The Astrophysical Journal Supplement Series, 216:25 (30pp), 2015 February doi:10.1088/0067-0049/216/2/25


# A NOVEL, FULLY AUTOMATED PIPELINE FOR PERIOD ESTIMATION IN THE EROS 2 DATA SET

Pavlos Protopapas[1,2], Pablo Huijse[3,4], Pablo A. Estévez[3,4], Pablo Zegers[5], José C. Príncipe[6], and Jean-Baptiste Marquette[7]
[1] Institute for Applied Computational Science, Harvard University, Cambridge, MA 02138, USA; pavlos@seas.harvard.edu
[2] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[3] Millennium Institute of Astrophysics, Chile
[4] Department of Electrical Engineering and the Advanced Mining Technology Center, Universidad de Chile, Santiago, Chile
[5] Universidad de los Andes, Facultad de Ingeniería y Ciencias Aplicadas, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile
[6] Computational Neuroengineering Laboratory, University of Florida, Gainesville, FL 32611, USA
[7] UPMC-CNRS, UMR7095, Institut d'Astrophysique de Paris, F-75014 Paris, France
*Received 2013 September 2; accepted 2014 December 3; published 2015 January 27*

## ABSTRACT

We present a new method to discriminate periodic from nonperiodic irregularly sampled light curves. We introduce a periodic kernel and maximize a similarity measure derived from information theory to estimate the periods and a discriminator factor. We tested the method on a data set containing 100,000 synthetic periodic and nonperiodic light curves with various periods, amplitudes, and shapes generated using a multivariate generative model. We correctly identified periodic and nonperiodic light curves with a completeness of ∼90% and a precision of ∼95%, for light curves with a signal-to-noise ratio (S/N) larger than 0.5. We characterize the efficiency and reliability of the model using these synthetic light curves and apply the method on the EROS-2 data set. A crucial consideration is the speed at which the method can be executed. Using a hierarchical search and some simplification on the parameter search, we were able to analyze 32.8 million light curves in ∼18 hr on a cluster of GPGPUs. Using the sensitivity analysis on the synthetic data set, we infer that 0.42% of the sources in the LMC and 0.61% of the sources in the SMC show periodic behavior. The training set, catalogs, and source code are all available at http://timemachine.iic.harvard.edu.

*Key words:* methods: data analysis – stars: variables: general

## 1. INTRODUCTION

Characterization of the dynamic optical sky is one of the observational frontiers in astrophysics. Variable sources, defined as any source whose apparent magnitude changes over time, have historically led to fundamental insights into subjects ranging from the structure of stars and the most energetic explosions in the universe to cosmology. These changes and their characteristics tell us a lot about sources such as pulsating stars and supernovae, as well as about the interaction of the source with its surroundings, such as active galactic nuclei (AGNs) or light being blocked by something between the source and the observer. However, no optical telescope to date has had the capability to search for transient phenomena at faint levels over enough of the sky to fully characterize variable sources.

A subcategory of the variable sources are the periodic variables. These are variables that in general repeat at regular intervals. While astronomers historically have been able to study variable and transient phenomena by examining the behavior of individual sources, the amount of data and the large number of sources have exponentially grown in the past decade (Hodapp et al. 2004; Ivezic et al. 2011; Larson et al. 2003; Law et al. 2009), making this task daunting.

Although most stars have at least some variation in luminosity, current estimations indicate that 3% of the stars are varying more than the sensitivity of the instruments and ∼1% are periodic (Eyer 1999). EROS-2 (Tisserand et al. 2007), MACHO (Alcock et al. 2000), and OGLE (Udalski et al. 1997) were among the first generation of large-scale surveys, monitoring millions of sources for many years. Pan-STARRS (Hodapp et al. 2004) is currently monitoring the whole visible sky repeatedly, and it will be doing this for a total of three years. In the future Sloan Digitial Sky Survey (SDSS; York et al. 2000), LSST (Ivezic et al. 2011) will monitor even more sources, and more frequently, generating billions of light curves. It is because of this explosion of data that there is a need for efficient and well-characterized period-finding techniques.

The problem of period estimation from noisy and irregularly sampled observations has been studied before. Most approaches identify the period by some form of grid search. That is, the problem is solved by evaluating a criterion $\Phi$ at a set of trial periods and selecting the period $p$ that yields the best value for $\Phi(p)$. Commonly used techniques vary in the form and parameterization of $\Phi$, the evaluation of the fit quality between model and data, the set of trial periods searched, and the complexity of the resulting procedures. Two methods that are popular are the Lomb–Scargle (LS) periodogram (Scargle 1982; Reimann 1994) and the phase dispersion minimization (PDM; Stellingwerf 1978), both known for their success in empirical studies. The LS method is relatively fast and is equivalent to maximum likelihood estimation under the assumption that the function has a sinusoidal shape. It therefore makes a strong assumption on the shape of the underlying function. On the other hand, PDM makes no such assumptions and is more generally applicable, but it is slower and is less often used in practice.

In this paper we adopted the correntropy kernelized periodogram (CKP), an information theoretical criterion introduced in Huijse et al. (2012) to assess periodicity in light curves. The CKP combines the generalized autocorrelation function (Principe 2010) with a periodic kernel yielding a generalized periodogram. The CKP measures similarity over time using statistical information contained in the probability density function (pdf) of the samples. This gives the CKP an advantage over methods that rely on second-order statistical descriptors.[8]

[8] To fully characterize non-Gaussian random processes, the higher-order moments are needed.

By adjusting the kernel parameters of the CKP, one can adapt the metric to different noise regimes and periodicities. The selection of these parameters for the case of light curves is thoroughly discussed in the present work.

To fully qualify the method, we generated a large set of synthetic light curves (110,000) using parameter distributions motivated from the data. To do so, we used a model-free multivariate generative model and sampled the parameters. We also use a smaller but manageable subset from the real data in order to compare our results with reality. These subsets were used to optimize the free parameters of the pipeline and to characterize the efficiency and completeness of the process.

Astronomy and many experimental sciences are now collecting more data than can be possibly analyzed by human experts in a reasonable amount of time. We are not really interested in the data per se, but in the information they contain about the natural phenomena. Machine learning and signal processing are becoming an integral part of the process of extracting information from data, because they are quantitative methods based on statistics and function analysis methods. This synergism is in its early stages, and this paper shows an effective methodology to speed up the discovery of periodic stars in large databases such as the EROS-2.

Section 2 describes the theoretical framework that this work is based on, Section 3 describes the pipeline and methodology, Section 4 describes the synthetic data set, Section 5 describes the data, Section 6 contains the results obtained from our runs, and conclusions are given in Section 9.

## 2. THEORETICAL FRAMEWORK

The structure of a time series can be quantified by measuring the signal similarity over time. The first measure that comes to mind is the autocorrelation function of the time series (Jenkins & Watts 1968). Let us define the time series as a realization of a stochastic process $\{x_n, n = 0, 1, \ldots, N\}$, where $x$ is a random variable in $\mathbb{R}$. The autocorrelation function for stationary processes is defined as

$$R[m] = \mathbb{E}[\langle x_n, x_{n-m}\rangle], \tag{1}$$

where $\mathbb{E}[\cdot]$ indicates the expectation value. The autocorrelation coefficient (covariance normalized by the variance) normally is estimated for stationary and ergodic time series as a simple sum of lagged products over a window of data:

$$\hat{R}[m] = \frac{1}{N+1-m}\,\frac{1}{\sigma^2}\sum_{n=m}^{N}(x_n-\mu)(x_{n-m}-\mu), \tag{2}$$

where $N+1$ is the number of measurements in the time series and the true mean $\mu$ and true variance $\sigma^2$ are time independent.

Looking more closely at the autocorrelation definition, one finds out that only second-order information of the random variable $x$ is utilized in the definition, and as is well known, only a few distributions such as the Gaussian are fully described by their (first- and) second-order moments. Therefore, one compromises the simplicity of the autocorrelation definition with a loss of a more in-depth description of the signal similarity. This paper will use more powerful definitions of similarity for a better quantification of time series structure, which is pivotal to achieve the reported results. The ideas are founded in the mathematical theory of information and a descriptor of entropy that exploits the full statistical information from samples (Principe 2010), which is utilized to define similarity metrics.

Let us consider a stationary stochastic process $\{x_n\}$ and define the generalized autocorrelation as

$$V[m] = \mathbb{E}[\kappa(x_n, x_{n-m})], \tag{3}$$

where $\kappa(x, y)$ is a positive definite function of two arguments called a kernel (Schölkopf & Smola 2002; Taylor & Cristianini 2004). If we define $\kappa(x, y) = \langle x, y\rangle$, i.e., the first-order polynomial kernel, one obtains the autocorrelation function of Equations (1) and (2). Instead, let us select $\kappa(x, y)$ as a translation-invariant kernel (Schölkopf & Smola 2002), i.e., $\kappa(x, y) = \kappa(x - y, 0)$. For simplicity we will use $\kappa(x - y)$ for translation-invariant kernel functions. The Gaussian kernel, defined as

$$G_\sigma(x-z) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right), \tag{4}$$

is a popular kernel that fits the conditions, where $\sigma$ is the covariance, and will be called in this context *the kernel size*. In Principe (2010) this class of functions is called autocorrentropy, or more simply correntropy, and here we will always assume the use of the Gaussian kernel. One of the advantages of correntropy is that it is still very easy to estimate directly from data assuming that the random process is ergodic. Using the sample mean, we can estimate Equation (3) as

$$\widehat{V}_\sigma[m] = \frac{1}{N+1-m}\sum_{n=m}^{N} G_\sigma(x_n - x_{n-m}). \tag{5}$$

The difference between autocorrelation and autocorrentropy seems pretty minor, but it is very significant, as fully discussed in Principe (2010). For this work, the important correntropy properties are the following.

1. Correntropy with the Gaussian kernel includes a weighted sum of all the even moments of the random variable, including the second-order moment (the autocorrelation) of $\|x_n - x_{n-m}\|$.
2. Correntropy is a positive definite function, and therefore it can replace the autocorrelation function in the definition of the Power spectrum, yielding the correntropy spectral density (CSD) (Principe 2010), as

$$P_\sigma[f] = \sum_{m=-\infty}^{\infty} U_\sigma[m]\cdot\exp\left(-i\,2\pi f\frac{m}{F_s}\right), \tag{6}$$

   where $F_s$ corresponds to the sampling frequency. The function $U_\sigma[m]$ corresponds to $V_\sigma[m] - IP$, where $IP$ corresponds to the mean value of the autocorrentropy function over the lags.[9]
3. Correntropy has a free parameter that can be interpreted as a scale parameter and therefore needs to be defined according to the time series data.
4. Correntropy quantifies similarity using the correntropy induced metric (CIM), defined as

$$\mathrm{CIM}(x, y) = (\kappa(0,0) - \mathbb{E}[\kappa(x, y)])^{1/2}. \tag{7}$$

   The CIM is a metric very different from the $L_p$ norms that define the Minskowski spaces, where the distances are

[9] This is also the argument of Renyi's quadratic entropy (Principe 2010).

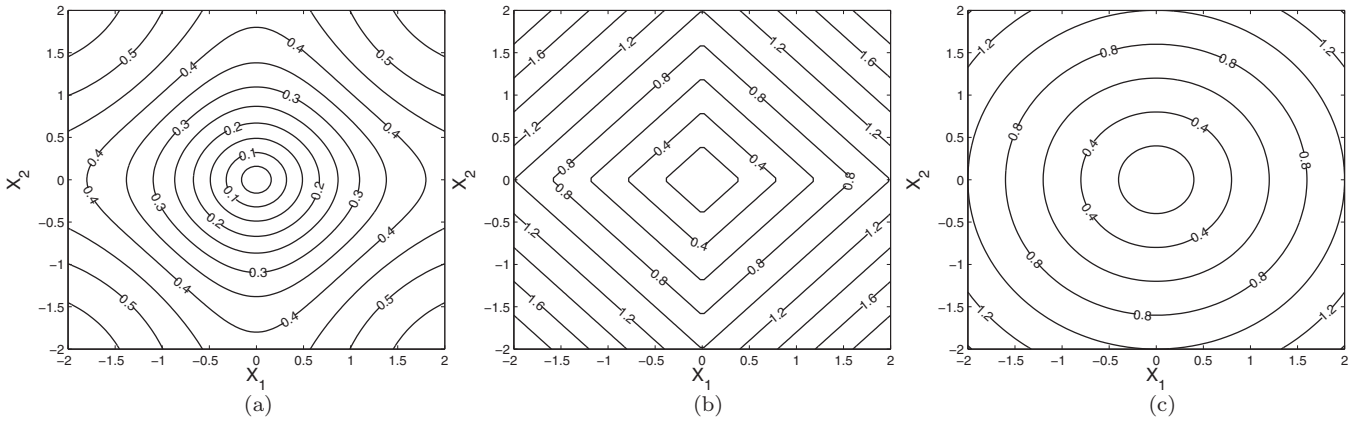


**Figure 1.** Distances to the origin (contours) in a bidimensional sample space using (a) the CIM(X,0), (b) $L_2$ norm, and (c) $L_1$ norm. For the CIM (Equation (7)) a Gaussian kernel function with $\sigma = 1$ is considered. Note how the CIM incorporates the $L_1$, $L_2$, and $L_0$ norms at different scales.

always weighted the same (Figure 1).[10] This means that distances between the arguments of the CIM are weighted nonuniformly, i.e., if the distance between the arguments is small, then the CIM approximates the $L_2$ norm, but if the difference is larger, then it will approximate the $L_1$ norm, and for very large difference between the arguments, the CIM tends to the $L_0$ norm. The transitions between the norms are smooth, and the assessment of small and large, the scale in this space, is controlled by the kernel size, which impacts drastically the assessment of similarity.

It is appropriate to present a synthetic example to illustrate the difference between autocorrelation and autocorrentropy in assessing similarity over time, as well as to elucidate the role of the kernel size. Let us take the case of the stochastic process with uniform random amplitude in $[-A, A]$ and a random phase in $[-\pi, \pi]$ defined as $x_n = A\sin(w_0 n + \varphi)$. As is well known, the autocorrelation function of sine waves is a sine wave with the same period. However, should it be a sine wave if we are interested in assessing the degree of similarity of the signal time structure? Since the sine wave is periodic, the similarity is maximum when the delay is exactly one period, but for intermediate shifts, the two functions are very dissimilar, and autocorrelation does not show this very clearly (and the similarity is neither normalized nor always positive, hence the use of the correlation coefficient). Therefore, if we are seeking a discriminative measure of similarity, the autocorrelation function is not exploiting optimally the information available in the statistics of the data. It turns out that correntropy is more discriminative, as shown in Figure 2. The autocorrentropy of a sine wave (or any other periodic function) is a periodic pulse train defined by the data period, where the pulses can be made arbitrarily sharp by decreasing the kernel size to zero. This can be easily explained by observing Equation (5). When $x_n$ and $x_{n-m}$ are similar, the argument is close to zero and the Gaussian yields a value close to the argument square; when the difference increases, the Gaussian function produces exponentially smaller results proportional to the difference in arguments ; and for larger differences, the Gaussian gives back very small values close to zero (see Figure 1(a)). Of course, if white noise is added to the sine wave, one immediately sees that the kernel size cannot be made arbitrarily small; otherwise, the correntropy becomes always very small, not capturing the periodic nature of the noisy signal. However, if the kernel size needs to be made very large to accommodate large noises, then the autocorrentropy approaches the autocorrelation function.

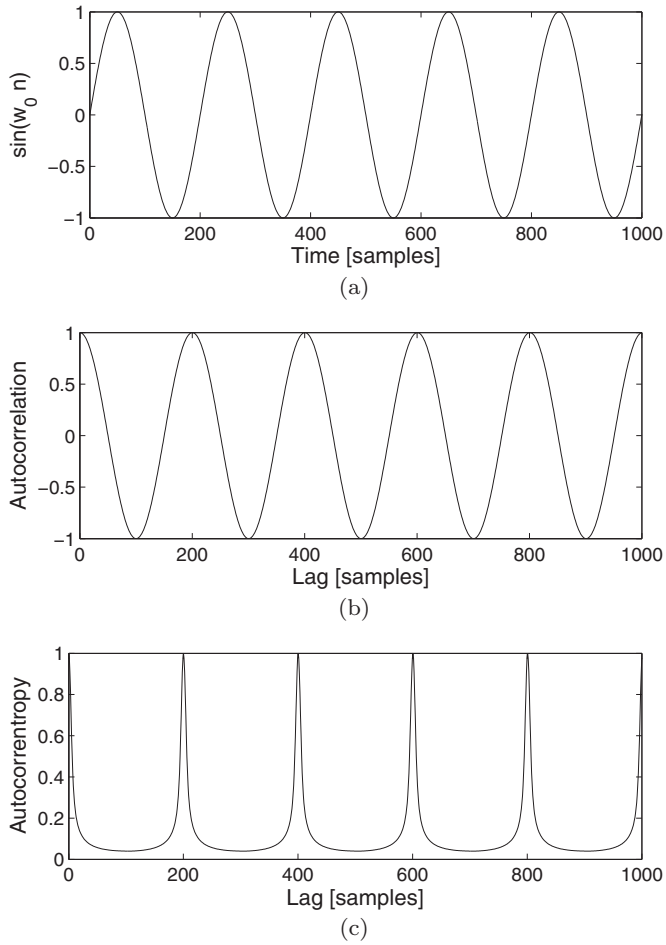


**Figure 2.** (a) Plot of $x_n = A\sin(w_0 n + \varphi)$ with unit amplitude, $w_0 = 2\pi/200$, and where $\varphi$ is a random variable uniformly distributed in $[-\pi, \pi]$. (b) Autocorrelation of $x_n$; note that the autocorrelation function of a sine wave is a sine wave. (c) Autocorrentropy of $x_n$; note that the autocorrentropy of a sine wave is a train pulse in which the periodicity is represented by the peaks. The sharpness of the peaks can be controlled using $\sigma$.

