same correlation, λ. The probability density function (PDF) of the correlations of this set would be a bivariate normal distribution (which at large N becomes a Gaussian). In reality, our sets of light curves do not all have the same correlations. For simplicity, we assume that the light curves could be grouped in clusters with constant correlations and, then, the resulting PDF will be a superposition of bivariate normal distributions each with different λ. Therefore, the final PDF is data set dependent and may or may not resemble a Gaussian. For these reasons, the average of the values of $\mathcal{R}$ and its variance has proven not to be a reliable approach.[6] Consequently, we have concluded that simply selecting the light curves with lowest average correlation (of the order of 5 per cent) is the fastest and the most reliable approach.

3 LARGE DATA SETS

The numerical method shown in Section 2 works well to identify outliers. This will be demonstrated in Section 4 where outliers are identified in real light curve catalogues.

For a data set containing ∼5000 light curves, the run time on a typical desktop (3 GHz Intel® Xeon™) is ∼5 h.

[5] Finding the correlation of equation (11) will only require shifting the zeroth component of the Fourier transforms.

[6] We have tested the above empirically and we found that in most cases the resulting PDFs are not invariant.

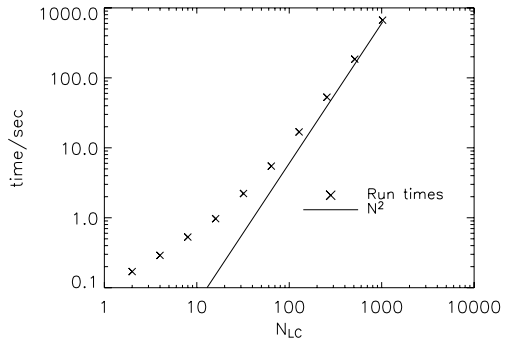

Figure 4. Time in seconds to complete the analysis as a function of number of light curves on log–log scale. The points correspond to the actual computational times, while the solid line corresponds to the N^2 relation. It is clear that for large values of N the computational time scales as N^2.

The real advantage of a method like this would lie in the ability to perform the analysis on much larger data sets. Unfortunately, our method scales as $N_{\rm LC}^2$ per star, where $N_{\rm LC}$ is the number of light curves. Fig. 4 gives a graphical representation of the performance of our model. It shows running times, in seconds, as a function of $N_{\rm LC}$ in a log–log scale. Superimposed on this plot is a curve that is proportional to $N_{\rm LC}^2$. For large $N_{\rm LC}$, we see our algorithm scales as $N_{\rm LC}^2$. Accordingly, for a data set of $\sim 10^5$ light curves, the analysis will take about 50 d! The program could be executed in parallel, and thus reducing the computational time by a factor of $n_{\rm CPU}$ ($n_{\rm CPU}$ being the number of CPUs). However, the data sets will soon grow to 10^6, and thus requiring a few thousands of CPUs in order to run the analysis in few days. Consequently, we must craft alternative, smarter algorithms to deal with larger data sets.

In the following subsections, we present alternative approaches to speed up the calculation, each one having advantages and disadvantages. In Section 3.1, we show how, in the case of a simple correlation coefficient (without the observation errors), the analysis in discovering outliers can be reduced from $N_{\rm LC}^2$ operations to $N_{\rm LC}$ operations. In Section 3.3, we will show a simple approximation that can be applied to the extended correlation coefficient in equation (14) (including observational errors and allowing time-lag to vary).

3.1 Simple correlation coefficient

The correlation coefficient between the two light curves i and j is given by

$$r_{ij} = \frac{\sum_t (f_i(t) - \bar{f}_i)(f_j(t) - \bar{f}_j)}{(N-1)\,\sigma_i \sigma_j} \tag{16}$$

where N is the number of observations. To identify outliers, we calculated the average correlation of each light curve with the rest of the set (see Section 2.3). This average correlation is given by

$$\begin{aligned} R_i &= \frac{1}{N_{\rm LC}-1}\left(\sum_j r_{ij} - 1\right) \\ &= \frac{1}{N_{\rm LC}-1}\left(\sum_j \left[\frac{\sum_t (f_i(t) - \bar{f}_i)(f_j(t) - \bar{f}_j)}{(N-1)\,\sigma_i\,\sigma_j}\right] - 1\right), \end{aligned} \tag{17}$$

where we sum over all j's and subtract 1 for the $i = j$ case.

Re-arranging the order of the sums, we get[7]

$$R_i = \frac{1}{N_{\rm LC}-1}\left(\sum_t \left[\frac{(f_i(t) - \bar{f}_i)}{(N-1)\,\sigma_i} \sum_j \frac{(f_j(t) - \bar{f}_j)}{\sigma_j}\right] - 1\right) \tag{18}$$

We define a centroid light curve as:

$$F(t) \equiv \frac{1}{N_{\rm LC}} \sum_j \frac{f_j(t)}{\sigma_j}, \tag{19}$$

and its average centroid light curve

$$\bar{F} = \frac{1}{N}\sum_t F(t) = \frac{1}{N_{\rm LC}} \sum_j \frac{\bar{f}_j}{\sigma_j}. \tag{20}$$

[7] Here we are making the assumption that all light curves have same t's. This is not true in general, but it is true after proper interpolation-something we performed in the preprocessing steps.

Substituting the definitions of F and $\bar{F}$ into equation (18), we get

$$R_i = \frac{N_{\rm LC}}{N_{\rm LC}-1}\frac{\sum_t (f_i(t)-\bar{f}_i)(F(t)-\bar{F})}{(N-1)\,\sigma_i} - \frac{1}{N_{\rm LC}-1}. \tag{21}$$

Note that at the limit where $N_{\rm LC} \gg 1$, $\frac{N_{\rm LC}}{N_{\rm LC}-1} \to 1$ and $\frac{1}{N_{\rm LC}-1} \to 0$ and, therefore,

$$R_i = \frac{\sum_t (f_i(t)-\bar{f}_i)(F(t)-\bar{F})}{(N-1)\,\sigma_i}. \tag{22}$$

Since $F(t)$ and $\bar{F}$ need to be calculated only once, the number of operations necessary to find all the values of R_i is $O(N_{\rm LC} + N \times N_{\rm LC}) \sim O(N \times N_{\rm LC})$ which is a significant improvement over the $O(N \times N_{\rm LC}^2)$ which was necessary before.

This gain does not come without disadvantages. First, note that we *cannot* apply the same transformation from 'average-of-the-correlations' to 'correlation-to-the-average' in the case of correlation coefficients using observational errors, since in this case the magnitudes and the errors are mixed. Nevertheless, this is not a major disadvantage, since the observational errors can be partially taken into account in the averaging/smoothing operations. The second major shortcoming is the fact that the time-lag cannot be considered as a free parameter. This is because the time-lag depends on both light curves and, thus, $F(t)$ cannot be calculated once for all light curves. To circumvent this problem, we need to find a priori an *absolute phase for all light curves*.

3.2 Universal phasing

To do just that, we have devised the following algorithm of adjusting the epoch of all light curves using clustering methods. The basic concept is to find where the signal with the highest/lowest magnitude dip occurs for each light curve and set it to a particular phase by time-shifting the folded light curve. Since the data are noisy, it will not be practical to just finding the maximum/minimum value of the magnitude. On the contrary, we must find a statistical measure of the signal.

Our method is based on a clustering technique that divides the data (here data refers to a single light curve) into clusters (cluster here means a subset of observations within a light curve) based on the magnitude, and then finds the cluster with the maximum average.

To find the clusters, we required that both the density within the clusters and the separation between clusters should be maximum. In other words, we want the clusters to be as compact as possible and be as separated from other clusters as possible.

We measure the cluster compactness or intercluster measure of two clusters as

$$S_{\rm inter} \equiv \sum_{t_i \in {\rm C1}} (t_i - \bar{t}_{\rm C1})^2 + \sum_{t_i \in {\rm C2}} (t_i - \bar{t}_{\rm C2})^2\,, \tag{23}$$

where $C1$ and $C2$ denote the clusters, t_i are the times of observations in the particular cluster and $\bar{t}_{\rm C1}$ is the average time in cluster $C1$. We also define the intradistance between the two clusters as

$$S_{\rm intra} \equiv \frac{|\bar{t}_{\rm C1} - \bar{t}_{\rm C2}|}{\sqrt{\sigma_{\rm C1}^2/N_{\rm C1} + \sigma_{\rm C2}^2/N_{\rm C2}}}, \tag{24}$$

where $N_{\rm C1}$ is the number of points in cluster $C1$ and

$$\sigma_{\rm C1}^2 = \frac{1}{(N_{\rm C1}-1)} \sum_{t_i \in {\rm C1}} (t_i - \bar{t}_{\rm C1})^2\,. \tag{25}$$

We define the following measure which by minimizing gives us a measure of goodness of clustering:

$$S \equiv \frac{S_{\rm inter}}{S_{\rm intra}}. \tag{26}$$

The actual algorithm is described below.