### 2.1. Periodic Kernel

With this introduction in mind, we move on to specifying the kernel that best encapsulates the information in the data

[10] For $\boldsymbol{x} \in \mathbb{R}^n$, the $L_p$ norms are defined as $L_p = \|\boldsymbol{x}\|_p = (\sum_{i=1}^{N} x_i^p)^{\frac{1}{p}}$, $p \in (0, \infty)$. In the limit $p \to 0$, the $L_0$ norm is defined as the number of nonzero components in the vector (counting norm).

for periodic signals. Periodic kernel functions are known to be appropriate for nonparametric estimation, modeling, and regression of periodic time series (Michalak 2010). A kernel function is periodic with period $P$ if it repeats itself for inputs separated by $P$. Periodic kernel functions have also been proposed in the Gaussian process literature (Rasmussen & Williams 2006; Mackay 1998; Wang et al. 2012).

A periodic kernel function can be obtained by applying a nonlinear mapping (or warping) $u(t)$ to the input vector $t$. In Mackay (1998) a periodic kernel function was constructed by mapping a unidimensional input variable $t$ using a periodic two-dimensional warping function defined as

$$u_f(t) = (\cos(2\pi f t), \sin(2\pi f t)).$$

The periodic kernel function $G_\sigma^P(f, t_z - t_y)$ with period $1/f$ is obtained by applying this warping function to the inputs of the Gaussian kernel function (Equation (4)). The periodic kernel function is defined as

$$\begin{aligned} G_\sigma^P(f, t_z - t_y) &= G_\sigma(u_f(t_z) - u_f(t_y)) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left( - \frac{2\sin^2(\pi f(t_z - t_y))}{\sigma^2} \right), \end{aligned} \tag{8}$$

where the following expression is used:

$$\|u_f(z) - u_f(y)\|^2 = 4\sin^2(\pi f(z - y)).$$

Note that the periodic kernel is a function of $\delta t = (t_z - t_y)$ and frequency, the inverse of the period. The Taylor series expansion at $\delta t = 0$ of Equation (8) is defined as

$$\begin{aligned} G_\sigma^P(f, \delta t) = &\lim_{N\to\infty} \sum_{k=0}^{N} \frac{(-1)^k}{k!\sigma_t^{2k} 2^{k-1}} \\ &\times \left[ \sum_{m=0}^{k} \binom{2k}{k-m} (-1)^m g_m \cos(2\pi m f \delta t) \right], \end{aligned} \tag{9}$$

where

$$g_m = \begin{cases} 1/2, & \text{if } m = 0. \\ 1, & \text{otherwise.} \end{cases}$$

Note that for large values of $\sigma$, only the first terms contribute to the sum, and thus the periodic kernel tends to a constant plus $\cos(2\pi f \delta t)$, which corresponds to the real part of the Fourier basis.

## 3. METHOD

We base our methodology on the work described in Huijse et al. (2012). In this section we summarize the key points from that work, which are based on the overall description given in the previous sections, and then introduce the new concepts, particularly an intuitive interpretation of the parameters of the CKP, simple rules to select these parameters, and a normalization term that is needed to perform ensemble comparisons.

The CKP used in Huijse et al. (2012) is a period detection function developed for unevenly sampled time series. The CKP is computed from the available samples following a direct quadratic estimator approach, as proposed in Marquardt & Acuff (1984).[11] For a discrete unidimensional random process $\{x_n, n = 1, \ldots, N\}$ with kernel sizes $\sigma_t$ and $\sigma_y$ and a period $1/f$, the CKP is computed as

$$\begin{aligned} \mathrm{CKP}_{\{\sigma_t,\sigma_y\}}(f) = &\frac{1}{N^2} \\ &\times \sum_{i=1}^{N}\sum_{j=1}^{N} \left( G_{\sigma_y}(\Delta y_{ij}) - IP_{\sigma_y} \right) G_{\sigma_t}^P(f, \Delta t_{ij}), \end{aligned} \tag{10}$$

where $\Delta y_{ij} = y_i - y_j$, $\Delta t_{ij} = t_i - t_j$, $G_{\sigma_y}(\cdot)$ is the Gaussian kernel function (Equation (4)), $G_{\sigma_t}^P(\cdot,\cdot)$ is the periodic kernel function (Equation (8)), and $IP_{\sigma_y}$ is the information potential

$$IP_{\sigma_y} = \frac{1}{N^2} \sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma_y}(\Delta y_{ij}). \tag{11}$$

Note that Equation (10) is similar to the CSD (Equation (6)) with two main differences: (1) the CKP is estimated in a direct approach; and (2) the basis functions, $\exp(-i\,2\pi f m/F_s)$, have been replaced by the periodic kernel (Equation (8)). In this sense the CKP can be interpreted as the result of transforming the autocorrentropy function through a basis defined by the periodic kernel.

By comparing magnitude values through the autocorrentropy function, the CKP is effectively using a CIM (Equation (7)) metric to measure magnitude distances. The kernel size $\sigma_y$ has influence in the assessment of magnitude similarities, as explained in the previous section. The CKP compares time differences with the trial period through the periodic kernel. The periodic kernel size $\sigma_t$ allows the user to choose how this comparison is made.

By summing in the time and magnitude index, a function of the trial period is obtained; thus, the CKP can be considered a generalized periodogram. Consequently, in order to detect periods in light curves the CKP is maximized over the frequency (inverse of the period) for a given combination of parameters, namely, the two kernel bandwidths ($\sigma_y$, $\sigma_t$).

One of the major advantages of the CKP over conventional methods is its adaptability given by the kernel parameters. In what follows, we describe heuristic approaches that use the available information on the light curve to set the kernel sizes. Without them, the maximization of the CKP would have been a very expensive procedure.

The kernel bandwidth, $\sigma_y$, controls the observation window that is used to compare the magnitude values of the light curve. This parameter needs to be set small enough so that outliers are filtered, but large enough to compensate for the observational and other measurement errors. Conveniently, those errors are usually available for most measurements in light curves (these are the magnitude errors). For a given light curve the Gaussian kernel bandwidth is selected as

$$\sigma_y = \mathrm{med}(\{e\}), \tag{12}$$

where med is the median and $\{e\}$ are the error bars of the light-curve measurements. Figure 3(a) shows a synthetic periodic light curve with random error bars. Samples $y_1$ and $y_2$ are compared using the Gaussian kernel, where the median of the error bars is 0.08 and the $\sigma_y$ is set to be 0.08. Figure 3(b) shows the equivalent Gaussian kernel value for this pair. In reality, the observational errors are not constant, and therefore Equation (12) should not be the same for all pairs and should be a combination of the two observational errors added in quadrature.

[11] The basic idea is that for uneven samples, one can calculate the periodogram without having to regularize the data.

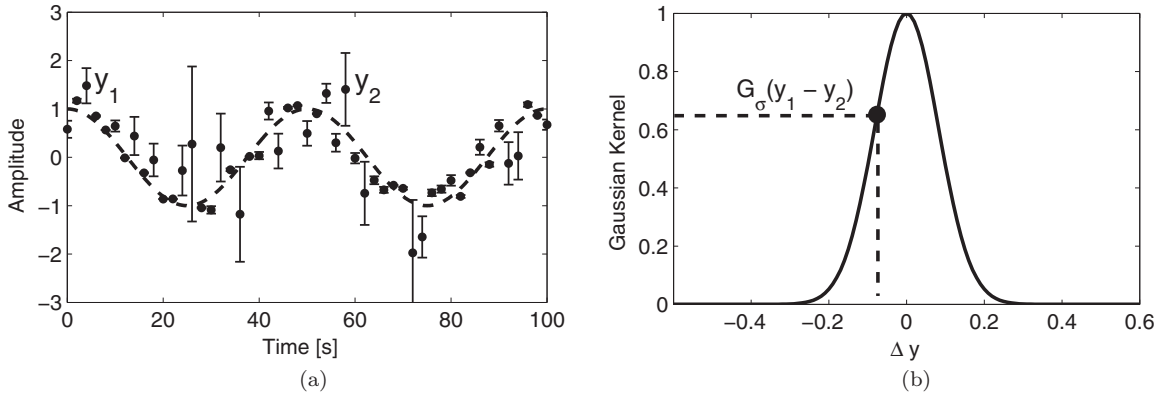


**Figure 3.** (a) Periodic synthetic time series $\cos(2\pi f t) + N(0, 0.5)$. The dotted line corresponds to the underlying signal. In this example the median of the error bars is 0.08. Samples $y_1$ and $y_2$ are compared using the Gaussian kernel (b). The kernel size is set to 0.08.

Practically the difference of this approximation and the correct approach is insignificant.

The kernel bandwidth, $\sigma_t$, controls the observation window that is used to compare the time differences of the light curve with the trial period. When $\sigma_t \to 0$, only the samples whose time differences are equal to the trial period will be picked by the periodic kernel. The smaller the $\sigma_t$ is, the more precise the estimation will be, although in practice fewer samples will be available. When $\sigma_t$ grows large, the exponential in Equation (8) has less relevance and the periodic kernel tends to a sinusoidal function (as shown in Section 2.1 through the Taylor expansion of Equation (8)). Intuitively, this parameter has influence on the periodicity's shape. A smaller $\sigma_t$ is beneficial to pick up shapes that have many features or abrupt changes, such as the narrow eclipses of an Algol-type eclipsing binary. On the contrary, a large $\sigma_t$ is used for smoother shapes, i.e., wiggles and high derivatives are ignored. In summary, the $\sigma_t$ needs to be set small enough so that the features of the periodicity will not be missed, but large enough so that there will be enough samples representing the period and to avoid picking up structures due to the noise.

Since $\sigma_t$ describes the smoothness of the shape of the light curve, a way to estimate $\sigma_t$ is to find the variation of $\delta t$ values in a given $y$ band. Empirically, we observed that for almost all periodic light curves, the CKP is maximized for $\sigma_t \sim 0.1$–$0.6$ and that the value of $\sigma_t$ is strongly correlated with the third moment or the skewness of the distribution of the magnitudes of the light curves. Light curves with skewed distributions, such as those corresponding to eclipsing binaries (Figure 4(a)), get a small $\sigma_t$ value. On the other hand, light curves with very symmetric distributions (Figure 4(b)) will get a larger $\sigma_t$. Finally, we will address ensemble comparisons for period discrimination. The kernel sizes are selected for each light curve differently as described above, and in order to compare different light curves, the CKP is required to be invariant under $\sigma_y$, $\sigma_t$, and the sample size.

For that we propose a properly normalized CKP metric as

$$\mathrm{nCKP}_{\{\sigma_t,\sigma_y\}}(f) = \frac{\sqrt{N\sigma_t}}{I P_{\sigma_y}} \frac{1}{N^2} \times \sum_{i=1}^{N}\sum_{j=1}^{N} \left(G_{\sigma_y}(\Delta y_{ij}) - I P_{\sigma_y}\right) G^P_{\sigma_t}(f, \Delta t_{ij}), \quad (13)$$

where $1/I P_{\sigma_y}$ normalizes against $\sigma_y$, $\sqrt{\sigma_t}$ normalizes against $\sigma_t$, and $\sqrt{N}$ normalizes against the number of samples. The normalization factors were confirmed empirically by comparing the distribution of the CKP across different sets of surrogate light curves, generated with the procedures described in Section 4. Figure 5(a) shows a histogram of $\max \mathrm{CKP}_{\{\sigma_t,\sigma_y\}}(f)$ for three sets of surrogates generated with different $N$ values. In this figure the unnormalized CKP is used (Equation (10)). For the histogram shown in Figure 5(b) the normalized CKP (Equation (13)) is used. In this case the distribution of the CKP is equivalent; thus, it is invariant to the different $N$ of the surrogates.

### 3.1. Trial Period Extraction, the Bands Method

The parameter to be estimated by maximizing the CKP is the period. Unfortunately, the dependence of CKP on period is not uniform and is difficult to model (Huijse et al. 2012); therefore, any clever optimization technique fails to converge faster than the brute-force approach.

To alleviate this problem, a fast search algorithm is adopted. The basic idea is that two points in an ideal light curve having the same magnitude have to be apart in time by an integer multiple of the period. For the ideal light curve case, finding the period is as simple as finding the greatest common divisor of the times of two points with the same magnitude.[12] However, the ideal case is not applicable to astronomical data because (a) light curves are composed of a nominal part and a signal part, as in the case of planetary transits and eclipsing binaries; (b) the observations are not performed continuously; and (c) measurements are not perfect but suffer from observational errors.

What follows is an approximation tailored for real light curves. Instead of looking at pairs of points with the same magnitude, subsets of points with similar magnitudes are selected. These subsets, called bands, should contain points that have time differences that are multiples of the period; therefore, in Fourier space these periods are enhanced. To avoid bands where the light curve is in its nominal state, we select bands where the derivatives are higher.

The details of the method are as follows.

For a unidimensional time series $\{t_i, x_i\}$ with $i = 1, \ldots, N$,

1. compute the first derivatives $d_i = (x_{i+1} - x_i)/(t_{i+1} - t_i)$;
2. divide the ordinate axis into 10 uneven-width bands, such that each band has 10% of the light curve samples;
3. compute the sum of the first derivatives that belong to band $j$ ($B_j$), $D_j = \sum_{i \in B_j} |d_i|$, with $j = 1, \ldots, 10$;

[12] This is the famous Euclid algorithm (oldest known).

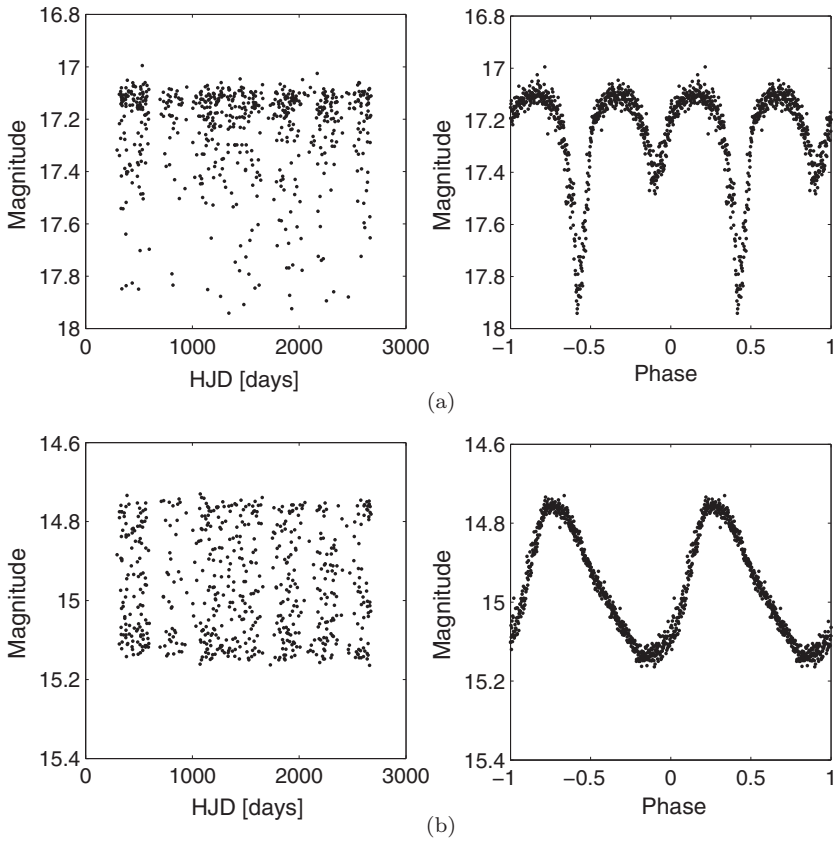


**Figure 4.** (a) Light curve lm0090l7821 folded with a period of 1.4255 days. This light curve has a highly positive skewed distribution. A time kernel bandwidth of $\sigma_t = 0.115$ is selected for this light curve. (b) Light curve lm0090n9337 folded with a period of 4.3949 days. This light curve has a symmetric distribution. In this case a time kernel bandwidth of 0.475 is selected.

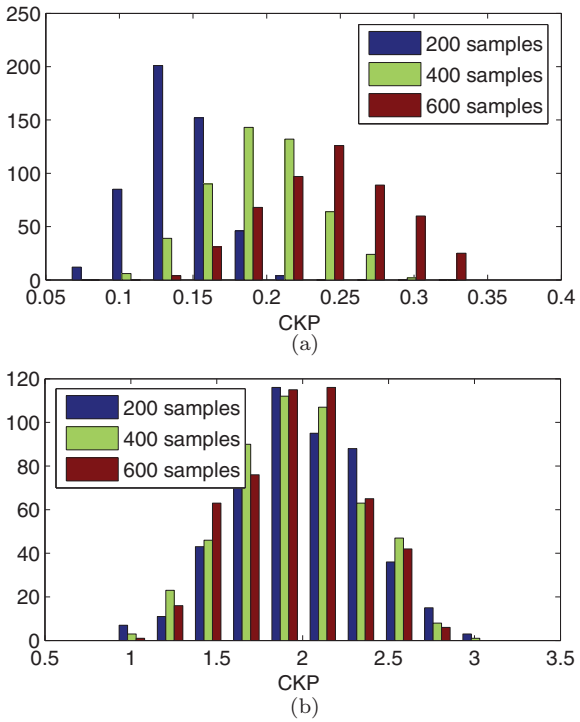


**Figure 5.** Distribution of the maximum CKP values on a set of 1500 synthetic light curves. The light curves are generated with the same period and S/N but using a different number of samples ($N$). Three sets of 500 light curves are generated using $N = 200$, 400, and 600. (a) Distribution of the unnormalized CKP. It is clear the CKP is not invariant to $N$. Light curves with higher $N$ have higher CKP values. (b) Distribution of the normalized CKP.

4. sort the bands in descending order of $D_j$ and keep the first $N_b$ bands;
5. for each band compute the spectral window function (Jenkins & Watts 1968) on a linearly spaced frequency grid from 0.00125 1/days to 3 1/days (periods between 0.3 days and 800 days),

$$S_j(f) = \left| \sum_{i \in B_j} \exp\left(\jmath 2\pi f t_i\right) \right|^2 ; \qquad (14)$$

6. save the frequencies associated with the $N_t$ highest local maxima of $S_j(f)$. Periods that comply with $\|P - 1\| < 1e - 4$ are omitted.[13] This gives a total of $N_b\, N_t$ trial frequencies.

The number of analyzed bands, $N_b$, and the amount of trial periods extracted per band, $N_t$, are user-defined parameters that represent a trade-off between efficiency and computational time. We expect to find the correct period in the first sorted bands; however, the true period may be captured by different bands although with different amplitudes, i.e., the rank of the true period may vary across bands. For example, the true period may be ranked 100th in the first band and 10th in the third band. Synthetic light curves (see Section 4) are analyzed with the period detection pipeline using different combinations of $N_t$ and $N_b$.

[13] The one day pseudo-sampling period is strongly represented in all the bands.

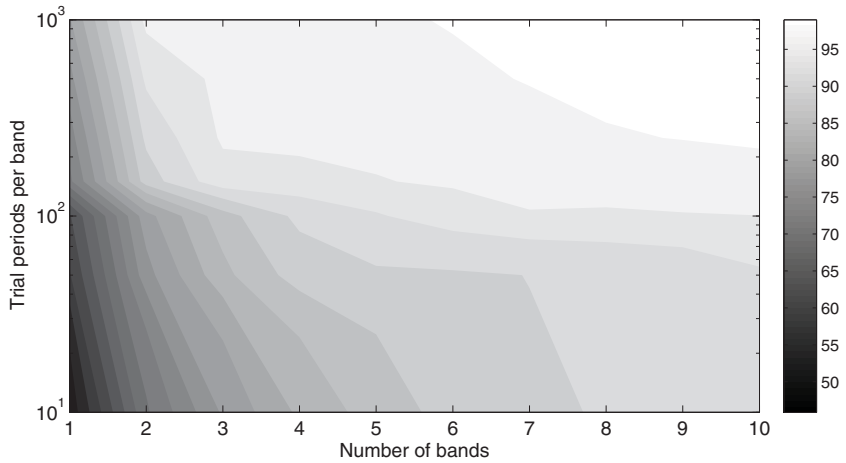


**Figure 6.** Hit rate as a function of the parameters of the bands methods. These parameters are the number of bands $N_b$ and the number of trial periods extracted per band $N_t$.

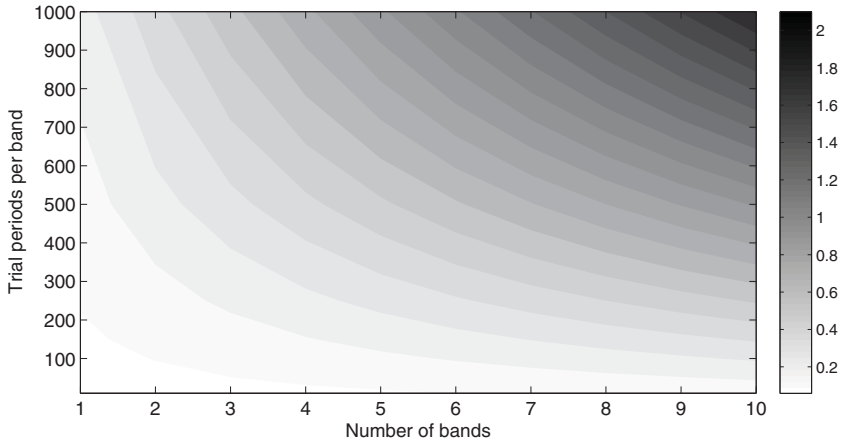


**Figure 7.** Computational time in seconds required to process one light curve (600 samples) as a function of the parameters of the bands methods. These parameters are the number of bands $N_b$ and the number of trial periods extracted per band $N_t$.

Figure 6 shows a contour plot of the hit rate as a function of $N_b$ and $N_t$. As expected, hit rates increase with $N_b$ and $N_t$. For every $N_t$, the hit rate gain obtained by adding additional bands decreases with $N_b$, which indicates that the bands are correctly sorted. Figure 7 shows a contour plot of the computational time required to analyze one light curve as a function of $N_b$ and $N_t$. For two points with equal $N_b N_t$ the point with lower $N_b$ requires less computational time. In terms of computational time, adding bands is less desirable than increasing $N_t$. The maximum hit rate achieved is 98.1%. We find the best operation point to be $N_b = 3$ and $N_t = 150$, which yields a hit rate of 95.1% with a computational time of 0.162 s per light curve. This point represents the best compromise between efficiency and computational time and is found by maximizing $\mathrm{HR} + 1/c_t$, where $c_t$ is the computational time.

Figure 8(a) shows a plot of an EROS-2 light curve, lm0090m4818. Figure 8(b) shows the same light curve folded with a period of 1.54192 days. The black dotted lines mark the band divisions on the magnitude axis. The shaded region shows the best band in terms of the first derivatives criterion. Figure 9 shows a plot of the spectral window function of the time instants extracted from the best band of lm0090m4818. The true period of the light curve is associated with the eighth-highest local maximum of the spectral window. In this case, if $N_t > 8$, then the underlying period will be within the trial period set that is to be evaluated by the CKP in the next step of the pipeline.

### 3.2. Performance Criteria

The task of discriminating periodic light curves can be viewed as a binary classification problem where the classes are periodic (true) and nonperiodic (false) light curves. In this case, true positives (TPs) are the periodic light curves classified as periodic, false positives (FPs) are the nonperiodic light curves classified as periodic, true negatives (TNs) are the nonperiodic light curves classified as nonperiodic, and false negatives (FNs) are the periodic light curves classified as nonperiodic.

To evaluate the performance of our method, we use the definitions of recall, $r$, precision, $p$,

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{15}$$

and $F$-score,

$$F_\beta = \frac{(1 + \beta^2)\, p\, r}{\beta p + r}. \tag{16}$$

The denominator of $r$ in Equation (15) corresponds to the number of periodic light curves in the data set. Recall is the ratio of recovered periodic light curves over the total number of periodic light curves in the data set. The denominator of $p$ in Equation (15) corresponds to the number of light curves that are classified as periodic. Precision, or completeness, is the ratio of recovered periodic light curves over the total number

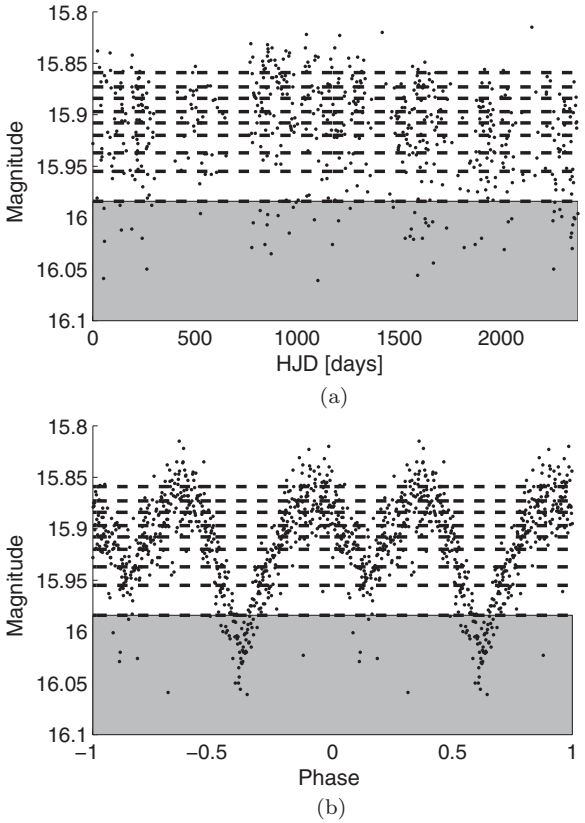


**Figure 8.** (a) EROS-2 light curve lm0090m4818. The dotted lines show the band divisions. The shaded region shows the best band in terms of the first derivatives criterion. (b) Same light curve folded with a period of 1.54192 days.

of light curves that are classified as periodic. The $F$-score (Equation (16)) is a weighted average of recall and precision. The parameter $\beta$ controls the importance of recall over precision on the weighted average. In what follows we use the $F_1$ score ($\beta = 1$).

We also define hit rate as

$$\mathrm{HR} = \frac{\mathrm{TP}^*}{\mathrm{TP}^* + \mathrm{FN}}, \tag{17}$$

where $\mathrm{TP}^*$ are the periodic light curves classified as periodic and at the same time the true period is recovered.[14]

## 4. SYNTHETIC LIGHT CURVES

In order to evaluate the actual efficiency of the system and determine the true number of periodics in our data set, we build a synthetic set containing both nonperiodic and periodic light curves.

*Periodic set.* The periodic synthetic light curves are generated using a multivariate Gaussian generative model with a covariance matrix similar to the periodic kernel in Equation (8). To generate a periodic synthetic light curve, with period $P$, signal-to-noise ratio (S/N) $S$, and smoothness $\sigma$, we follow the procedure below.

1. Randomly select a light curve from the database and extract its time instants $\{t_i\}$ and error bars $\{e_i\}$. This defines the number of samples, $N$, of the generated light curve.

[14] Note that a light curve can be classified as periodic even if the true period is not recovered, such as when a multiple of the true period is found.

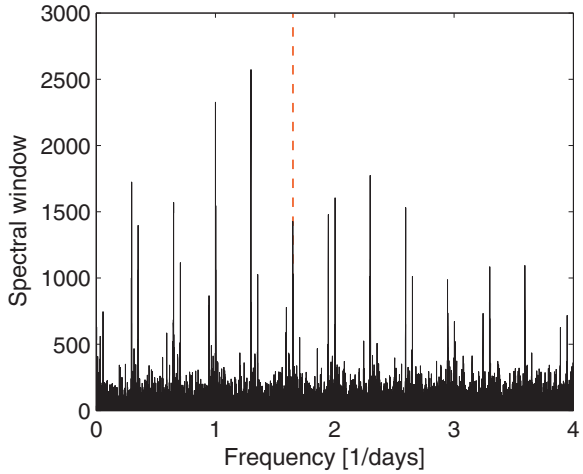


**Figure 9.** Spectral window of the tenth band from light curve lm0090m4818. The red dotted line shows the location of the underlying period ($1/P = 0.6485$). The underlying period is associated with the eighth-highest local maximum of the spectrum.

2. Use the time instants $\{t_i\}$, period $P$, and smoothness $\sigma$ and generate an $N \times N$ covariance matrix as

$$\Sigma_1(i, j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left( - \frac{2\sin^2(\pi(t_i - t_j)/P)}{\sigma^2} \right).$$

3. Generate a random periodic vector, $Y_s$, of length $N$ using a multivariate normal random generator with $N \times 1$ zero mean vector and $\Sigma_1$ covariance matrix.
4. Use the error bars to generate an $N \times N$ diagonal covariance matrix with diagonal elements,

$$\Sigma_2(i, i) = e_i^2.$$

5. Generate a random noise vector $Y_n$ of length $N$ using a multivariate normal random generator with an $N \times 1$ zero mean vector and $\Sigma_2$ covariance matrix.
6. The synthetic light curve $Y$ is obtained by summing the noise vector and the signal vector as follows:

$$Y = S\frac{\mathrm{med}(e_i)}{0.7413\mathrm{iqr}(Y_s)} Y_s + Y_n, \tag{18}$$

where $S$ is the desired signal-to-noise ratio, med is the median function, and iqr is the interquartile range. Note that the resulting light curve has S/N $S$ by construction.

For our purpose we generated a set of 10,000 synthetic periodic light curves, using the following parameter ranges.

1. Ten linearly spaced values for $\sigma$ in the range [0.1, 0.6].
2. Twenty logarithmically spaced values for $P$ in the range [0.4, 1000] days.
3. Ten values for $S$ extracted from the distribution of the S/N of EROS-2 light curves.

Five synthetic light curves are generated for each combination of $S$, $P$, and $\sigma$.

We present examples of the synthetic light curves generated using this procedure in Figure 10. Figure 10(a) shows a synthetic light curve with a period of 2.432 days, a smoothness value of 0.2 and an S/N of 10. Using a low smoothness value yields a shape with many features. Owing to the high S/N, the periodicity is very clear. Figure 10(b) shows a synthetic light curve with a

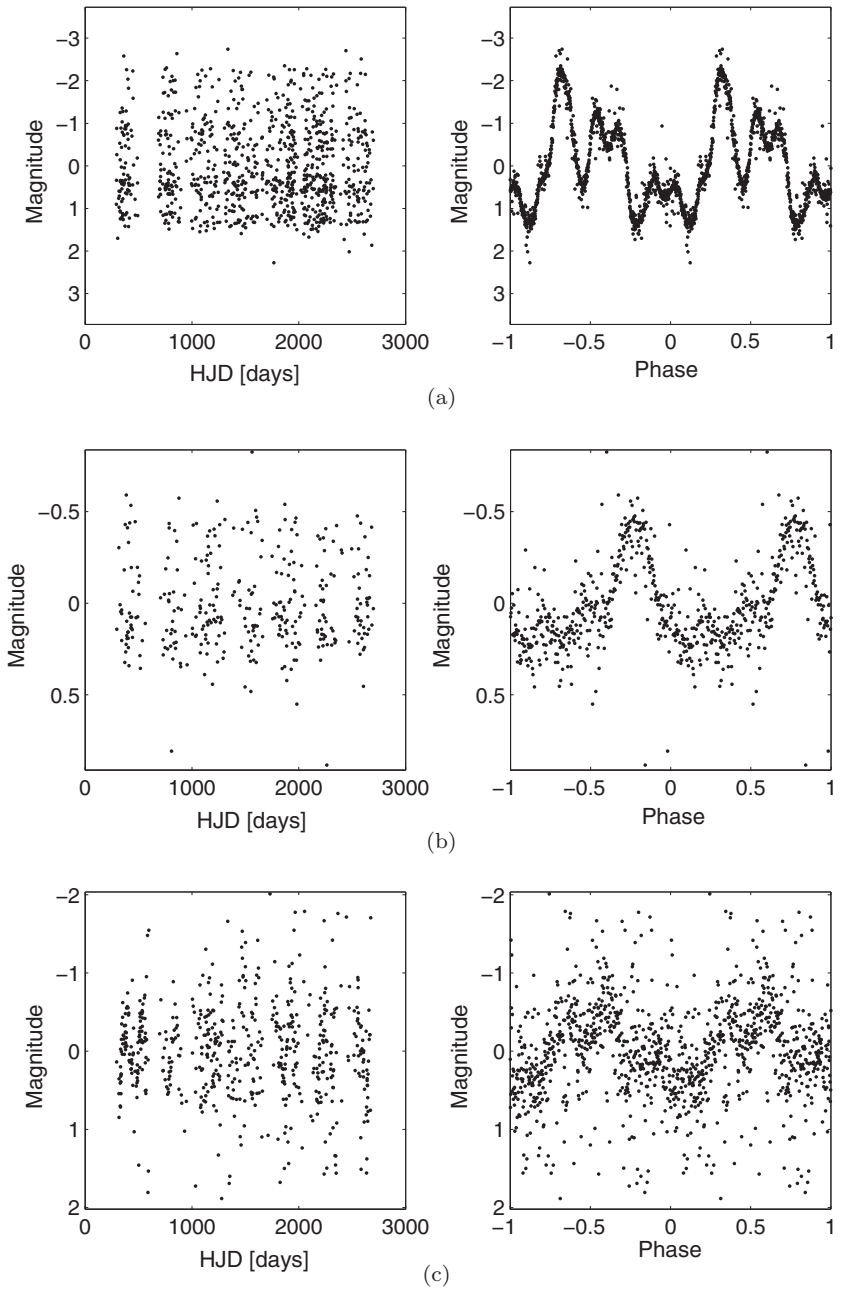


**Figure 10.** Example of synthetic periodic light curves. (a) Light curve created using $P = 2.432$ days, $\sigma_t = 0.2$, S/N = 10, and $N = 642$. (b) Light curve created using $P = 10.24$ days, $\sigma_t = 0.5$, S/N = 4, and $N = 342$. (c) Light curve created using $P = 154$ days, $\sigma_t = 0.4$, S/N = 2, and $N = 932$.

period of 10.42 days, smoothness of 0.5, and S/N of 4. In this case, a higher $\sigma$ value yields a smoother shape, as seen in the folded light curve. Figure 10(c) shows a synthetic light curve with a period of 154 days, smoothness of 0.4, and S/N of 2.

*Nonperiodic set.* The nonperiodic synthetic light curves are generated using block-bootstrap surrogates (Schmitz & Schreiber 1999; Schreiber & Schmitz 1999; Buhlmann 1999). The procedure to generate a nonperiodic synthetic light curve is as follows.

1. Randomly select a light curve and extract its time instants $\{t_i\}$ and error bars $\{e_i\}$. This defines the number of samples $N$ of the generated light curve.
2. Compute the slotted autocorrelation function (ACF; Edelson & Krolik 1988) of the light curve.
3. Find the time lag associated with the ACF value of $\exp(-1)$; this time lag is used as the block length (BL) for the block bootstrap method below.
4. Until at least $N$ magnitude values have been created, do the following.
   (a) Randomly select the block starting point $i_s$, such that $i_s \in [1, N - N']$. Find $N'$ as the last light curve sample that complies with
   $$t(N) - t(N') > \text{BL}.$$
   (b) Find the end point of the block $i_e$ as the first time instant that complies with
   $$t(i_e + 1) - t(i_s) > \text{BL}.$$

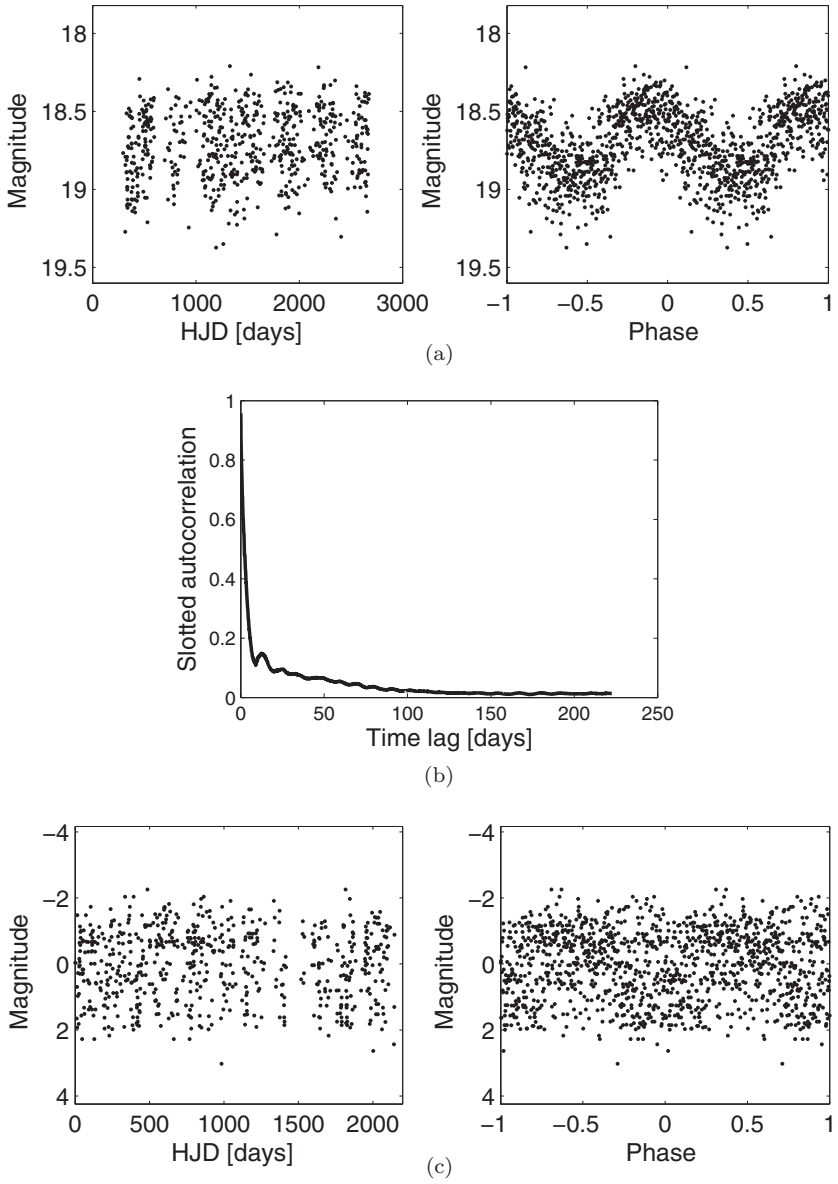


**Figure 11.** (a) Periodic light curve EROS-2 lm0090l27524 folded with the period of 0.337443 days; this period has a CKP value of 2.7424. (b) Slotted autocorrelation function of light curve lm0090l27524. Using the slotted ACF, a window length of 3.67 days is selected to create the surrogates. (c) Surrogate created from lm0090l27524. The CKP value of the surrogate is 0.4532, which is below the corresponding periodicity threshold.