(i) For each light curve, we select the highest/lowest 10 per cent magnitude data points.

(ii) We divide the data in two clusters as $t_{\rm C1} \in \{t_1, t_2, \ldots, t_s\}$ and $t_{\rm C2} \in \{t_{s+1}, t_{s+2}, \ldots, t_N\}$ where s is the index of the separator.

(iii) For each $s = \{1 \ldots N\}$, we calculate the goodness of clustering using equation (26). If S is minimum within the range $1 < s < N$, we keep the division of data into two clusters. We repeat this process in the subclusters until no more clustering is favourable.[8]

(iv) After the clustering is done, we calculate the mean magnitude and mean time in each cluster. We select the cluster of the highest mean magnitude.

(v) We translate time such as the mean time of the selected cluster is always at the same predefined time.

By phasing every light curve to a universal phase, the method of 'correlation-to-the-average' can be applied assuming that the observational errors are incorporated in the running average method. However, the method is an approximation, since it does not guarantee that the correlation between two light curves is maximum. Nevertheless, for most light curves where a maximum/minimum signal is well defined, this method

[8] Since the data are in a dimensional space, it is guaranteed that points in the same cluster are sequential. Therefore, a separation at a given iteration cannot alter the clustering measure of the previous iteration.

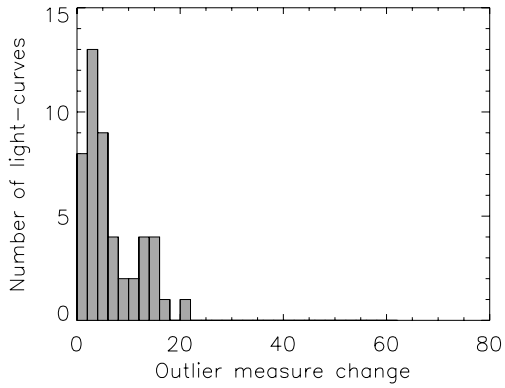

Figure 5. Histogram of the outlier measure difference between the full method and the approximate method for 500 EBs from the OGLE EB catalogue. Only the 10 per cent with the lowest average correlation were used.

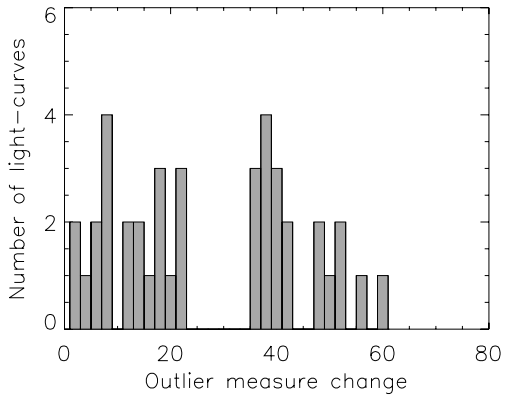

Figure 6. Histogram of the outlier measure difference between the full method and the approximate method for 1000 RR Lyræs from the OGLE RR Lyræ catalogue. Only the 10 per cent with the lowest average correlation were used.

should give us very similar results to the full method. We have tested this method on two sets; 500 light curves of OGLE EBs and 1000 light curves of OGLE RR Lyræs. Figs 5 and 6 show the runs on these two sets. In each figure, we show a histogram of the the rank differences between the full method and the approximation described in this section for the bottom 10 per cent of the light curves. EBs do have a much better defined minimum, so the approximation performs very well (most light curves are ranked with ± 10 of the original rank), whereas in the case of RR Lyræs, the approximation is not performing as well.

3.3 Outlier analysis within subsets

Another alternative approach which avoids the drawbacks of the method described above is based on a simple statistical argument. If a light curve is an outlier in the whole set, it will be an outlier in a large subset of the whole set. We could then, in principle, divide the whole set into large subsets and perform the analysis on each subset. If the subsets are randomly selected and the number of light curves is large enough, the outlier measure from each subset can be put together and, hence, we can rank all light curves as if they were in a single set.

Since each subset must be a substantial fraction of the full set ($\geqslant$10 per cent), the overall performance gain is about a factor of 10 at the best. In the case of large sets, this method will not scale very favourably, but it is an 'exact' method and it is very easily parallelizable.

Table 1. Main features of the catalogues used.

Type of variable star	Found by	Number of stars
Cepheids	MACHO	3177[a]
Cepheids	OGLE-II	1329[a]
RRLyræ	MACHO	16 020[a]
RRLyræ	OGLE-II	5327[a]
EBs	MACHO	6064[a,b]
EBs	OGLE-II	2580[a]

[a]Only light curves with at least 100 observations were included.
[b]EBs from both the LMC and SMC were included.

We have applied this method to 16 020 of the RR Lyræs from the MACHO survey (see Section 4).

4 RESULTS

We tested the validity of our method on various periodic star catalogues, both published and unpublished, compiled by the MACHO collaboration (Alcock et al. 2000)[9] and by the OGLE collaboration (Udalski, Kubiak & Szymański 1997).[10]

Both the MACHO and the OGLE projects were microlensing surveys devoted to finding gravitational microlensing events in the halo of the Milky Way by background stars in the Large and Small Magellanic Cloud (LMC and SMC) and the bulge of the Milky Way. These surveys also produced large catalogues of variable stars: details on the MACHO variable star research can be found in Alcock et al. (1995, 1996a,b, 1997a) and Cook et al. (1995). OGLE variable star catalogues found during part II of the project (OGLE-II), (Udalski et al. 1997) with accompanying papers, can be found on the group website (Udalski et al. 1999b,c; Soszyński et al. 2003; Wyrzykowski et al. 2003).

The variable stars considered were EBs of which catalogues were published by MACHO (Alcock et al. 1997b; Faccioli et al. 2005) and OGLE (Wyrzykowski et al. 2003), RR Lyræ and Cepheids from OGLE (Udalski et al. 1999b) and unpublished MACHO collections that were compiled at the Lawrence Livermore National Laboratory (LLNL) by Kem Cook, Doug Welch and Gabe Prochter. These lists have been generated from the MACHO data base by appropriate cuts in the period–luminosity diagram. This is only a first step in producing a catalogue and, thus, the resulting lists are expected to be contaminated.

MACHO observations were taken in two non-standard bandpasses: MACHO 'blue', hereafter indicated as $V_{\rm MACHO}$, with a bandpass of 440–590 nm and MACHO 'red', hereafter indicated as $R_{\rm MACHO}$, with a bandpass of 590–780 nm; transformations to standard Johnson V and Cousins R bands are described in detail in (Alcock et al. 1999).[11]

The average number of observations in both bands is several hundreds, with the centre of the LMC being observed more frequently than the periphery.

MACHO periods were found by applying the SUPERSMOOTHER algorithm (Reimann 1994), first published by Friedman (1984). The algorithm folds the light curve around different trial periods and selects the one that gives the smoothest folded light curve. Periods were found for the red and the blue bands separately, and usually agree with each other to better than 1 per cent. The algorithm may fail though, usually determining a period for one colour band that is a multiple of the period found for the other band. In these cases, the light curve with the incorrect period will often be flagged as an outlier; hence, the program can be useful in finding wrong periods in a large data set of variable stars (see MACHO Cepheids and RR Lyræs below).

OGLE observations were taken in the B, V and I bands and reduced via Difference Image Analysis (DIA) (Żebruń et al. 2001a); a catalogue of variable stars for the Magellanic Clouds was thus produced (Żebruń, Soszyński & Wożniak 2001b) and from it a sample of 2580 EBs was selected (Wyrzykowski et al. 2003); we used only I-band, DIA-reduced observations in our analysis, since the number of observations in this band was much higher (of the order of $\approx$200–300) compared to V and B.

The main features of the MACHO and OGLE variable star data sets are summarized in Table 1.

Results of these runs are presented in the following way: for each of the collections listed in Table 1, there are three figures and one table. The first figure shows the histogram of the outlier measure. The second figure shows the centroid light curve as defined in equation (19). The next nine-panel figure presents the lowest nine light curves, that is, our outliers. Each panel is labelled according to its position in the figure; from A1 to C3. Following that there is a table which summarizes the properties of these outliers including our interpretation. These interpretations were formed after further investigation including cross-correlation with other surveys, position in the Hertzsprung–Russell diagram, spectral types where available, etc.

Cepheids. Cepheids are periodic variables with periods ranging from about 1 d to about 50 d (with few extreme examples of 200 d) and which lie between the main-sequence and the giant stars. Detailed characteristics of their light curves varied depending on the period (Hertzprung progression). More details about Cepheids and other variable stars in general can be found in Petit (1987), Sterken & Jaschek (1996) and Richter, Wenzel & Hoffmeister (1985).