(c) Grab the time instants, magnitudes, and error bars of the original light curve segment in $[i_s, i_e + 1]$.
(d) Subtract the initial time $t_{i_s}$ from the selected time instants. After this, the block starts at zero days.
(e) Add the time from the previous block $t_{\rm PB}$ to the selected time instants ($t_{\rm PB} = 0$ for the first block). After this, the block starts where the last block ended.
(f) Update $t_{\rm PB} = t(i_e + 1)$. Delete the time instant, magnitude, and error bar of sample $i_e + 1$ from the block.
(g) Add the newly constructed block to the surrogate.

For each EROS-2 light curve selected, 10 surrogates were created. A total of 10,000 EROS-2 light curves were used to create a training set of 100,000 nonperiodic synthetic light curves. To demonstrate that the resulting surrogates are not periodic and to retain the same spectra characteristics as the original light curves, we perform the procedure described above with a light curve of a periodic star. Figure 11(a) shows EROS-2 light curve lm0090l27524 folded with a period of 0.337443 days. The associated CKP value is 2.7424. The block bootstrap method was used to create a nonperiodic synthetic light curve. Figure 11(b) shows the slotted ACF, and the block length selected for this light curve is 3.67 days. Ten surrogates are generated using the procedure described above. Figure 11(c) shows one of the surrogates. The surrogate is folded with its best period, and clearly the periodicity of the original light curve is not retained by the surrogate.

### *4.1. Obtaining the Periodicity Discrimination Thresholds*

A light curve is labeled as periodic if the CKP value associated with its best trial period is above a given periodicity discrimination threshold. We determine the threshold by optimizing the $F_1$ score (Equation (16)) with a training set created as described above and following the guidelines in Section 3.2. The periodicity threshold is a function of the S/N, and therefore we obtain a periodicity threshold per S/N. To do so, the S/N values are

**Table 1**
Periodicity Thresholds and Associated Precision and Recall Values for Each S/N Bin

| $S$ | th($S$) | max $F$-score | $p(S)$ [%] | $r(S)$ [%] |
|---|---|---|---|---|
| [0, 1.5] | 0.4584 | 0.92 | 94.26 | 89.15 |
| [1.5, 2] | 0.4565 | 0.94 | 95.14 | 92.15 |
| [2, 2.5] | 0.4537 | 0.95 | 96.42 | 92.98 |
| [2.5, 3.5] | 0.4581 | 0.96 | 96.82 | 94.26 |
| [3.5, 5] | 0.5875 | 0.97 | 97.52 | 96.12 |
| [5, 10] | 1.1153 | 0.98 | 98.12 | 97.51 |
| [10, 20] | 1.6464 | 0.98 | 98.22 | 97.81 |
| [20, ∞] | 2.4112 | 0.97 | 98.54 | 96.15 |

discretized in nine bins, $S = \{[0, 1.5], [0, 1.5], [1.5, 2], [2, 2.5], [2.5, 3.5], [3.5, 5], [5, 10], [10, 20], [20, \infty]\}$, and we compute the periodicity threshold according to the following procedure.

1. Evaluate the CKP values for each light curve in the training set whose S/N falls in bin $S$.
2. Construct a threshold array of 5000 points in [min(CKP), max(CKP)].
3. Compute the $F_1$ score (Equation (16)) at each threshold value.
4. Select the threshold th($S$) as the CKP value that maximizes the $F_1$ score.

Once the thresholds have been computed, a light curve whose S/N falls in bin $S$ is labeled as periodic if

$$\mathrm{CKP}(P_{\mathrm{best}}) > \mathrm{th}(S),$$

where $P_{\mathrm{best}}$ is the detected period that maximizes the CKP for the given light curve.

### 4.2. Estimating the True Number of Periodic Light Curves

In this section, we elaborate on how to estimate the number of periodic light curves in a data set. This is not to be confused with the number of light curves labeled as periodic by the proposed method. The true number of periodic light curves in a data set, $N_p$, is the number of true positives plus the false negatives, which is equivalent to the denominator of $r$ in Equation (15). The number of light curves classified as periodics, $\tilde{N}_p$, is the number of true positives plus false positives, which is equivalent to the denominator of $p$ in Equation (15).

Using Equation (15), we can estimate the actual number of periodics in a given S/N bin $S$ as

$$N_p(S) = \tilde{N}_p(S)\frac{p(S)}{r(S)}, \tag{19}$$

where $p(S)$ and $r(S)$ are the precision and recall values for bin $S$, respectively, which we assume we can determine from the training set. The precision and recall values are computed following the procedure given in Section 4.1. Given an $\tilde{N}_p$, we can estimate the true number of periodic light curves in a data set as

$$\tilde{N}_p = \sum_S \tilde{N}_p(S)\frac{p(S)}{r(S)}. \tag{20}$$

Table 1 shows the thresholds th($S$), associated $F$-score, and recall and precision values obtained for each S/N bin $S$. The overall precision and recall (across the S/N bins) are 95.3% and 92.7%, respectively.

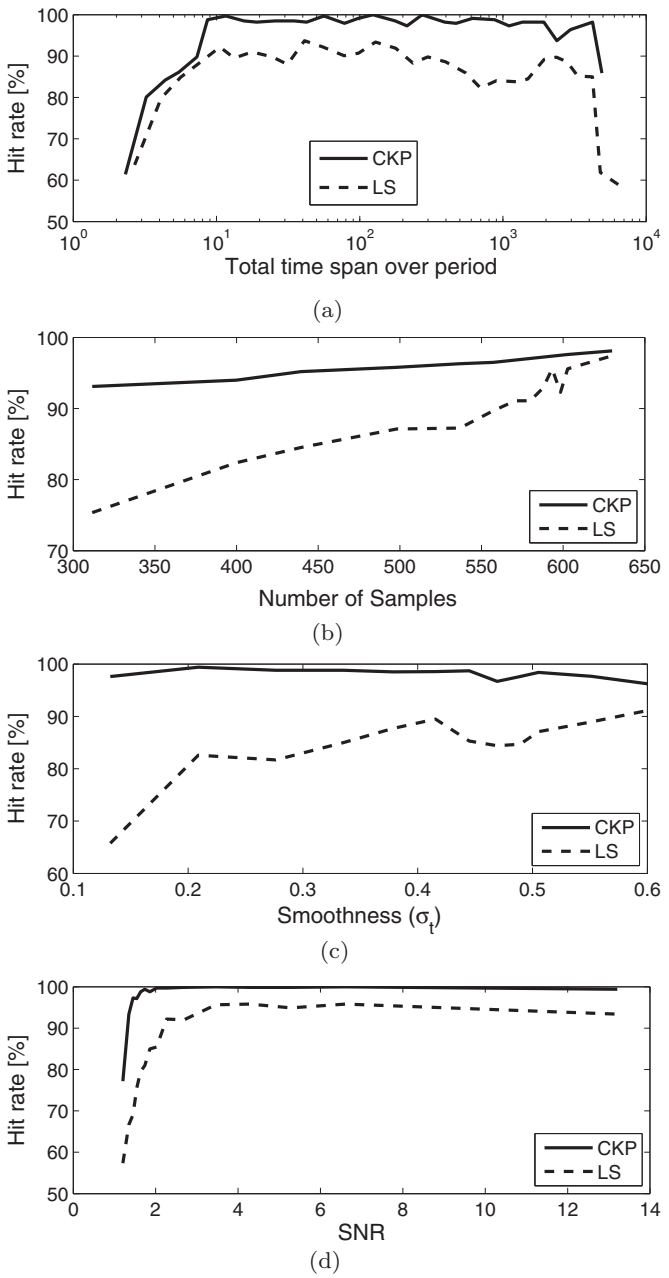


**Figure 12.** Hit rate in the synthetic periodic light curves as a function of the value of the parameters used to generate the set. The parameters are (a) the number of samples, (b) the smoothness, (c) period over total time span, (d) and S/N. The proposed method is compared with the LS periodogram.

### 4.3. Efficiency as a Function of Parameters

In the following tests, we assess the efficiency of the proposed method as a function of the parameters of the synthetic light curves. Hit rate (Equation (17)) is measured as a function of the total time span divided by the period, number of samples, smoothness, and S/N for the 10,000 synthetic periodic light curves. Hit rates are computed as a function of one of the parameters while summing for the other three. The CKP is compared with the LS periodogram on each test.

Figure 12(a) shows a plot of the HR as a function of the ratio between the total time span of the light curve and its period (T/P). The total time span of the light curves in the EROS-2 survey is approximately 2500 days, and the sampling rate is approximately 1.2 samples per day. The ratio T/P can

be viewed as the number of times the underlying signal repeats itself. The period range in the training set goes from 0.4 days to 1000 days. HR is stable across the given range except for T/P below 10 and above 2300. Intuitively, the fewer times a signal is repeated across T, the more difficult it is to assess its periodicity. This can be seen in the plot for periods above 280 days. There is also a limit in the resolution due to the sampling rate, which is reflected as a hit rate drop for periods below 0.5 days. The same hit rate drop can be observed for the LS periodogram.

Figure 12(b) shows a plot of HR as a function of the number of samples of the synthetic light curve. HR increases with the number of samples. The hit rate rises by 5% when the number of samples increases from 300 to 600. In comparison with the LS periodogram, the CKP is less affected by $N$. Intuitively, the less information available on the process, the harder it is to assess its periodicity.

Figure 12(c) shows a plot of the hit rate as a function of the smoothness ($\sigma$) of the synthetic light curves. The hit rate is stable across the given range, decreasing slowly for the very large and very small values of $\sigma$. Overall, the smoothness does not have great influence on the CKP hit rate. The LS-periodogram hit rate increases with $\sigma$. This is expected, as smaller values of $\sigma$ produce light curves with highly nonsinusoidal shapes, as shown in Figures 10(a), (b) and (c).

Finally, Figure 12(d) shows a plot of the hit rate as a function of the S/N (Equation (22)) of the synthetic light curves. HR is stable for the given S/N range, dropping abruptly for S/N below 1.8. For S/N of 1.2 the hit rate has decreased by almost 25%. A similar behavior can be seen for the LS periodogram.

## 5. DATA

### 5.1. Description of the Data

The EROS-2 project (Tisserand et al. 2007; Rahal et al. 2009) was designed to search for gravitational microlensing events caused by massive compact halo objects (MACHOs) in the halo of the Milky Way. To do this, 32.8 million stars in the Magellanic Clouds were surveyed over 6.7 yr. The objective of the EROS-2 survey was to test the hypothesis that MACHOs were a major component of the dark matter present in the halo of our galaxy.

The EROS-2 project surveyed 28.8 million stars in the Large Magellanic Cloud (LMC) and 4 million stars in the Small Magellanic Cloud (SMC), distributed in 88 and 10 observational fields, respectively. Each field is divided into 32 chips (8 CCDs and 4 quadrants per CCD). Each light curve file has five columns: time instant, red channel magnitude, red channel error bars, blue channel magnitude, and blue channel error bars. In what follows, only the blue channel is used. The average number of samples per light curve is 430 and 780 in the LMC and SMC, respectively.

### 5.2. Preprocessing and Intricacies of the Data

*Fixing the error bars.* As described above, the kernel size was estimated using the error bars of the magnitudes or the estimate of the observational errors. If these observational errors were underestimated or overestimated (as is often the case), the kernel size will be also wrongly estimated. For example, if the error bars are for some reason underestimated, then the kernel bandwidth will also be underestimated and will not account of the true scatter of the light curve, resulting in low CKP values.

For a light curve that is not variable the sample variance and the error bars should have very similar values. Another way of expressing this is that for a given nonvariable light curve the median of the error bars should be equal to the interquartile range. Since we know that most sources are not variable, a plot of those two quantities should be distributed around the bisector (line with slope of 1). Figure 13(a) shows a plot of the median of the error bars as a function of the interquartile range of the magnitudes for a randomly selected chip, lm0090k. Each dot corresponds to a light curve. The locus of the points (light curves with magnitudes between 17 and 21) is over the bisector, i.e., the error bars are larger than the dispersion of the light curve. This is an example of a field with overestimated error bars.

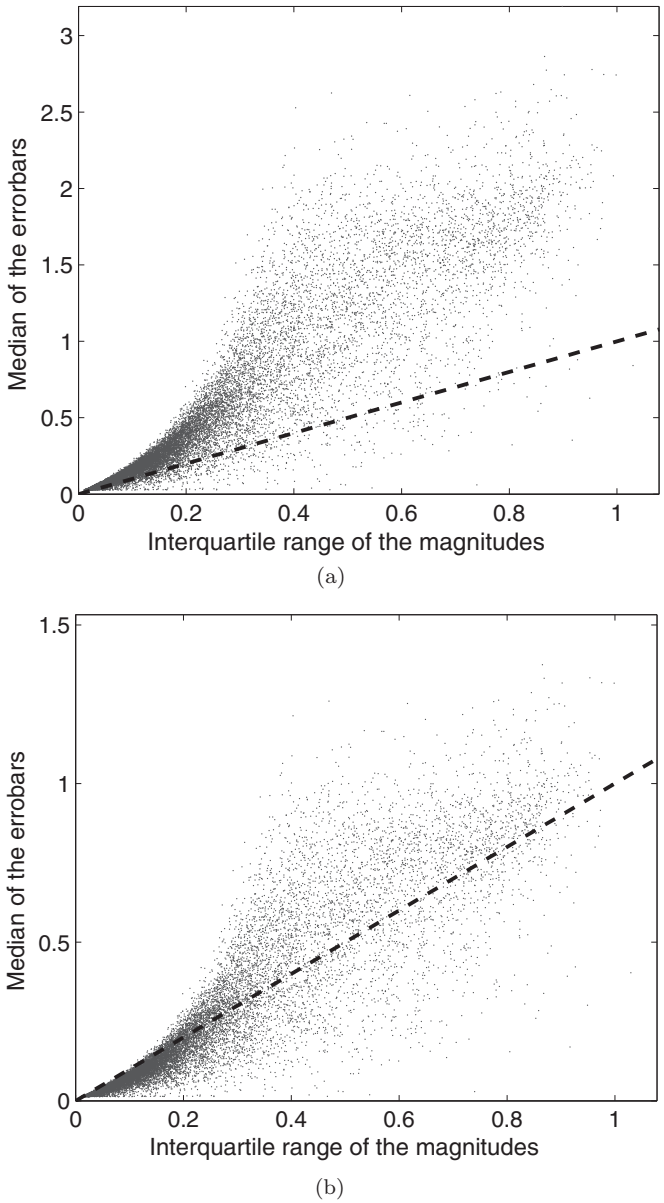


**Figure 13.** (a) Median of the magnitude's error bars as a function of the interquartile range of the magnitudes for chip lm0090k. The dotted line has a slope of 1. The error bar correction factor for lm0090k is 0.42. (b) Same plot after correcting the error bars.

For a given field with $N_{\rm lc}$ light curves, the error bar *correction factor* is defined as the constant that minimizes

$$\alpha_{\rm cf} = \arg\min_{\alpha} \sum_{k=1}^{N_{\rm lc}} (\mathrm{iqr}(\{y\}_k) - \alpha \mathrm{med}(\{e\}_k))^2 , \qquad (21)$$

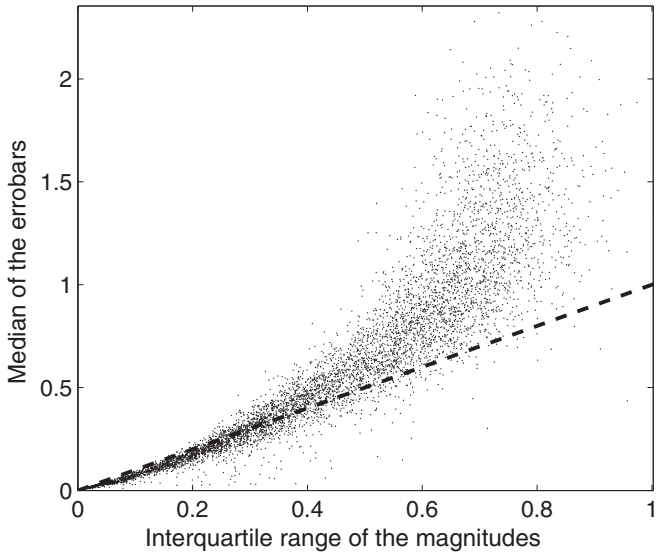


**Figure 14.** Median of the magnitude's error bars as a function of the interquartile range of the magnitudes for chip lm0140k. The error bar correction factor for lm0140k is ∼1.

where $\{y\}_k$ and $\{e\}_k$ are the magnitudes and error bars of light curve $k$, respectively, iqr is the interquartile range, and med is the median.

For the field shown in Figure 13(a) an error bar correction factor of 0.42 is obtained. Figure 13(b) shows the plot of the same field after correcting the error bars. Figure 14 shows the same plot for chip lm0140k. This chip is on the periphery of the LMC. The error bar correction factor for this field is ∼1, i.e., there is no need for correction.

Using the error bar correction factor, we define the pseudo-signal-to-noise ratio (pS/N) of a given light curve as

$$\mathrm{pS/N} = \frac{0.7413\mathrm{iqr}(\{y\})}{\alpha \mathrm{med}(\{e\})}, \tag{22}$$

where $y$ and $e$ are the magnitudes and error bars, respectively, and $\alpha$ is computed per field using Equation (21).

*Removing outliers and bad points.* The mean $\bar{e}$ and the standard deviation $\sigma_e$ of the error bars are computed per light curve, and samples that do not comply with

$$e_i < \bar{e} + 3\sigma_e,$$

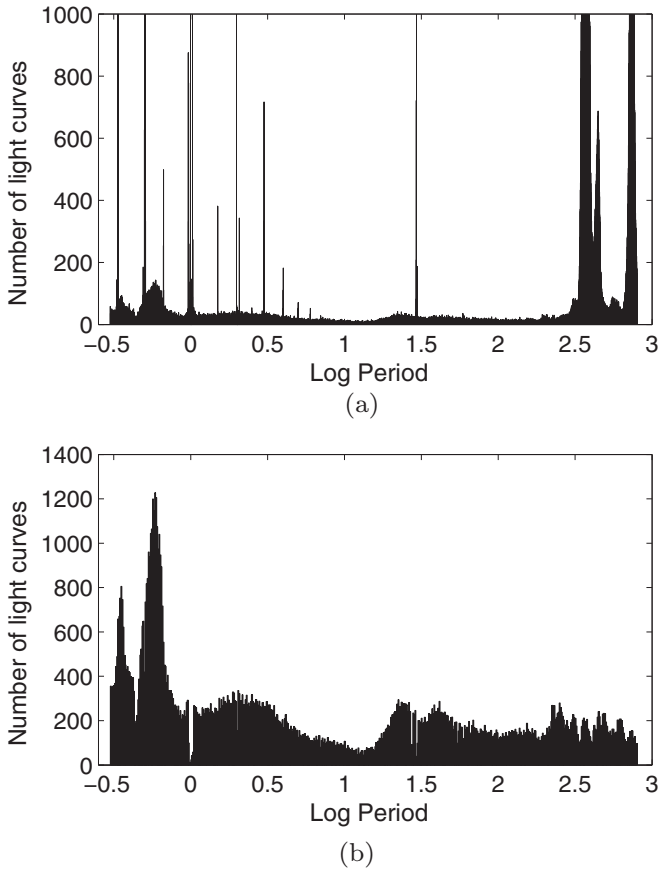


**Figure 16.** (a) Histogram of the periodic light curves detected with the proposed method on the LMC. The spurious periods have not been filtered in these results. The vertical columns correspond to the spurious periods, their multiples, and their aliases. (b) Histogram of the periodic light curves detected in the LMC after carrying out the spurious period removal scheme.

where $e_i$ is the error bar of a sample $i$, are removed from the light curve. At this point, light curves with less than 50 samples are discarded from the analysis.

*Simple detrending.* After that, the coefficients of a least-squares linear $\chi^2$ regression on the magnitudes are computed:

$$\chi^2 = \sum_{i=1}^{N} \frac{(a_0 + a_1 t_i - x_i)^2}{e_i^2}, \tag{23}$$

where $a_0$ is the intercept and $a_1$ is the slope. The coefficients of the linear fit are obtained by differentiating Equation (23) with

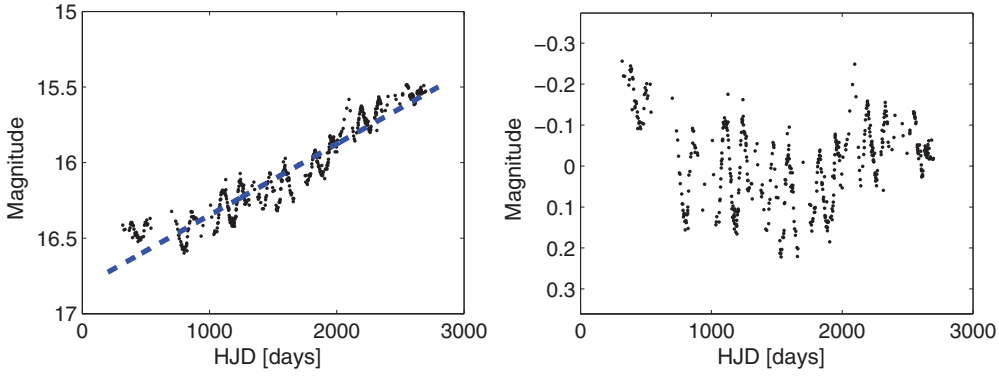


**Figure 15.** (a) Light curve lm0324k13673 from the EROS-2 survey. A linear $\chi^2$ fit is computed for this light curve (blue dotted line). The correlation coefficient for the linear fit is 0.9493. (b) Light curve lm0324k13673 after the linear trend subtraction.

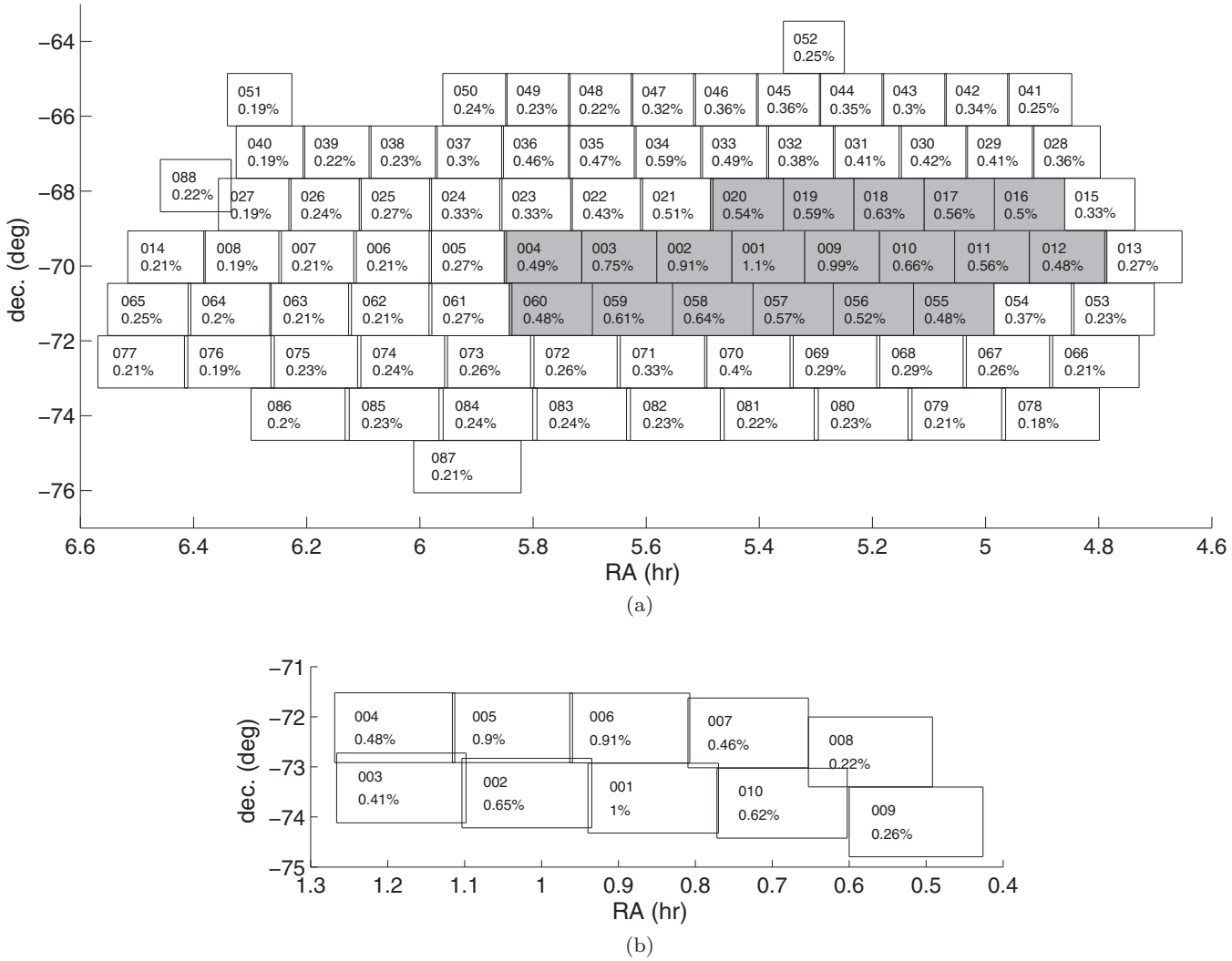


**Figure 17.** Maps of the (a) EROS-2 LMC and (b) SMC fields. The percentage of periodic light curves is shown below the name of the field.

respect to $a_1$ and $a_0$. The linear $\chi^2$ fit is subtracted from the light curve only if the correlation coefficient between the light curve and its linear fit is above 0.5 (goodness of fit). Figure 15(a) shows EROS-2 light curve lm0324k13673. The signal is mounted on a monotonically increasing linear trend. The dotted line in Figure 15(a) shows the $\chi^2$ linear fit. Figure 15(b) shows the light curve after the linear fit subtraction; further evaluation shows that the light curve is periodic with a period of 120.38 days.

## 6. RESULTS

### 6.1. Filtering of Spurious Periods

The set of trial periods obtained with the bands method and evaluated using the CKP (Equation (13)) contain spurious periods related to the solar day, the moon phase, the year, and their multiples. These spurious periods need to be filtered in order to find the true underlying periodicity of the light curve. Additional spurious periods were found by analyzing the histogram of the periodic light curves detected by the proposed method (Figure 16(a)). These additional spurious periods, which are given in Table 2, correspond to aliases of the known spurious periods.

A Gaussian mask centered around the spurious period is created for each of the spurious periods. Periods whose CKP falls inside the masks are filtered as spurious periods. The standard deviation and the amplitude of the masks are set so that the associated spurious peak in the period histogram is flattened.[15] The trial period that maximizes the CKP and does

**Table 2**
Description of the Spurious Periods

| Period (days) | Description |
|---|---|
| 1 | Solar day ($P_d$) |
| 29.5305 | Moon phase or Synodic month ($P_m$) |
| 365.24 | Tropical year ($P_y$) |
| 2, 335 | Average time span of EROS-2 light curves ($T$) |
| 0.4917 | $((P_d/2)^{-1} + P_m^{-1})^{-1}$ |
| 0.5086 | $((P_d/2)^{-1} - P_m^{-1})^{-1}$ |
| 0.9672 | $(P_d^{-1} + P_m^{-1})^{-1}$ |
| 1.0351 | Lunar day, $(P_d^{-1} - P_m^{-1})^{-1}$ |
| 0.9973 | Sidereal day, $(P_d^{-1} + P_y^{-1})^{-1}$ |
| 1.0027 | $(P_d^{-1} - P_y^{-1})^{-1}$ |
| 27.31 | Sidereal month, $(P_m^{-1} + P_y^{-1})^{-1}$ |
| 32.13 | $(P_m^{-1} - P_y^{-1})^{-1}$ |
| 315.65 | $(P_y^{-1} + T^{-1})^{-1}$ |
| 432.63 | $(P_y^{-1} - T^{-1})^{-1}$ |

[15] The parameters of the filters can be found alongside the catalogs at http://timemachine.iic.harvard.edu.

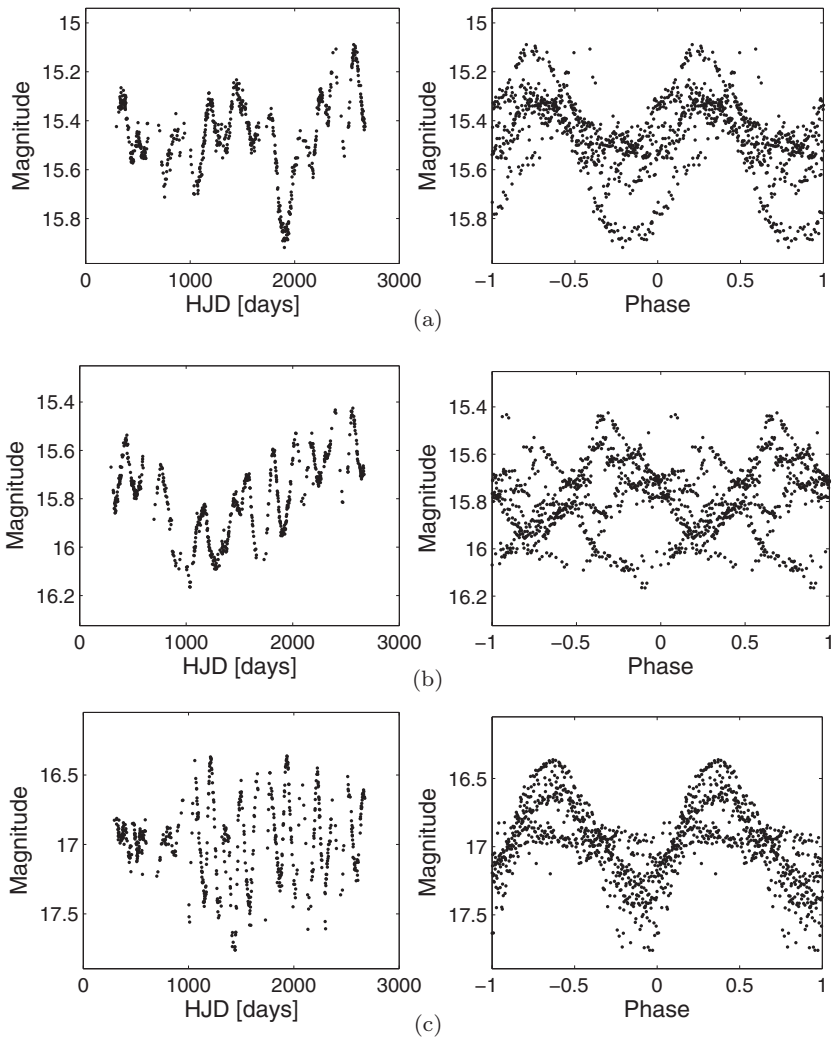


**Figure 18.** These light curves are examples of the false positives found in the catalogs. (a) Light curve lm0090n29655, folded with the detected period of 278 days, is an example of quasi-periodic behavior. (b) Light curve lm0091119300, folded with the detected period of 264 days, is mounted on a polynomial trend in the mean. (c) Light curve lm0090n6107, folded with the detected period of 144 days, varies in amplitude across the time span.

**Table 3**
Characteristics of Selected Fields

| Field | Number of Light Curves | Average $N$ | Average S/N |
|---|---|---|---|
| lm009 | 109,802 | 548 | 1.628 |
| lm012 | 95,010 | 447 | 0.959 |
| sm001 | 92,666 | 830 | 1.505 |

**Table 4**
Results in the Selected EROS-2 Survey Fields

| Field | $\tilde{N}_p$ | FP | Prec. (%) | FN | Recall (%) | Observed $N_p$ | Synthetic $N_p$ |
|---|---|---|---|---|---|---|---|
| lm009 | 1160 | 41 | 96.47 | 66 | 94.43 | 1185 | 1189 |
| lm012 | 718 | 30 | 95.82 | 51 | 93.10 | 739 | 743 |
| sm001 | 1564 | 69 | 95.59 | 99 | 93.79 | 1594 | 1637 |

not fall in any of the spurious period masks is selected as the best trial period for the light curve.

### *6.2. Results for Selected Fields*

In this experiment the proposed method is evaluated on three fields from the EROS-2 survey. The objectives are to measure the accuracy of the method and to compare the number of periodic light curves in the fields with the expected number of periodic light curves computed from the synthetic results by performing visual inspection to a large but manageable number of light curves. The first six chips from fields lm009, lm012, and sm001 are used in this experiment. Table 3 shows the number of light curves, the average number of samples, and the average S/N from the selected fields.