The MACHO Cepheid data set (Table 2) contains a small number of light curves where the folded period is an integer multiple of the 'correct' period. This can be seen in Fig. 9(A1), (A2), (B3), (C2) and (C3). Also there is a second bump in the histogram of average correlations (Fig. 7) at about 0.1. These light curves are mostly light curves folded with integer multiple period of the 'true' period. Notwithstanding, the light curve shown in A3 in the same figure is clearly an EB and not a Cepheid. B1 is evidently a periodic light curve (apparent from the distinct pattern in the folded light curve), but the shape in both R and V bands (only R shown) does not match that of a Cepheid (or all subtypes; Fig. 8). Further investigation (e.g. spectral type) is needed to determine the type of variable. Note our goal in this work is to identify the outliers, and thus demonstrate that this method can lead us to a few interesting cases. It is not our intention to do an in-depth investigation for

[9] http://www.macho.mcmaster.ca/.
[10] http://sirius.astrouw.edu.pl/ogle/.
[11] Transformation to standard magnitudes is given for the LMC by (Kem Cook, private communication):

$$V = V_{\rm MACHO} + 24.22 - 0.1804(V_{\rm MACHO} - R_{\rm MACHO})$$

$$R = R_{\rm MACHO} + 23.98 + 0.1825(V_{\rm MACHO} - R_{\rm MACHO}).$$

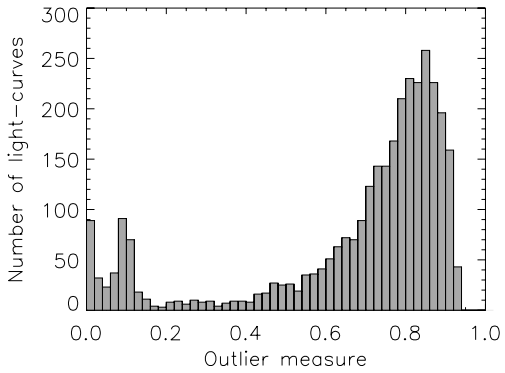

Figure 7. The histogram of the outlier measure for 3297 Cepheids in the MACHO sample.

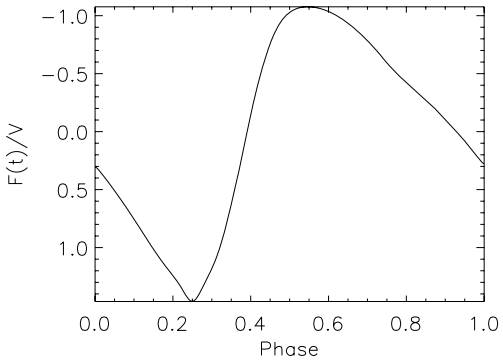

Figure 8. The centroid light curve for 3297 Cepheids in the MACHO sample.

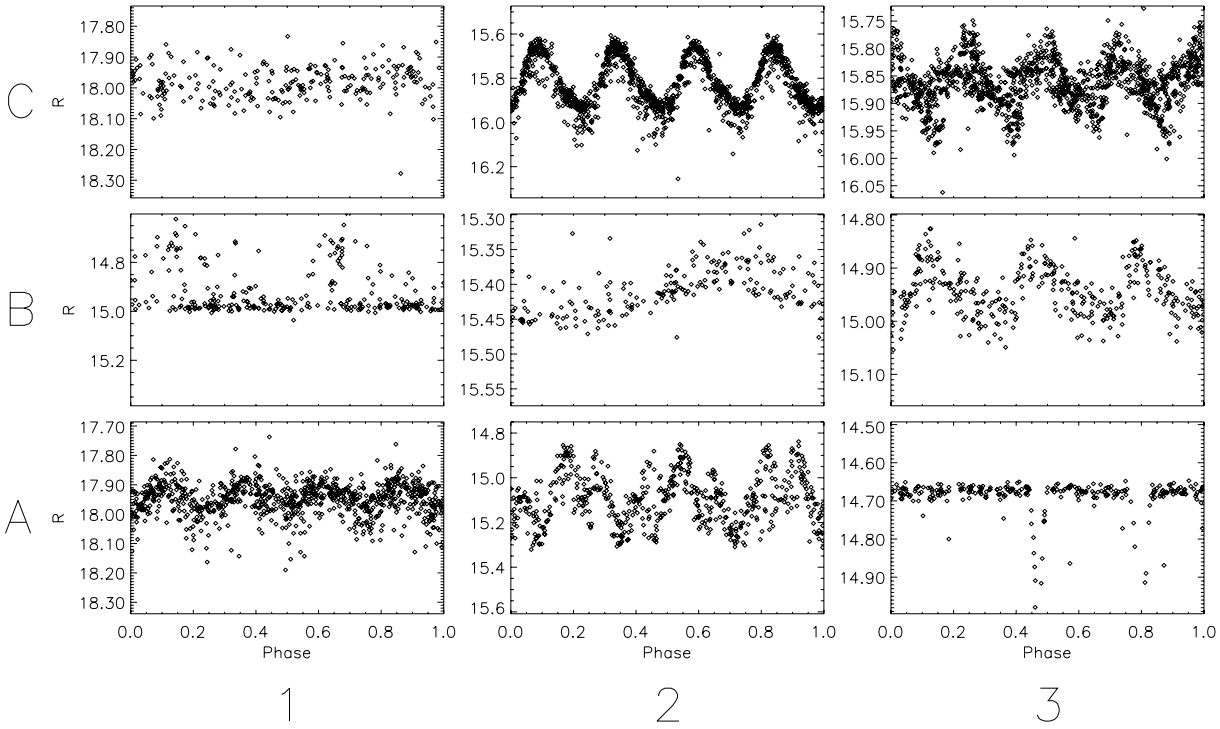

Figure 9. Light curves with the lowest measure of similarity from the MACHO Cepheid data set. Only the RED band is shown.

Table 2. MACHO Cepheid outliers.

Survey	Type	ID	Plot COORD	Period (d)	Days of observation	Number of observations	Interpretation
MACHO	Ceph	14.9223.221	A1	242.498 01	2721.04	883	Multiple period.
MACHO	Ceph	9.4511.14	A2	12.191 21	2720.82	595	Multiple period.
MACHO	Ceph	4.7459.14	A3	6.854 45	2718.04	278	**EB.**
MACHO	Ceph	60.7467.9	B1	2.001 74	2717.75	273	**Periodic, but unlikely to be Cepheid.**
MACHO	Ceph	81.9490.26	B2	1.145 35	2715.86	204	**Blue band suggests an EB.**
MACHO	Ceph	61.8562.27	B3	7.333 85	2715.84	366	Multiple period.
MACHO	Ceph	20.4309.2977	C1	0.707 94	2715.71	241	**Not periodic/variable.**
MACHO	Ceph	77.7067.41	C2	8.005 20	2709.83	1333	Multiple period.
MACHO	Ceph	79.4659.3452	C3	6.966 15	2708.91	1352	Multiple period.

each unidentified light curve, but only to point out the obvious misclassifications and interesting cases. The light curve shown in C1 does not look periodic or variable for that matter in both bands and, thus, we classify it as 'likely not periodic' star.

The OGLE Cepheid catalogue (Udalski et al. 1999b) has few true outliers (Table 3). Only three interesting cases did make it into our list (see Figs 10 and 12). From the histogram in Fig. 10, we see that there is no second bump, but three light curves are clearly on the lowest bin. A1 is vaguely a periodic light curve, but there are not enough data and they are too noisy. Even if we agree to the periodicity, we find that the asymmetry is atypical of Cepheids of all types, with slow rise and fast decline. Similarly, A2 exhibits a clear periodic signal, but wrong asymmetry. The light curve A3 is interesting. The overall shape, period and colour are consistent with a Cepheid. The extra regularly spaced spikes are too regular in folded space to be ignored. The possibility to be an EB with Cepheid variable is highly unlikely, since the periods are synchronized (1:5) which suggest some fundamental dynamical process. A more careful study is needed to understand the physical process underlying this light curve. The rest of the outliers have much higher average outlier measure and they are only shown here for consistency (nine light curves per catalogue).

RR Lyræs. RR Lyræs come in many different types, but most predominately in two subclasses: the majority are in RRAB which are asymmetric and RRC which have a sinusoidal shape light curves. These pulsating stars have very well-defined period (0.5–0.3 d). It is usually hard to distinguish them from Cepheids just from the characteristics of the shape of the light curves. More details about RR Lyræ stars can be found in Petit (1987) and Sterken & Jaschek (1996).