Table 4 shows the results obtained for the selected fields. Column (2) ($\tilde{N}_p$) corresponds to the number of light curves labeled as periodic by our method. These light curves are folded with the detected period and visually checked in order to find the number of false positives (Column (3)). Column (4) is the precision in the detected periodic light curve set. Column (5) gives an estimate of the false negatives (FNs) in the field. The FNs are estimated by visually inspecting the folded light curves of the objects that are below the periodicity thresholds. Because it is impracticable to check all the nonperiodic objects, the search for FNs is stopped if 50 consecutive nonperiodic light curves are found for each S/N bin. Column (6) is the recall calculated using the observed number of true positives ($\tilde{N}_p$-FP) and the

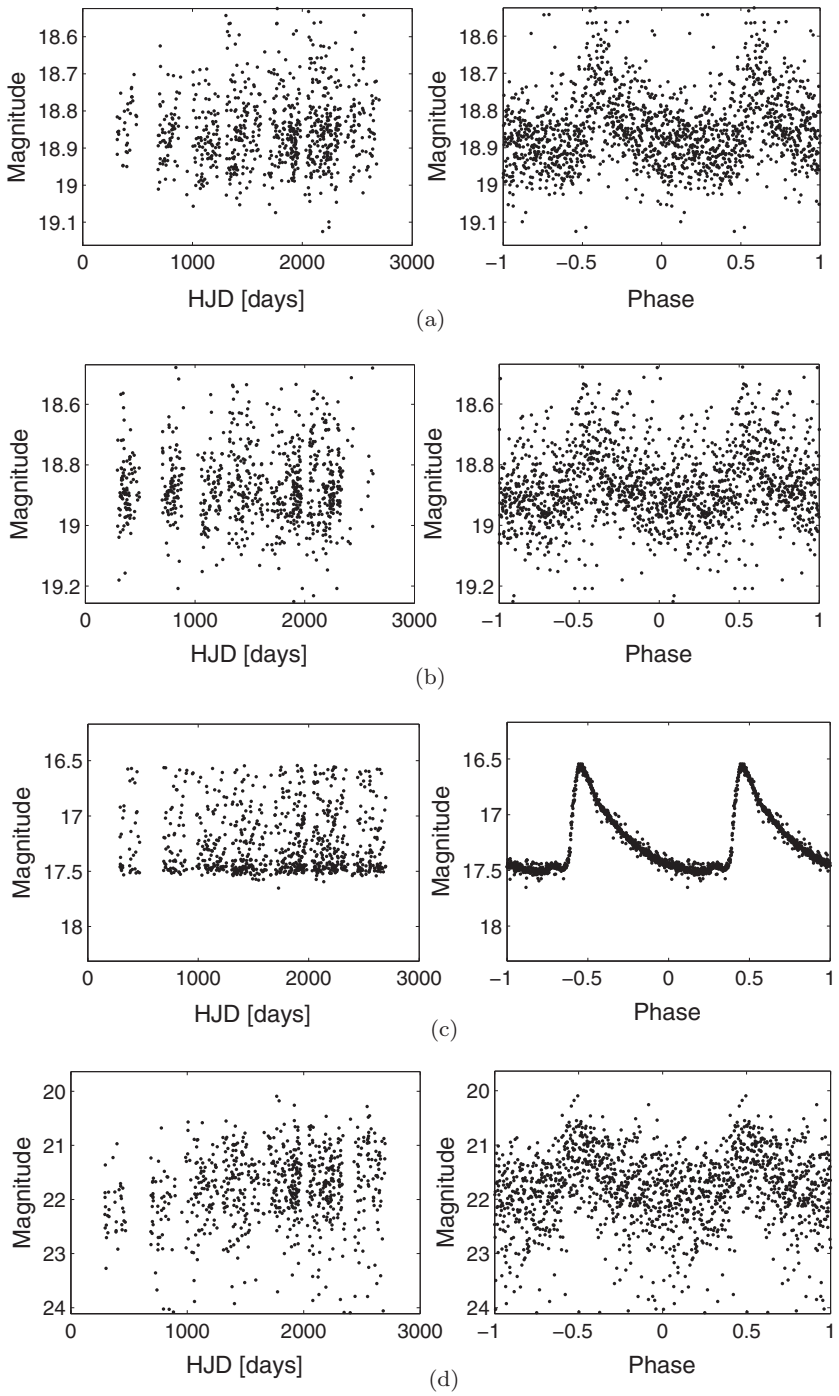


**Figure 19.** Examples of overlapping and blending. A period of 2.4796 days is detected for light curves (a) sm0077n17908 and (b) sm0010k3199. The angular distance between these light curves is 0.″5. Their differences in magnitude and CKP value are 0.03 and 0.23, respectively. These light curves are associated with a star that is in an overlapped region between fields sm001 and sm007. Light curves (c) sm0023n10183 and (d) sm0023n10325 are also found to have the same period (1.2535 days), but they reside in the same field. Their angular distance, $\delta$-magnitude, and $\delta$-CKP are 4.″9, 4.5, and 4.1, respectively. In this case the light from sm0023n10183 (c) introduces a periodicity in its neighbor (d).

FN. Column (7) corresponds to the observed number of periodic light curves ($\tilde{N}_p$ -FP+FN). Column (8) shows an estimation of the true number of periodic variables ($N_p$) using the synthetic precision and recall values given in Section 4.2. Column (7) is also an estimation of $N_p$ because the true amount of FNs is not known.

A grand total of 1160 periodic light curves is recovered from field lm009, which corresponds to 1.06% of the field. The percentage of periodic light curves in lm012 and sm001 is 0.75% and 1.69%, respectively.[16] The overall precision and recall in all the fields are within 2% of the overall precision and recall found in the synthetic data set. For comparison we ran the LS periodogram[17] on the lm009 field. The spurious periods are filtered as described in previous sections. The filtered

[16] These chips have a higher number of periodics than the average found in the LMC and SMC, as can be seen in Figure 17(a). This issue is discussed in the next section.

[17] The vartools software with the -LS option is used.

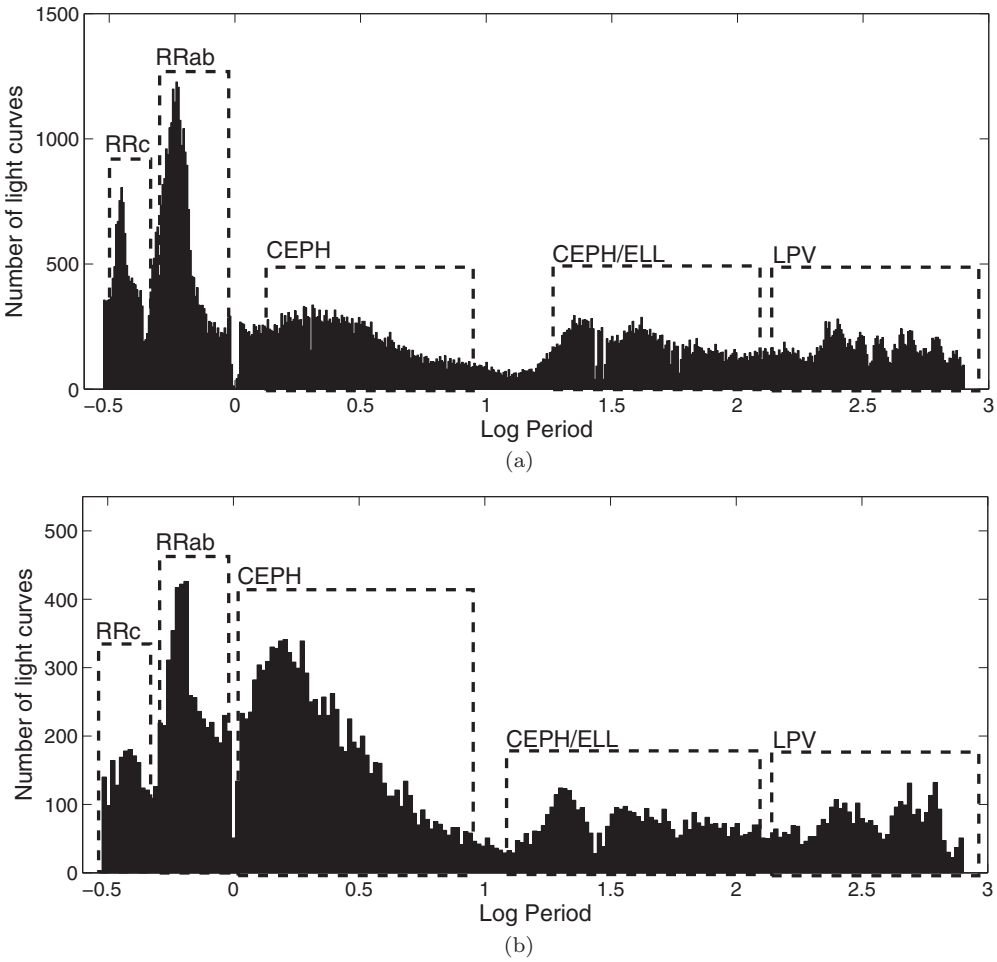


**Figure 20.** Histogram of the periods found in the (a) LMC and (b) SMC blue channel data. The regions marked with dotted boxes are associated with clusters of a given type periodic variable star.

periods found with the LS periodogram are sorted according to their normalized LS statistic. By imposing a threshold on this statistic, the periodic light curves obtained the CKP plus 298 false positives, and 14 additional true positives are obtained. This corresponds to a drop of 16.5% in precision and a negligible increase in recall (1%) with respect to the CKP.

It is important to note that there are periodic behaviors that are not captured in the proposed synthetic light curve set. Examples of these are periodicities mounted on polynomial trends, objects with more than one oscillation period, objects that are not periodic in the whole time span, and objects whose oscillation amplitude changes irregularly or following a modulation pattern, such as semiregular and irregular long period variables (LPVs). These cases are considered as nonperiodic during the inspection. Examples of these cases are shown in Figures 18(a), (b), and (c), which correspond to false positives found in field lm009. Currently, the proposed method is not able to discriminate quasi-periodicities or other irregular periodics.

### *6.3. Results on EROS-2 LMC and SMC Fields*

A total of 32.8 million light curves from the EROS-2 survey were processed with the proposed periodicity discrimination pipeline, 28.8 million from the LMC and 4 million from the SMC. Table 5 shows the summary of the results for the LMC and SMC. $\tilde{N}_p$ corresponds to the number of light curves labeled as periodic by our method. The *Discarded* column corresponds to the number of periodic light curves that appear twice in the list, owing to field overlapping and blending. Column $N_p$ corresponds to an estimation of the true number of periodic variables using the synthetic precision and recall values given in Section 4.2.

**Table 5**
Periodic Light Curve Discrimination Result Summary on the EROS-2 Survey

| | $N_{\rm LC}$ | $\tilde{N}_p$ | Discarded | $N_p$ | Periodics (%) |
|---|---|---|---|---|---|
| LMC | 28,797,305 | 120,983 | 2,663 | 121,147 | 0.42 |
| SMC | 4,064,179 | 24,920 | 1,817 | 24,855 | 0.61 |

To select the duplicate light curves, the nearest neighbor for each object in terms of angular distances is first identified. If the distance to the nearest neighbor is less than 10″ and both objects have the same period, then the light curve with the lowest magnitude is added to the discarded set. Using this criterion, 2663 pairs of light curves are selected from the LMC. From this set 336 correspond to light curves that reside in different chips. The average delta magnitude in this set is 0.281, and the average delta CKP is 0.744. Each pair of light curves corresponds to the same star that appears twice in the survey owing to the overlapping in the observational fields. The other 2327 cases correspond to light curves that are neighbors in the same chip. The average delta magnitude in this set is 2.15, and the average

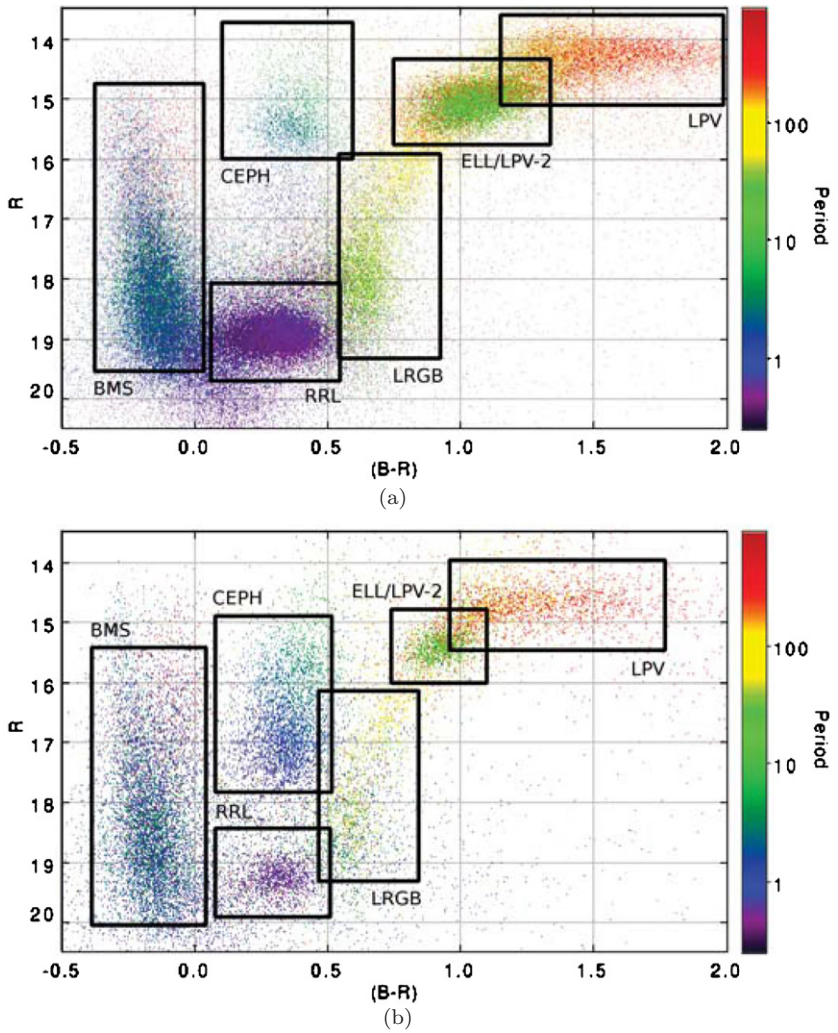


**Figure 21.** Color–magnitude diagram showing the periodic light curves found in the (a) LMC and (b) SMC. BMS corresponds to blue main sequence. LRGB corresponds to lower red giant branch. Black boxes mark the location of Cepheid, RR Lyrae, LPV, and ellipsoidal variable populations.

delta CKP is 3.02, much higher than in the previous set. In this set the more luminous star of the pair injects its periodicity in the light curve of the less luminous star (blending). Figure 19 shows an example of an overlapped pair and blended pair. It is interesting to note that 72% of the blended light curves are found in the fields within the LMC bar where the star density is the highest, while the overlapped light curves are equally distributed between bar and nonbar fields. In the SMC 1817 pairs of light curves are selected to be discarded. In this case 386 are due to field overlapping and 1431 are due to blending. The average delta magnitude in the overlapped light curves is 0.21, and the average delta CKP is 0.78. The average delta magnitude in the blended light curves is 2.34, and the average delta CKP is 4.86. The percentage of discarded light curves in the SMC is 7.2%, which is higher than the 2.3% found in the LMC. This again was attributed to the fact that SMC seeing is worse than LMC seeing, resulting in an overlapping PSF, which in turn resulted in correlated light curves.

Figure 17(a) shows a map of the 88 fields of the LMC. The shaded fields correspond to the LMC bar. The percentage of periodic light curves is shown for each field below its name. The fields corresponding to the LMC bar have a higher percentage of periodics. The percentage of periodics tends to drop the farther the field is from the LMC bar. Figure 17(b) shows a map of the 10 fields of the SMC where the same pattern is apparent. Because the cores of the LMCs have an older population of stars, it is known that one would expect more periodic stars in those regions.

A grand total of 118,320 and 23,103 periodic light curves are found from the LMC and SMC blue channel data, respectively. Using the recall and precision from the training data set, we estimate that the true number of periodic light curves is 121,147 for the LMC and 24,855 for the SMC. The percentage of light curves that are periodic is 0.42% in the LMC and 0.61% in the SMC.

Figure 20(a) shows the histogram of the periods found in the LMC blue channel data. Some of the known populations of periodic variables are identified in the histogram. The most notable populations correspond to c-type RR Lyrae (period centered in 0.3 days) and ab-type RR Lyrae (period centered in 0.6 days). These results are consistent with the RR Lyrae period histogram from the MACHO survey results on the LMC (Cook et al. 1995).

Figure 21(a) shows a color–magnitude diagram of the periodic light curves found in the LMC blue channel. The third axis corresponds to the detected period. The regions of interest are marked with black dotted squares. Examples of the periodic variable stars found in these regions are shown in Figures 29–32.

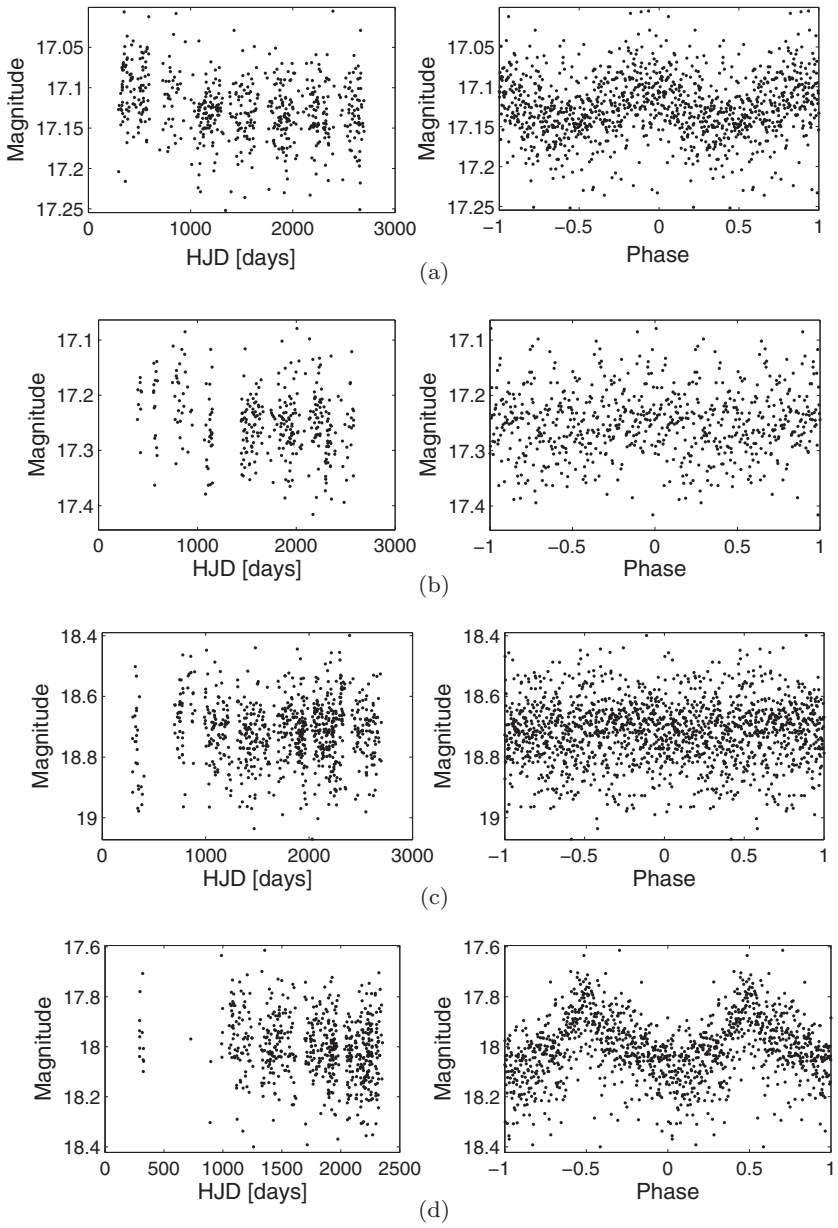


**Figure 22.** Examples of periodic light curves detected only in one of the EROS-2 channels. Panels (a) and (b) correspond to light curve lm0012k17912. (a) Blue channel light curve folded with the detected period of 0.48004 days. Using the red channel data, no strong periodicity is found. (b) Red channel light curve folded with the 0.48004 day periods. Panels (c) and (d) correspond to light curve sm0010l10270. (c) Blue channel data folded with the period detected in the red channel. (d) Red channel data folded with the detected period of 10.4453 days. Using the blue channel data, no strong periodicity is found.

These results are consistent with the color–magnitude diagram of the LMC periodic variables from the OGLE survey (Spano et al. 2009).

Figures 20(b) and 21(b) show the histogram of periods and the color–magnitude diagram of the periodic light curves found in the SMC blue channel, respectively. By comparing the histogram and color–magnitude diagram with those of the LMC, the following differences arise: the relative size of the Cepheid population is larger in the SMC, and the relative size of the c-type RR Lyrae population is larger in the LMC.

The red channel light curves are also analyzed for comparison purposes. A grand total of 87,025 and 14,501 periodic light curves are collected from the LMC and SMC red channel data, respectively. This represents a decrease of 30% with respect to the amount of periodics collected from the blue channel. By cross-matching the lists obtained from the blue and red channels in the LMC, we found that 68,179 objects appear in both lists, 50,141 objects are found only in the blue channel, and 18,846 objects are found only in the red channel. For the SMC, 12,536 objects appear in both lists, 1965 appear exclusively in the red, and 10,567 appear exclusively in the blue. For a given object the S/N may change between channels, as shown in the examples of Figure 22. By inspecting the histogram of the color $(B - R)_{\rm eros}$ of the EROS-2 light curves, it is clear that it is skewed to the blue side. The average color value in the LMC and SMC is 0.46 and 0.31, respectively, and therefore the S/N is higher in the blue channel; this explains why more periodics are found in the blue channel data.[18]

[18] Another reason could be related to the training scheme, in which only blue channel light curves where used to create the synthetic database.

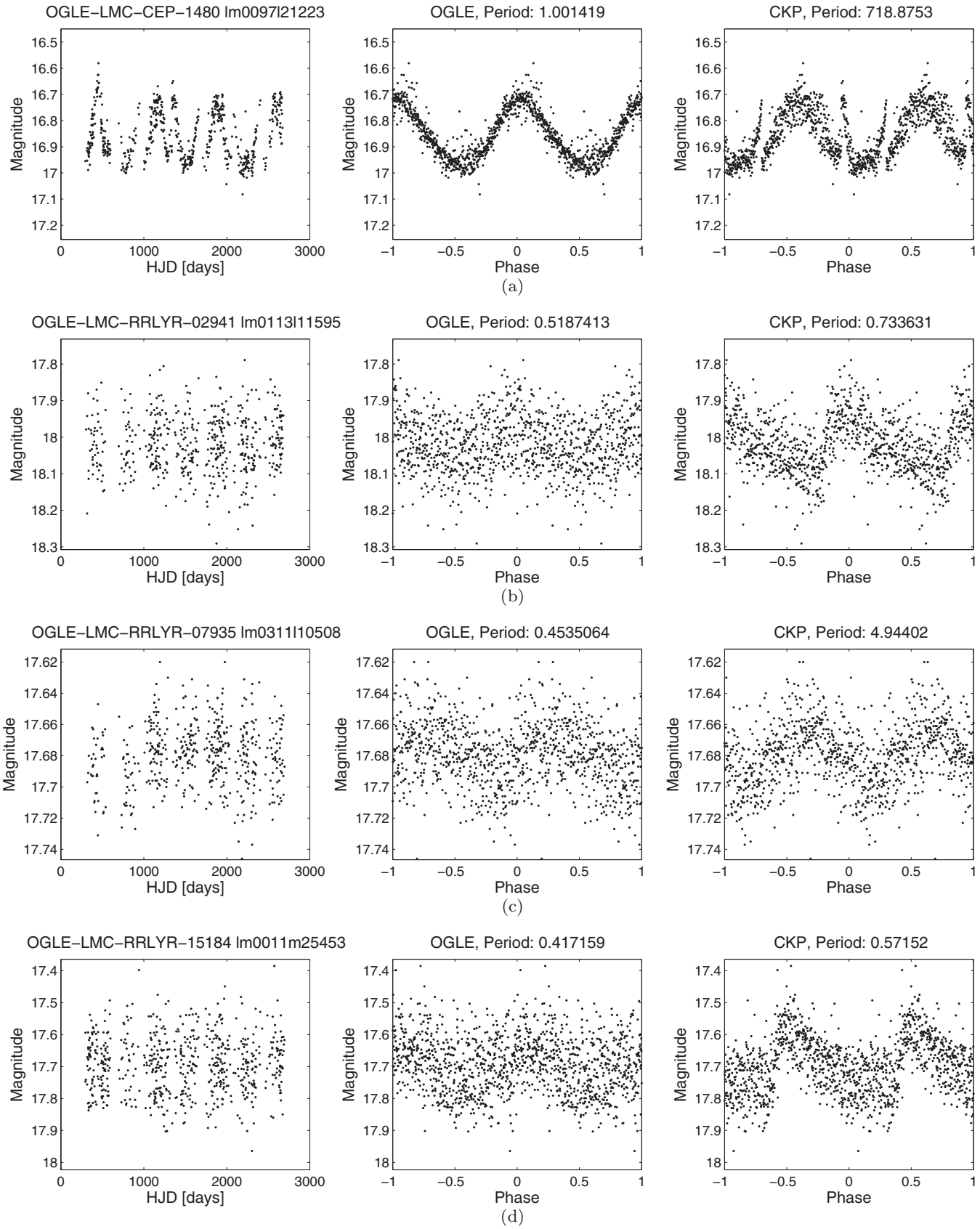


**Figure 23.** Light curves in which the reported period is in disagreement with the OGLE period. The EROS and OGLE labels, along the periods, are shown in the title of each light curve.

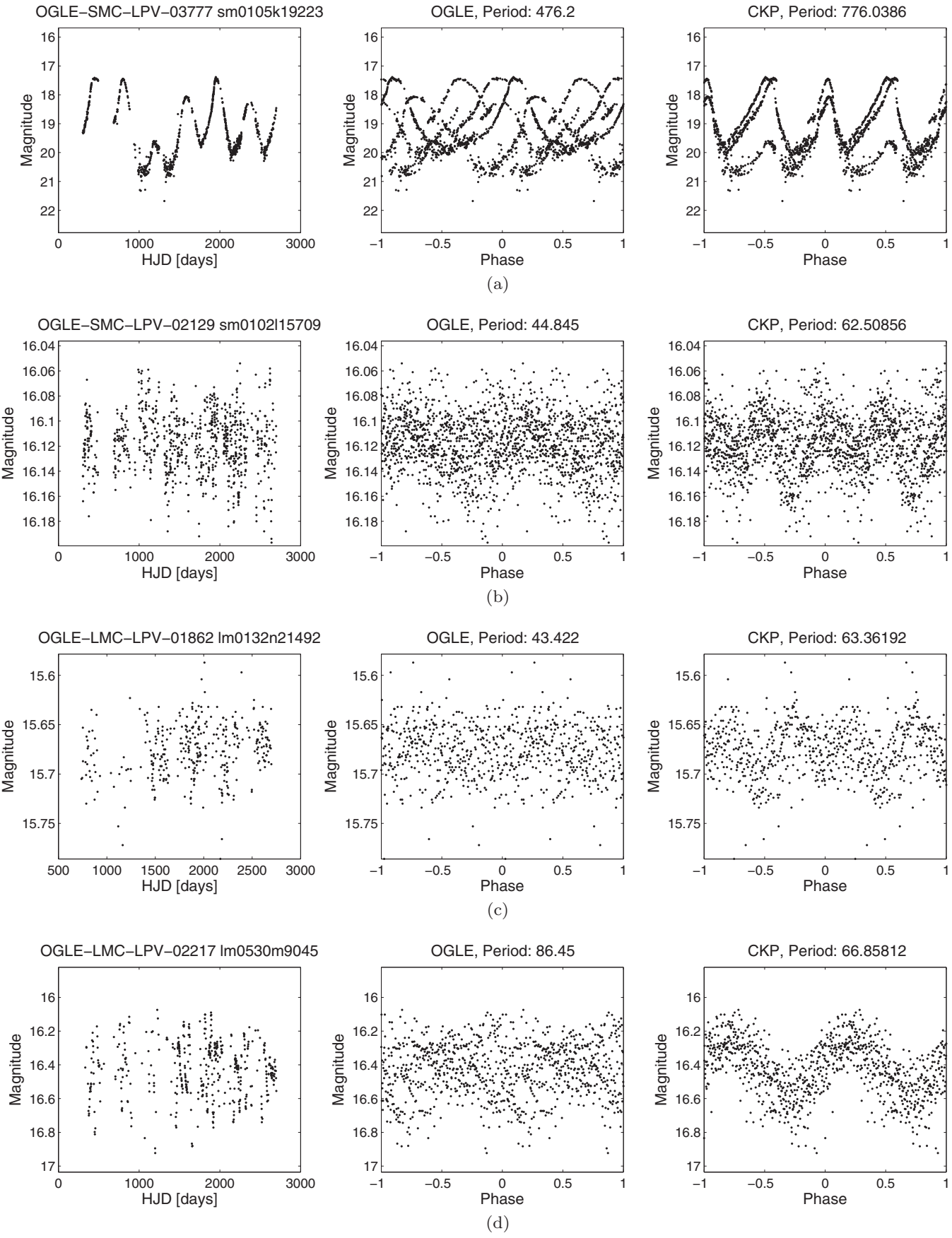


**Figure 24.** Examples of LPVs in which the reported period is in disagreement with the OGLE period. The EROS and OGLE labels, along the periods, are shown in the title of each light curve.

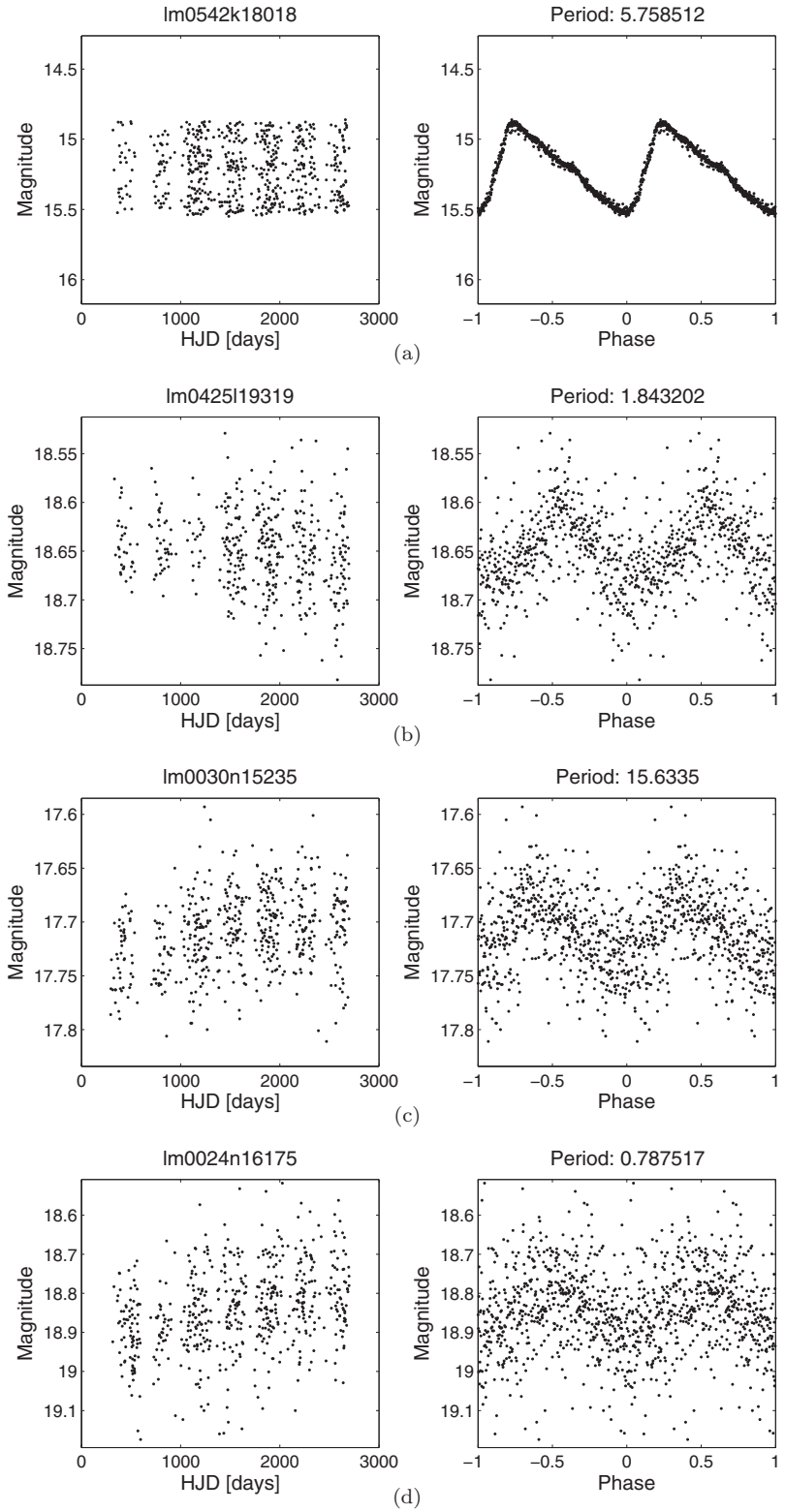


**Figure 25.** Examples of periodic light curves not found by OGLE. Panel (a) corresponds to a Cepheid variable with high S/N not found by OGLE. The majority of these light curves have a low CKP value, which translates roughly to low S/N. Panels (b)–(d) are low-S/N examples.

The catalogs are compared with existing periodic variable star catalogs for the LMC and SMC. We first test against the published OGLE catalogs for Cepheids (Soszynski et al. 2008a; Soszyński et al. 2010a), type II Cepheids (Soszyński et al. 2008b; Soszyński et al. 2010c), RR Lyrae (Soszyński et al. 2009b; Soszyński et al. 2010b), and LPV (Soszyński et al. 2009a; Soszyński et al. 2011) in the LMC and SMC. The OGLE team performed an extent period search using Fourier-based methods, analysis of variance, and visual inspection. In this test the objective is to reveal how many of the periodic variables

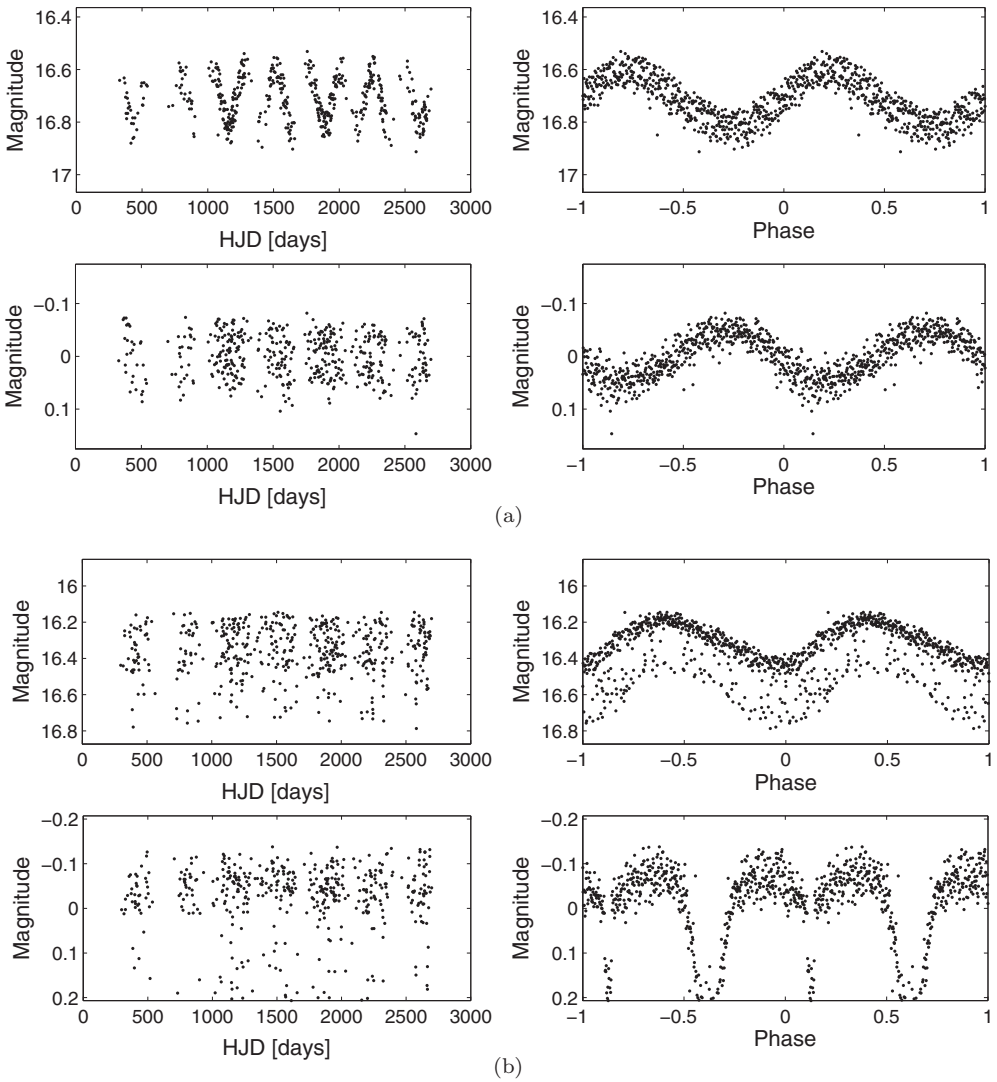


**Figure 26.** Light curves (a) lm0356k24082 and (b) lm0100m7313 are selected as dual-mode candidates. On each plot, the first and second rows correspond to the original and whitened light curve, respectively. In (a) the original light curve is folded with $P_0 = 244.06$ days. The whitened light curve is folded with $P_1 = 3.6399$ days. In (b) the original light curve is folded with $P_0 = 6.3419$ days. The whitened light curve is folded with $P_1 = 84.19$ days.