The published OGLE catalogue (Udalski et al. 1999b) is 'cleaner' (does not contain the wrong types or wrongly stated period of variables) than the unpublished MACHO collection (Tables 4 and 5). This can be seen from Figs 13 and 16 where it is clear that the correlation distribution of the MACHO data set is centred closer to zero than the distribution of the OGLE catalogue (this is due to contamination of the MACHO data set with other variable stars).

As in the case of Cepheids, the RR Lyræ MACHO data set contains light curves that are either folded using a multiple of the true period or folded simply with the wrong period in one of the two bands, and thus appear to be outliers. Nevertheless, some of the light curves were most likely misclassified as RR Lyræs. The light curves A2 and A3 in Fig. 15 have periods of 0.98 and 0.53 d, respectively, which are too large to belong to RRC group. Such periods can be from the RRAB group, but the shape, amplitude and symmetry of the light curves indicate a non-periodic light curve; hence, we ruled them as possibly misclassified. A1 was identified as the outlier of greatest degree. However, when we looked at the V-band light curve, it had the characteristics (period, amplitude, etc.) of an RRC. The light curves B1, B2 and C1 were simply folded with the wrong period in the red band. Looking in the V band, the periods were more in accordance to RRC group and the shape, amplitude characteristics are in accordance with that.

In the OGLE RR Lyræ catalogue, we identified three light curves that likely do not belong to this catalogue. The light curve A1 of Fig. 18 does not look periodic and the quoted period and amplitude do not correspond to a typical RR Lyræ. The light curve C1 has quoted period of 0.86 d and amplitude of <0.1 in the I band and hard to make out signal. C3 is a light curve that has period of 0.55 d, and thus most likely belonging to RRAB group, but the light curve is very symmetric, and thus belonging to the RRC group. This is one of the light curves on which further investigation should be performed.

EBs. EBs are not due to physical variation but, rather, due to occultation: one member of the pair of stars passes in front of the other.

MACHO EB catalogues are submitted for publication in Faccioli et al. (2005). We used the method presented in this paper to help free the submitted catalogues from outliers (Table 6). We found few cases of outliers that are shown here, but will not be in the final published catalogues. These are the light curves shown in Fig. 21(A1), (A2) and (A3) where all the three light curves have a symmetric single occultation and periods consistent more with RR Lyræs rather than EBs. The light curve in B1 shows no periodicity, however, after examining the V band we were convinced that it is a true EB. The light curve shown in C3 shows a very noisy light curve, but after cross-correlating with the OGLE catalogue we established that is a proper EB.

In the OGLE EBs catalogue most outliers are EBs with very eccentric orbits, and thus appear as outliers, since the second minimum will rarely be aligned with the second minimum of the rest of the light curves (Table 7). However, the light curve shown in panel C2 is not a typical EB. There is either a third body present in the system producing a second occultation or some form of atmospheric variation in one of the stars is synchronized with the binary system. Perhaps there is a large reflection effect. This occurs when the side of the dimmer star that is facing the Earth is illuminated by the brighter companion star, and thus increasing the luminosity of the system (Pollacco & Bell 1993). This effect also includes radiative brightening. For example, the system could be a small hot star with a much cooler subgiant or giant component. This light curve warrants further investigation.

The reason why the algorithm identifies highly eccentric EBs as outliers is well understood. At the same time, it is well understood that this is an indication that cross-correlation may not be the best choice of similarity measure. In cases like these a different measure of similarity must be employed. These and other potential extensions will be investigated in future works.

5 FUTURE WORK

This paper is not intended to study all possible methods for finding outliers in data sets of light curves but, rather, to help demonstrate and hopefully convince others how an automatic method like this can be applied to facilitate the discovery of new, interesting variable objects. Special emphasis should be given to the choice of measure of similarity. An attempt to study this issue will be made in a second paper where we will study how to employ more than one measure of similarity.

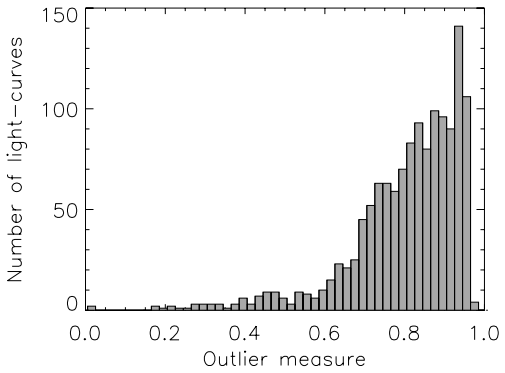

Figure 10. The histogram of the outlier measure for 1329 Cepheids in the OGLE sample.

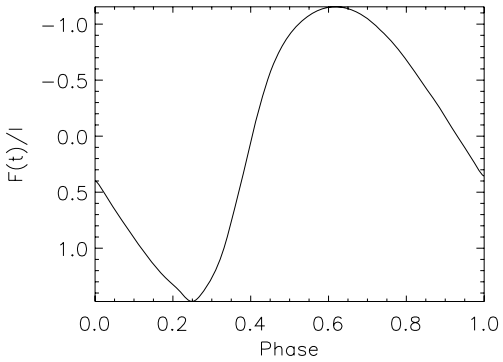

Figure 11. The centroid light curve for 1329 Cepheids in the OGLE sample.

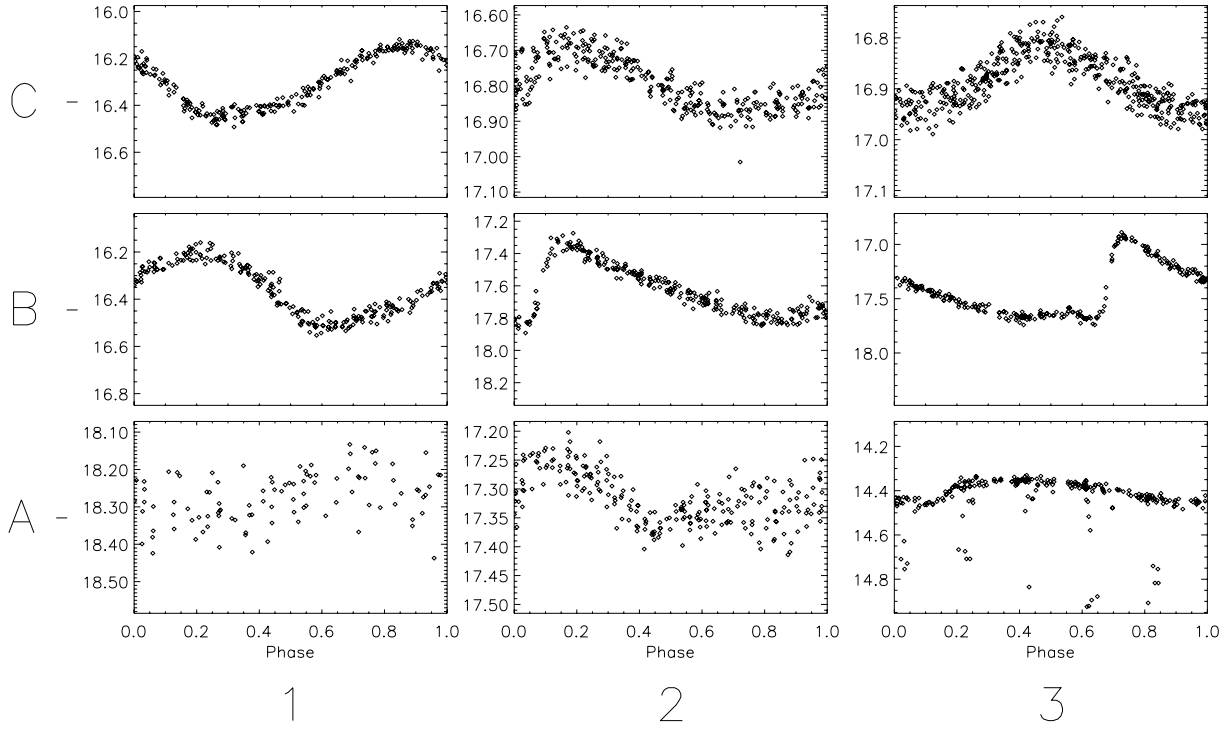

Figure 12. Light curves with the lowest measure of similarity from the OGLE Cepheid catalogue. Only the OGLE *I* band is shown.

Table 3. OGLE Cepheid outliers.

Survey	Type	Field	Number	Plot COORD	Period (d)	Days of observation	Number of observations	Interpretation
OGLE-II	Ceph	17	701 23	A1	28.966 83	1419	105	**Not enough/noisy data.**
OGLE-II	Ceph	13	184 117	A2	13.640 83	1419	245	**Atypical asymmetry.**
OGLE-II	Ceph	21	408 76	A3	4.973 38	1418	248	**Needs further study.**
OGLE-II	Ceph	17	221 134	B1	11.228 65	1418	243	
OGLE-II	Ceph	21	119 037	B2	0.878 13	1417	264	
OGLE-II	Ceph	14	114 046	B3	0.9094	1415	238	
OGLE-II	Ceph	18	185 847	C1	12.200 18	1414	244	
OGLE-II	Ceph	4	168 269	C2	0.729 23	1417	327	
OGLE-II	Ceph	4	427 313	C3	0.674 13	1414	454	

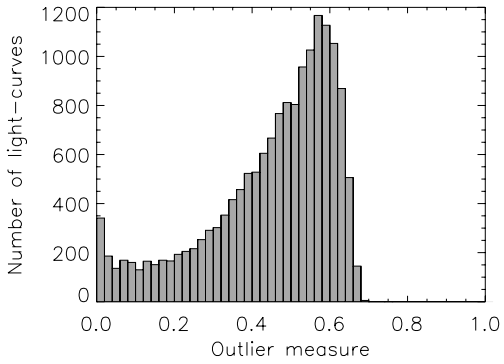

Figure 13. The histogram of the outlier measure for 16 080 RRL in the MACHO sample.

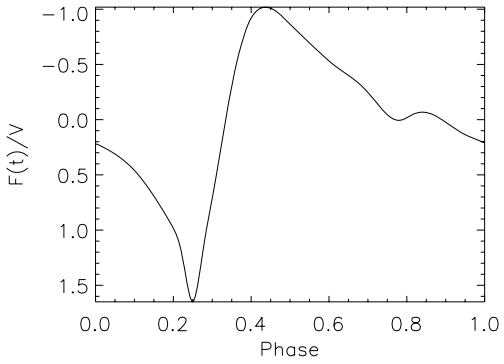

Figure 14. The centroid light curve for 16080 RRL in the MACHO sample.

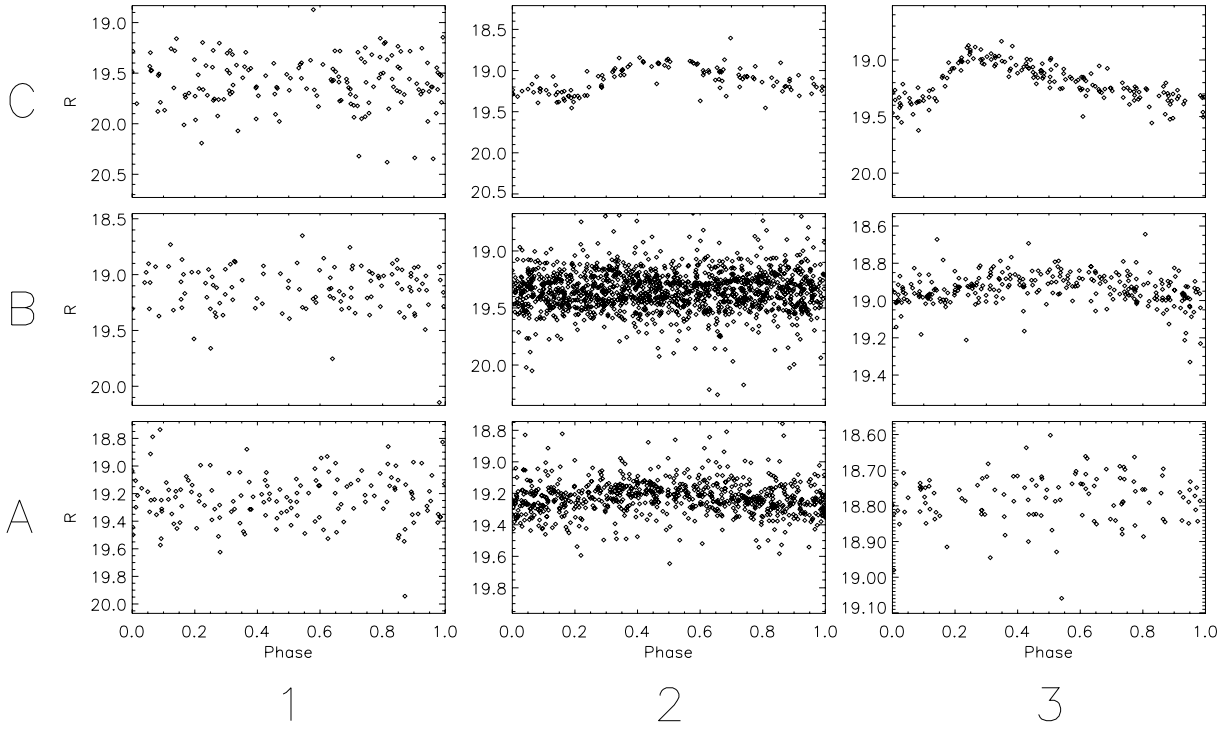

Figure 15. Light curves with the lowest measure of similarity from the MACHO RRL data set. Only the *R* band is shown.

Table 4. MACHO RRL outliers.

Survey	Type	ID	Plot COORD	Period (d)	Days of observation	Number of observations	Interpretation
MACHO	RRL	48.2992.463	A1	0.118 28	2720.99	151	RRC.Incorrect period in *R* band. $P_V = 0.35484$.
MACHO	RRL	82.8772.705	A2	0.980 11	2717.79	747	**Unlike RRL.**
MACHO	RRL	73.13488.41	A3	0.534 64	2716.77	121	**Unlike RRL.**
MACHO	RRL	76.10942.176	B1	0.467 59	2714.75	121	RRC. Incorrect period in *R* band. $P_V = 0.350\,42$.
MACHO	RRL	79.5499.2627	B2	0.254 65	2710.73	1387	RRC. Incorrect period in *R* band. $P_V = 0.337\,56$.
MACHO	RRL	67.10489.79	B3	0.297 36	2707.9	273	
MACHO	RRL	34.9080.261	C1	0.307 95	2700.79	156	RRC. Incorrect period in *R* band. $P_V = 0.461\,93$.
MACHO	RRL	37.6316.471	C2	0.620 69	2697.96	125	
MACHO	RRL	49.6623.336	C3	0.620 01	2689.9	178	

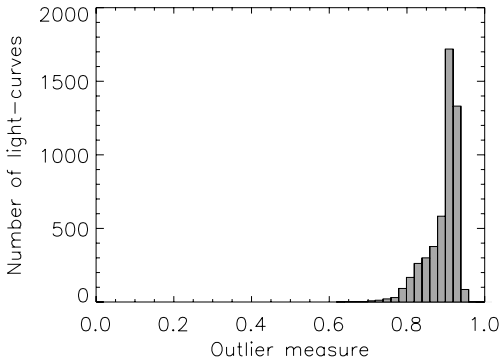

Figure 16. The histogram of the outlier measure for RRL in the OGLE sample.

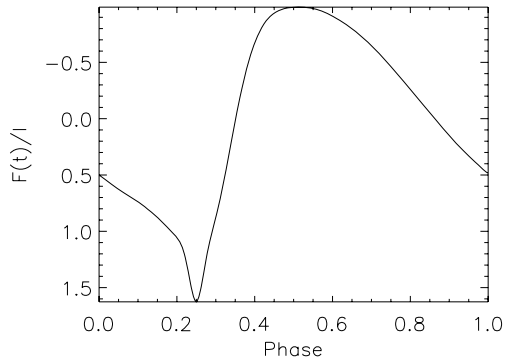

Figure 17. The centroid light curve for 5327 RRLs in the OGLE sample.

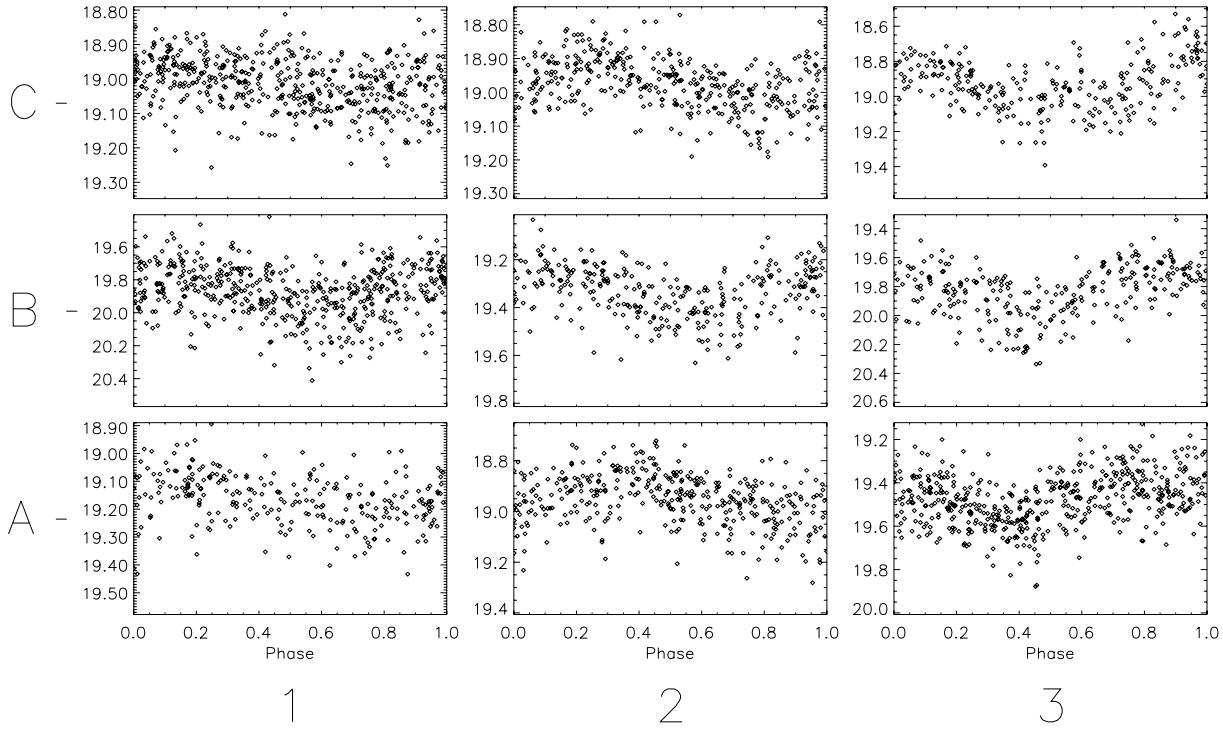

Figure 18. Light curves with the lowest measure of similarity from the OGLE RRL catalogue. Only the OGLE *I* band is shown.

Table 5. OGLE RRL outliers.

Survey	Type	ID(RA-Dec.)	Plot COORD	Period (d)	Days of observation	Number of observations	Interpretation
OGLE	RRL	053803.42-695656.4	A1	0.332 3824	1420	267	**Unknown; noisy data.**
OGLE	RRL	053325.94-701109.8	A2	0.287 6012	1420	371	
OGLE	RRL	052447.86-694319.0	A3	0.258 5634	1420	495	
OGLE	RRL	052436.03-694541.8	B1	0.223 2339	1420	504	
OGLE	RRL	053525.67-702210.2	B2	0.216 4212	1420	298	
OGLE	RRL	054036.89-701424.8	B3	0.236 1195	1420	268	
OGLE	RRL	052219.98-691907.1	C1	0.861 6601	1420	503	**Unknown.** **RRAB period, but amplitude too small.**
OGLE	RRL	053241.91-702718.9	C2	0.274 9622	1420	373	
OGLE	RRL	054609.21-702316.7	C3	0.549 4448	1420	263	**Unknown.** **RRAB period, but symmetric.**

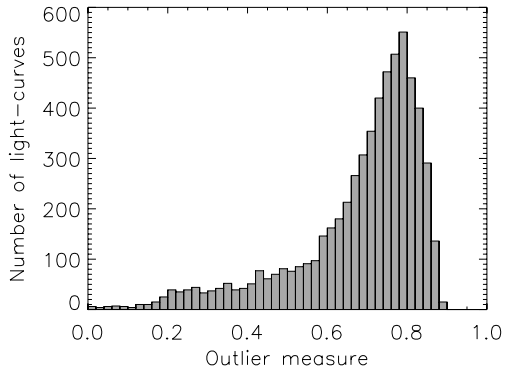

Figure 19. The histogram of the outlier measure for 6064 EBs in the MACHO sample.

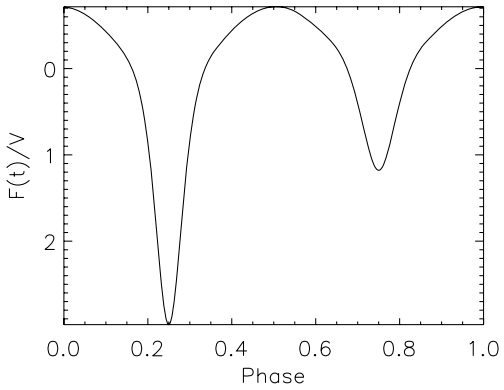

Figure 20. The centroid light curve for 6064 EBs in the MACHO sample.

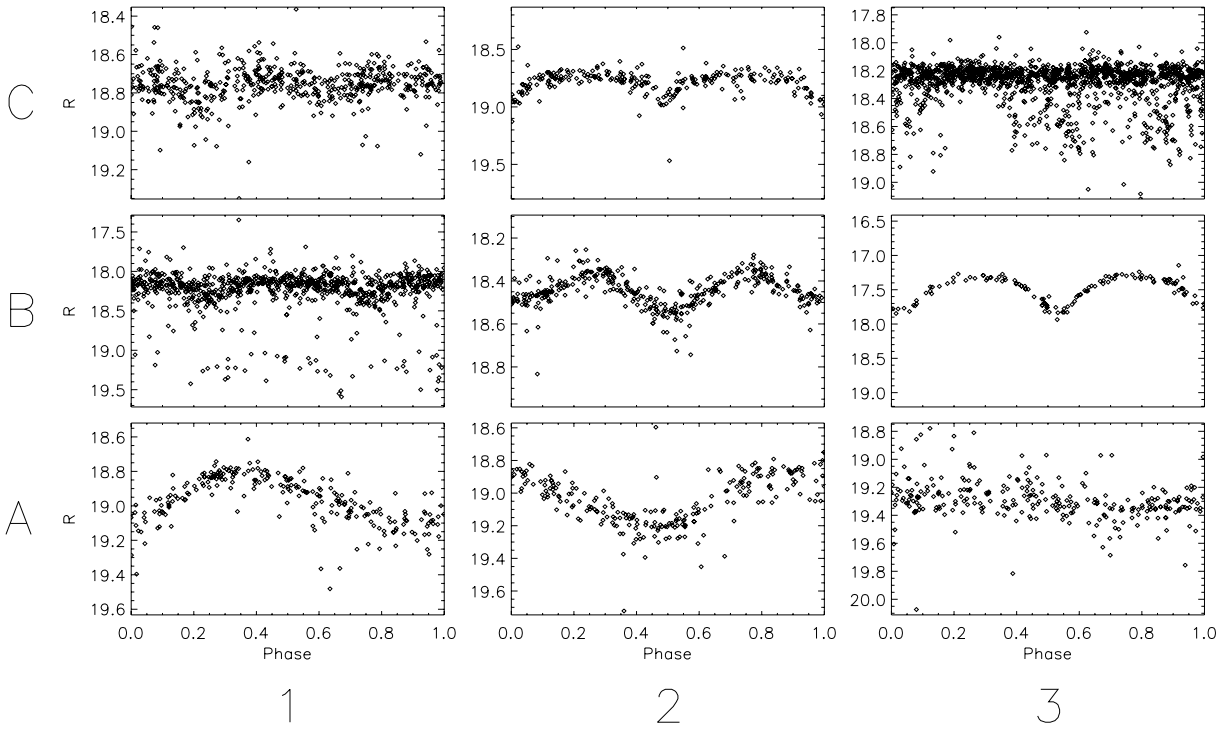

Figure 21. Light curves with the lowest measure of similarity from the MACHO EB catalogue. Only the MACHO *R* band is shown.

Table 6. MACHO EB outliers.

Survey	Type	ID	Plot COORD	Period (d)	Days of observation	Number of observations	Interpretation
MACHO	EB	64.7964.375	A1	0.322 79	2728.73	263	**RRAB Asymmetry, period.**
MACHO	EB	68.10485.363	A2	0.368 04	2714.83	205	**RRAB Asymmetry, period.**
MACHO	EB	27.10782.248	A3	0.287 17	2713.96	294	**RRAB Asymmetry, period.**
MACHO	EB	212.15797.121	B1	0.677 19	2711.93	910	Red band is noisy. Blue band is OK.
MACHO	EB	25.3836.269	B2	2.274 36	2711.82	341	
MACHO	EB	36.7395.92	B3	0.316 33	2702.76	276	
MACHO	EB	22.4871.431	C1	3.048 61	2702.73	530	
MACHO	EB	57.4953.114	C2	1.254 21	2660.68	278	
MACHO	EB	80.7194.423	C3	2.600 78	2649.06	1370	

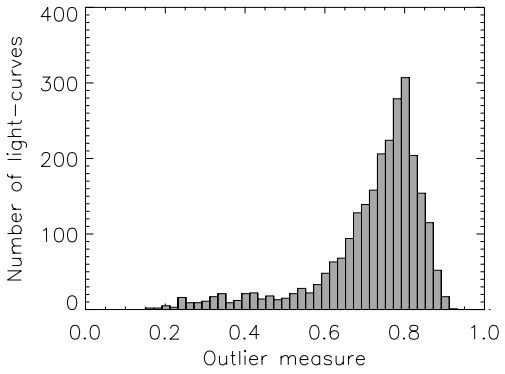

Figure 22. The histogram of the outlier measure for 2580 EBs in the OGLE sample.

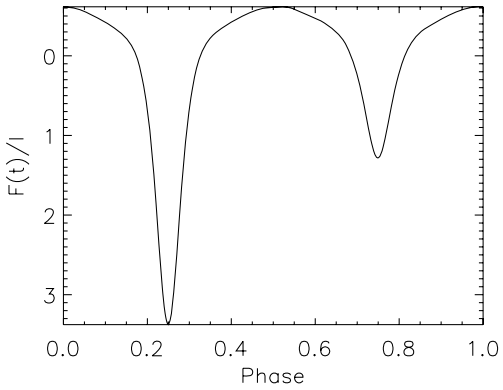

Figure 23. The centroid light curve for 2580 EBs in the OGLE sample.

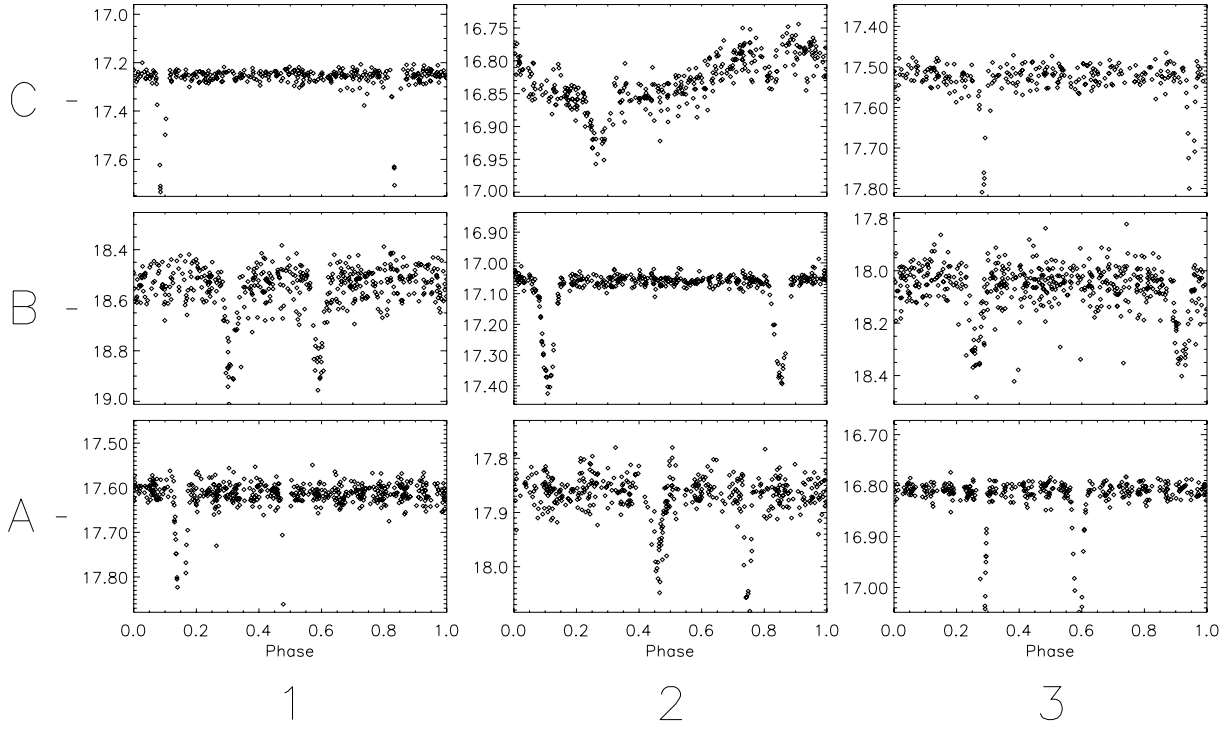

Figure 24. Light curves with the lowest measure of similarity from the OGLE-II EB catalogue. Shown here is only the OGLE *I* band is shown.

Table 7. OGLE EB outliers.

Survey	Type	ID(RA-Dec.)	Plot COORD	Period (d)	Days of observation	Number of observations	Interpretation
OGLE-II	EB	052937.78-700903.4	A1	15.033 14	1239	503	Eccentric orbit.
OGLE-II	EB	051915.79-693808.1	A2	8.033 76	1238	432	Eccentric orbit.
OGLE-II	EB	051519.31-692640.3	A3	15.962 56	1235	360	Eccentric orbit.
OGLE-II	EB	051858.34-693946.4	B1	2.295 55	1235	473	Eccentric orbit.
OGLE-II	EB	051700.39-691813.8	B2	5.291 29	1235	368	Eccentric orbit.
OGLE-II	EB	052521.32-694858.9	B3	4.120 88	1234	500	Eccentric orbit.
OGLE-II	EB	051734.54-692736.5	C1	14.582 52	1234	325	Eccentric orbit.
OGLE-II	EB	051657.87-690328.1	C2	5.661 41	1238	365	**EB with reflection effect.**
OGLE-II	EB	050646.85-683700.4	C3	12.149 88	1420	264	Eccentric orbit.

In this paper, we have used particular preprocessing tools and we tweaked our preprocessing steps for each catalogue. We are planning a full release of the software which will include many preprocessing options and optimized algorithms as a downloadable software and as an on-line tool and web services in the near future (http://darwin.cfa.harvard.edu/LightCurves/s/).

6 CONCLUSIONS

In this paper we presented a methodology based on cross-correlation as a measure of similarity that enables us to discover outliers in catalogues of periodic light curves. We established the methodology in Fourier space and extended the cross-correlation to accommodate observational errors.

The results from the application of our method on catalogues of classified periodic stars from the MACHO and OGLE projects are encouraging, and establish that our method correctly identifies light curves that do not belong to these catalogues as outliers.

We have identified light curves that were simply misclassified, light curves that were folded with the wrong period and so appear different, and light curves that emerged as unique.

We show how with careful approximations our method can be applied to very large catalogues, and thus making it a useful tool for the upcoming new surveys: the Panoramic Survey Telescope & Rapid Response System (Pan-STARRS) (http://pan-starrs.ifa.hawaii.edu) and Large Synoptic Survey Telescope (LSST) (http://www.lsst.org).

We have, none the less, also concluded that a single measure of similarity is not adequate to capture all features for all types of light curves and we understand that an extension of our method that utilizes more measures (comparison of Fourier components, wavelet coefficients etc) or combinations of measures have to be carried out; these will be presented in a future paper.

It is worth mentioning that other works performing automated classification of light curves (Brett, West & Wheatley 2004) can also, in principle, find outliers. However, since their focus is classification, there is no guarantee that an outlier will be identified. This is because a light curve must be clearly decoupled from all clusters in order to be considered as an outlier, whereas in our case, since we do not have clusters, any light curve can be classified as an outlier. This distinction is important in order to appreciate the advantage of our method. Moreover, a classification method cannot scale as N, whereas our method can do so in some approximation schemes.

We would like to make one last point. The situation of data sets that are not fully processed is going to become more common as the larger surveys come on-line. In the near future, it will become nearly impossible to fully 'clean' data sets without the use of automated methods such as the one presented here. We believe we have shown that our method has great utility at a number of steps along the processing pipeline.

ACKNOWLEDGMENTS

This work uses public domain data from the MACHO project whose work was performed under the joint auspices of the US Department of Energy, National Nuclear Security Administration by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48, the National Science Foundation through the Centre for Particle Astrophysics of the University of California under cooperative agreement AST-8809616, and the Mount Stromlo and Siding Spring Observatory, part of the Australian National University. This work also uses public domain data obtained by the OGLE project. We thank Kem Cook, Doug Welch and Gabe Prochter for compiling the lists of potential RR Lyræ and Cepheids and to Kem Cook for providing MACHO to standard magnitude transformations. We also thank Edward Guinan for his insight into some of the outlier light curves.

REFERENCES

Alcock C. et al., 1995, AJ, 109, 1654
Alcock C. et al., 1996a, ApJ, 470, 583
Alcock C. et al., 1996b, AJ, 111, 1146
Alcock C. et al., 1997a, ApJ, 482, 89
Alcock C. et al., 1997b, AJ, 114, 326
Alcock C. et al., 1999, PASP, 111, 1539
Alcock C. et al., 2000, ApJ, 542, 281
Allen C. W., 1973, Astrophysical Quantities, 3rd edn. Athlone Press, London
Belokurov V., Evans N. W., Du Y. L., 2003, MNRAS, 341, 1373
Belokurov V., Evans N. W., Le Du Y., 2004, MNRAS, 352, 233
Brett D. R., West R. G., Wheatley P. J., 2004, MNRAS, 353, 369
Cook K. H. et al., 1995, in Stobie R. S. W. P. A., ed., ASP Conf. Ser. Vol. 83, Astrophysical Applications of Stellar Pulsation Variable Stars in the MACHO Collaboration Database. Astron. Soc. Pac., San Francisco, p. 221
Engle S. G., Guinan E. F., Koch R. H., 2004, American Astronomical Society Meeting Abstracts, 06.09, 204
Evans N. R., Sasselov D. D., Short C. I., 2002, ApJ, 567, 1121
Eyer L., Blake C., 2005, MNRAS, 358, 30
Faccioli L., Alcock C., Cook K., Prochter G., Protopapas P., Syphers D., 2006, AJ, submitted
Ferlet R., Maillard J., Raban B., eds, 1997, Variables Stars and the Astrophysical Returns of the Microlensing Surveys. Editions Frontières, Paris
Friedman J. H., 1984, Technical report, A Variable Span Smoother, Technical Report No. 5. Laboratory for Computational Statistics. Department of Statistics, Stanford University

Gorry P. A., 1990, Am. Chem. Soc., 62, 570
Lomb N. R., 1976, Ap&SS, 39, 447
Luo J., Ying K., He P., Bai J., 2005, Digital Signal Processing, 15, 122
Mizerski T., Bejger M., 2002, Acta Astron., 52, 61
Paczyński B., 2001, in Banday A. J., Zaroubi S., Bartelmann M. eds, Proc. MPA/ESO Workshop, Mining the Sky Massive Variability Searches: the Past, Present and Future. Springer Verlag, Heidelberg, p. 481
Petit M., 1987, Variable stars. Chichester, England and New York, John Wiley and Sons, P. 268, Translation.
Pojmanski G., 2000, Acta Astron., 50, 177
Pojmanski G., 2002, Acta Astron., 52, 397
Pollacco D. L., Bell S. A., 1993, MNRAS, 262, 377
Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. T., 1992, Numerical recipes in C. The art of scientific computing, 2nd edn. Cambridge Univ. Press, Cambridge
Reimann J., 1994, PhD thesis, Department of statistics, University of California, Berkeley
Richter G., Wenzel W., Hoffmeister C., 1985, Variable stars. Springer-Verlag, New York, p. 339
Riemann J., 1994, PhD thesis, University of California, Berkeley
Soszyński I. et al., 2003, Acta Astron., 53, 93
Sterken C., Jaschek C., 1996, Light Curves of Variable Stars: A Pictorial Atlas. Cambridge Univ. Press, Cambridge
Udalski A., Kubiak M., Szymański M., 1997, Acta Astron., 47, 31
Udalski A., Soszyński I., Szymański M., Kubiak M., Pietrzyński G., Wożniak P., Żebruń K., 1999a, Acta Astron., 49, 437
Udalski A., Soszyński I., Szymański M., Kubiak M., Pietrzyński G., Wożniak P., Żebruń K., 1999b, Acta Astron., 49, 223
Udalski A., Soszyński I., Szymański M., Kubiak M., Pietrzyński G., Wożniak P., Żebruń K., 1999c, Acta Astron., 49, 437
Weisstein E. W., 1999, Correlation Coefficient. From MathWorld–A Wolfram Web Resource
Wolf M., Diethelm R., Hornoch K., 2001, A&A, 374, 243
Wolf M. et al., 2004, A&A, 420, 619
Wożniak P. et al., 2002, Szalay A. S., ed., Proc. SPIE Vol. 4846, SkyDOT (Sky Data base for Objects in the Time Domain): A Virtual Observatory for Variability Studies at LANL. SPIE, Bellingham, p. 147
Wyrzykowski L. et al., 2003, Acta Astron., 53, 1
Żebruń K. et al., 2001a, Acta Astron., 51, 317
Żebruń K., Soszyński I., Wożniak P. R., 2001b, Acta Astron., 51, 303

APPENDIX A: CONVOLUTION IN FOURIER SPACE

Let $x(n)$ and $y(n)$ be arbitrary functions of discrete time n with Fourier transforms. Take

$$x(n) = \boldsymbol{\mathcal{F}^{-1}}\left[\mathcal{X}(\nu)\right](n) = \frac{1}{N}\sum_{n=0}^{N-1}\mathcal{X}(\nu)\mathrm{e}^{2\pi \mathrm{i}\nu n/N} \tag{A1}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}\bar{\mathcal{X}}(\nu)\mathrm{e}^{-2\pi \mathrm{i}\nu n/N}, \tag{A2}$$

$$y(n) = \boldsymbol{\mathcal{F}^{-1}}\left[\mathcal{Y}(\nu)\right](n) = \frac{1}{N}\sum_{n=0}^{N-1}\mathcal{Y}(\nu)\mathrm{e}^{2\pi \mathrm{i}\nu n/N} \tag{A3}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}\bar{\mathcal{Y}}(\nu)\mathrm{e}^{-2\pi \mathrm{i}\nu n/N}, \tag{A4}$$

where $\bar{\mathcal{X}}$ and $\bar{\mathcal{Y}}$ are the complex conjugates Fourier transforms and $\boldsymbol{\mathcal{F}^{-1}}(n)$ is the inverse Fourier transform. The correlation, given a time-lag τ,

$$r^2_{xy}(\tau) = \sum_{n=0}^{N-1} x(n)\, y(n-\tau), \tag{A5}$$

is

$$\begin{aligned} r^2_{xy}(\tau) &= \sum_{n=0}^{N-1} x(n)\frac{1}{N}\sum_{\nu=0}^{N-1}\bar{\mathcal{Y}}(\nu)\mathrm{e}^{\mathrm{i}2\pi(\tau-n)\nu/N} \\ &= \frac{1}{N}\sum_{\nu=0}^{N-1}\bar{\mathcal{Y}}(\nu)\mathrm{e}^{\mathrm{i}2\pi\nu\tau/N}\sum_{n=0}^{N-1} x(n)\mathrm{e}^{-\mathrm{i}2\pi\nu n/N} \\ &= \frac{1}{N}\sum_{\nu=0}^{N-1}\bar{\mathcal{Y}}(\nu)\mathcal{X}(\nu)\mathrm{e}^{\mathrm{i}2\pi\tau\nu/N} \\ &= \mathcal{F}^{-1}[\bar{\mathcal{Y}}(\nu)\mathcal{X}(\nu)](\tau). \end{aligned} \tag{A6}$$

APPENDIX B: ERROR PROPAGATION

The SG smoothing can be written as a simple linear sum over neighbouring points

$$y_s = \sum_{i=-\frac{N-1}{2}}^{\frac{N-1}{2}} C_i \, y_i, \tag{B1}$$

where the coefficients C_i are difficult to deduce, but have no errors in them (they do not depend on the data). The error in the smoothed value is then given by

$$\sigma_{y_s} = \sqrt{\sum \left(\frac{\partial y_s}{\partial y_i} \, \sigma_{y_i} \right)^2}, \tag{B2}$$

implying,

$$\sigma_{y_s} = \sqrt{\sum (C_i \, \sigma_{y_i})^2}. \tag{B3}$$

To get the value of a measurement y for a given x, using linear interpolation between the two points (x_1, y_1) and (x_2, y_2), we have

$$y \; = \; \eta \, y_2 + (\eta - 1) \, y_1, \tag{B4}$$

where η is defined as

$$\eta = \frac{x - x_1}{x_2 - x_1}. \tag{B5}$$

Using the rules of error propagation,

$$\sqrt{\left(\frac{\partial y}{\partial y_1} \, \sigma_{y_1} \right)^2 + \left(\frac{\partial y}{\partial y_2} \, \sigma_{y_2} \right)^2}, \tag{B6}$$

calculating the derivatives we find,

$$\sigma_y = \sqrt{(1 - \eta)^2 \, \sigma_{y_1}^2 \; + \; \eta^2 \, \sigma_{y_2}^2}. \tag{B7}$$

Similarly, we can estimate the errors for the running averages where the running averages are

$$y = \sum_{i \in \text{window}} \mathrm{e}^{-(y - y_i)^2 / 2\omega^2} \, y_i, \tag{B8}$$

where ω is the window size. Estimating the derivatives, we get

$$\sigma_y^2 = \sum_{i \in \text{window}} \mathrm{e}^{-(y - y_i)^2 / \omega^2} \; \left[1 - \frac{y_i \, (y - y_i)}{\omega} \right]^2 \sigma_{y_i}^2. \tag{B9}$$

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