**Table 6**
Cross-matching with OGLE Periodic Variable Catalogs in the LMC and SMC

| OGLE Catalog | $N_{\rm catalog}$ | $N_{\rm inEROS}$ | $N_{\rm match}$ | Agree (%) | Multiple (%) | Disagree [%] |
|---|---|---|---|---|---|---|
| OGLE-LMC-CEPH | 3,375 | 2,727 | 2,711 | 98.8 | 1.0 | 0.2 |
| OGLE-LMC-t2CEPH | 203 | 161 | 148 | 94.6 | 4.1 | 1.3 |
| OGLE-LMC-RRLyr | 24,906 | 18,092 | 17,272 | 92.0 | 6.8 | 1.2 |
| OGLE-LMC-LPV | 91,995 | 74,960 | 20,430 | 77.2 | 2.0 | 20.8 |
| OGLE-SMC-CEPH | 4,630 | 3,413 | 3,395 | 99.3 | 0.6 | 0.1 |
| OGLE-SMC-t2CEPH | 43 | 30 | 30 | 93.4 | 3.3 | 3.3 |
| OGLE-SMC-RRLyr | 2,475 | 1,392 | 1,360 | 97.7 | 1.7 | 0.6 |
| OGLE-SMC-LPV | 19,384 | 14,103 | 4,413 | 70.3 | 2.6 | 27.1 |

reported by the OGLE team can be found in our catalogs and to analyze the discrepancies between the detected periods. Table 6 summarizes the results of the cross-matching. First, for each OGLE object, a nearest neighbor in the EROS catalog is found. Neighbors with a separation larger than 1.″5 are not considered. Column $N_{\rm inEROS}$ corresponds to the number of OGLE objects that were found in the EROS set within the search distance. The OGLE objects that did not have an EROS neighbor were either out of EROS bounds, located on interchip EROS zones, or located on corrupted EROS chips. Column $N_{\rm match}$ corresponds to the number of cross-matched OGLE–EROS objects that appear in our periodic variable catalog. The differences between $N_{\rm inEROS}$ and $N_{\rm match}$ are due to OGLE objects whose CKP is below the periodicity threshold (low-S/N light curves). There are cases in which the true period is within the spurious filter areas and was missed in our search. Finally, the periods

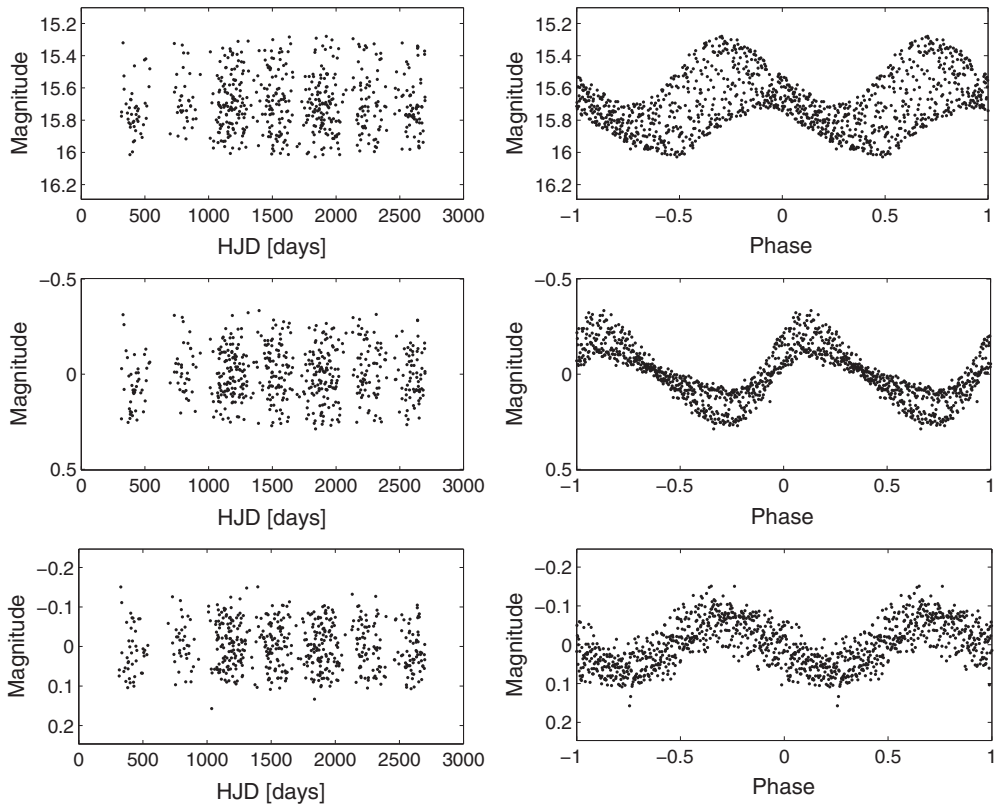


**Figure 27.** Light curve lm056518888 is selected as a triple-mode candidate. In the plot the first, second, and third rows correspond to the original, first whitened, and second whitened light curves, respectively. The original light curve is folded with the detected period $P_0 = 2.4725$ days. The first whitened light curve is folded with $P_1 = 3.4455$ days. The second whitened light curve is folded with $P_2 = 1.4395$ days.

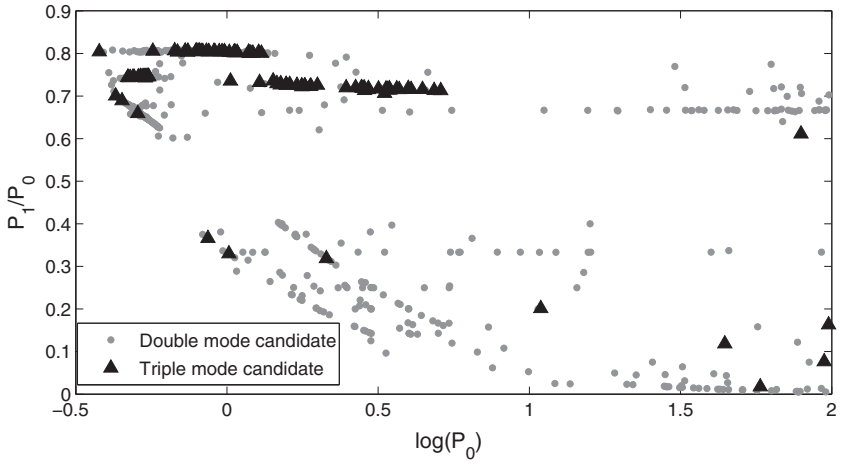


**Figure 28.** Petersen diagram of the 1165 dual-mode candidates found in the LMC. The triangles mark the location of 116 triple-mode candidates. Clear structures arise in the diagram.

reported by OGLE are compared with the periods found with our method. The agreement column corresponds to the percentage of light curves in which the OGLE period is equal to the period found in our catalog (a 1% relative error is considered). The Multiplecolumn corresponds to the cases in which the reported period is either a multiple, submultiple, or alias of the OGLE period. The Disagreementcolumn corresponds to the cases in which the reported period is not related to the OGLE period.

There is a high level of agreement between the reported and OGLE periods for Cepheids, type II Cepheids, and RR Lyrae classes, in both the LMC and SMC. The periods labeled as multiples were visually inspected. In these cases the OGLE period is the correct period, but it was not found by the proposed method because it was either below 0.3 days or filtered in the spurious period rejection stage. Examples of the light curves in which the reported period is in disagreement with the OGLE period are shown in Figure 23.

For the LPV class the difference between $N_{\text{inEROS}}$ and $N_{\text{match}}$ is larger than in other classes (i.e., more objects with CKP below periodicity threshold). This is expected as the LPVs are known to suffer from irregularities that affect their period. Additionally,

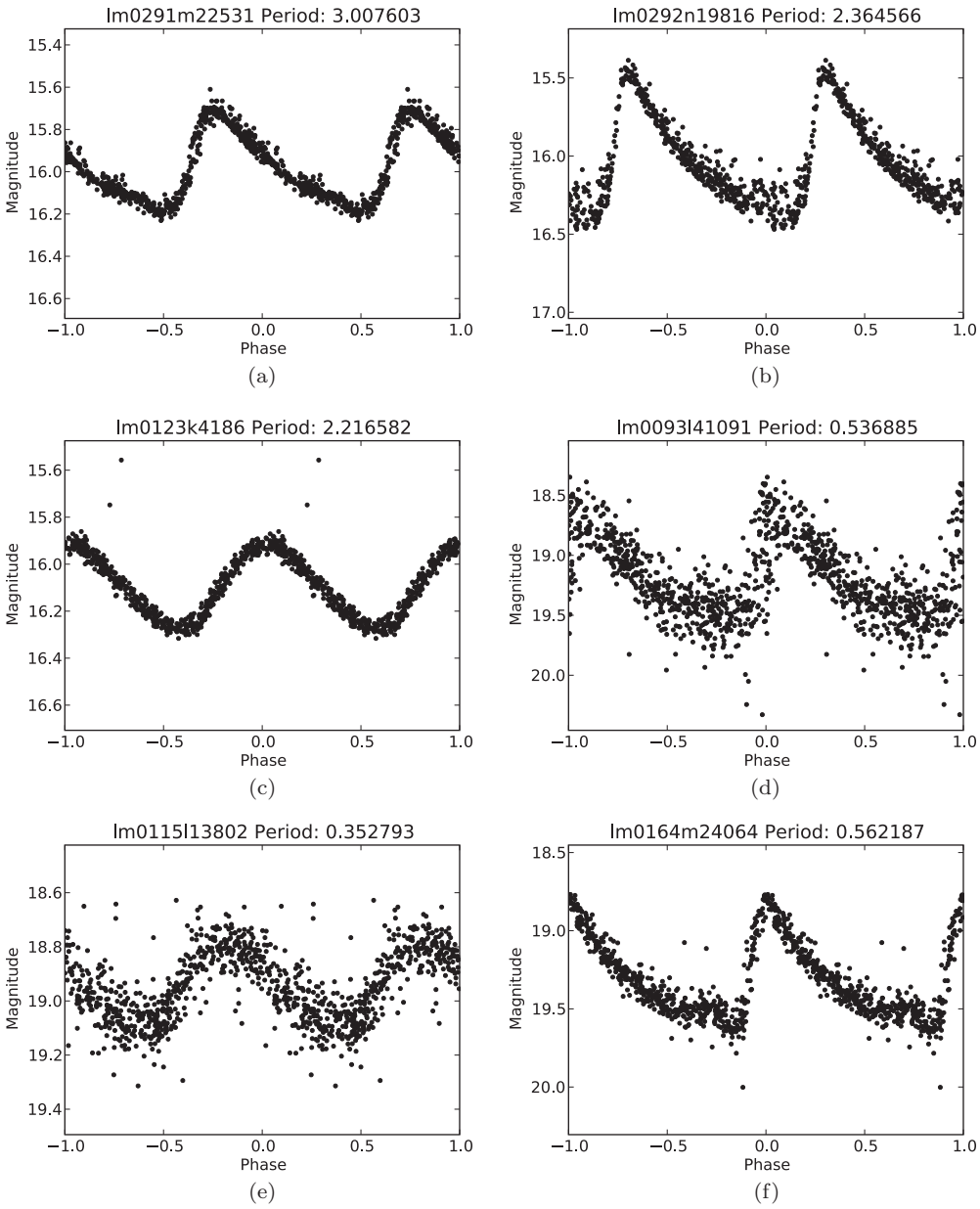


**Figure 29.** Examples of EROS-2 periodic light curves folded with their estimated period. (a)–(c) are Cepheids taken from the CEPH cluster (see Figure 21(a)). (d)–(f) are RR Lyrae taken from the RRL cluster. (d) and (f) are examples of RRab class stars. (e) is an example of an RRc class star.

the level of agreement between periods is lower than the other classes. Figure 24 shows examples of disagreeing periods in the LPV class.

There are 80,304 objects in our periodic catalog that do not have a neighbor from the OGLE periodic variable catalogs (within 2.″5). Some of these objects may not have been surveyed by the OGLE project, or they could belong to classes with currently unavailable catalogs such as eclipsing binaries. A total of 60% of these light curves have a low CKP value, which translates roughly to low S/N. This could indicate that the proposed method is more sensitive than the method used by the OGLE team. Figure 25 shows examples of periodic light curves found in the EROS catalog that do not appear in the OGLE catalogs.

The periodic variable catalogs are also compared to the lists of beat Cepheids found in the EROS-2 data by Marquette et al. (2009). The catalog contains Cepheids pulsating on their fundamental and first overtone (F/FO) and first and second overtone (FO/SO), respectively. The periods were obtained using a combination of Fourier decomposition, analysis of variance, and visual inspection. The results are summarized in Table 7. There are eight cases that do not appear in our catalog owing to their CKP value being below the threshold. In the remaining 409 cases, only three cases show disagreement with the reported period. The one case in which

**Table 7**
Cross-matching with EROS-2 Beat Cepheid Catalogs for the LMC and SMC

| Beat Cepheids Catalog | $N_{\rm catalog}$ | $N_{\rm match}$ | Agree (%) | Multiple (%) | Disagree (%) |
|---|---|---|---|---|---|
| F/FO pulsation | 115 | 109 | 100.0 | 0.0 | 0.0 |
| FO/SO pulsation | 302 | 300 | 99.0 | 0.66 | 0.33 |

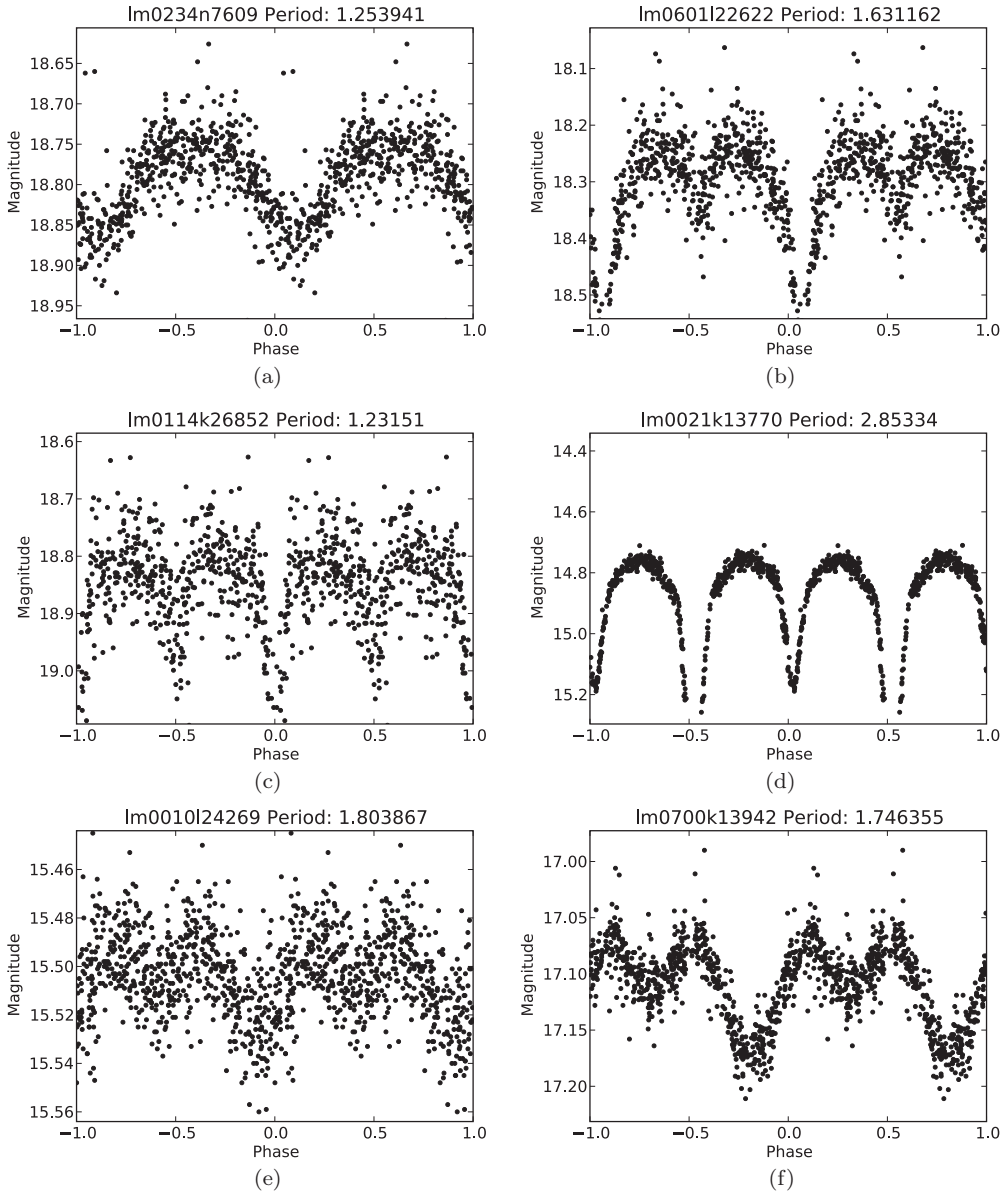


**Figure 30.** Examples of EROS-2 periodic light curves folded with their estimated period. These light curves correspond to eclipsing binary stars found in the blue main sequence (see Figure 21(a)).

the period is not a multiple of the EROS-2 period was shown in Figure 23(a).

## 7. BEYOND CKP

### *7.1. Multimodes*

It is known that periodic stars exhibit multimode oscillations that are manifested in the morphology of the light curves. Despite the fact that the methodology presented in this paper was not designed to find multimodes, we have explored the multimodes in a two-level search approach. For each periodic light curve the prime light curve $P_0$ is used to remove the periodic signal. This procedure is known as whitening and is performed as follows.

1. Fold the light curve with $P_0$.
2. Obtain a template of the periodicity by smoothing the folded light curve using a moving average of 30 samples.
3. Subtract the template from the folded light curve.
4. Rearrange the light curve samples to their original time order.

If the whitened light curve is found to be periodic with period $P_1$, that is, not multiple/submultiple or alias of $P_0$, then the light curve is selected as a dual-mode candidate. Subsequent oscillation modes can be found by repeating the procedure above.

This procedure is applied on 34,000 periodic light curves from the LMC with CKP values above 2.0.[19] From this set 1165 light curves are selected as dual-mode candidates. After evaluating the double-mode candidates, 116 are found to have a third oscillation mode. Examples of dual-mode and triple-mode candidates are shown in Figures 26 and 27, respectively. The lists of double- and triple-mode candidates can be found at http://timemachine.iic.harvard.edu.

[19] We only selected the most prominent periodic light curves.

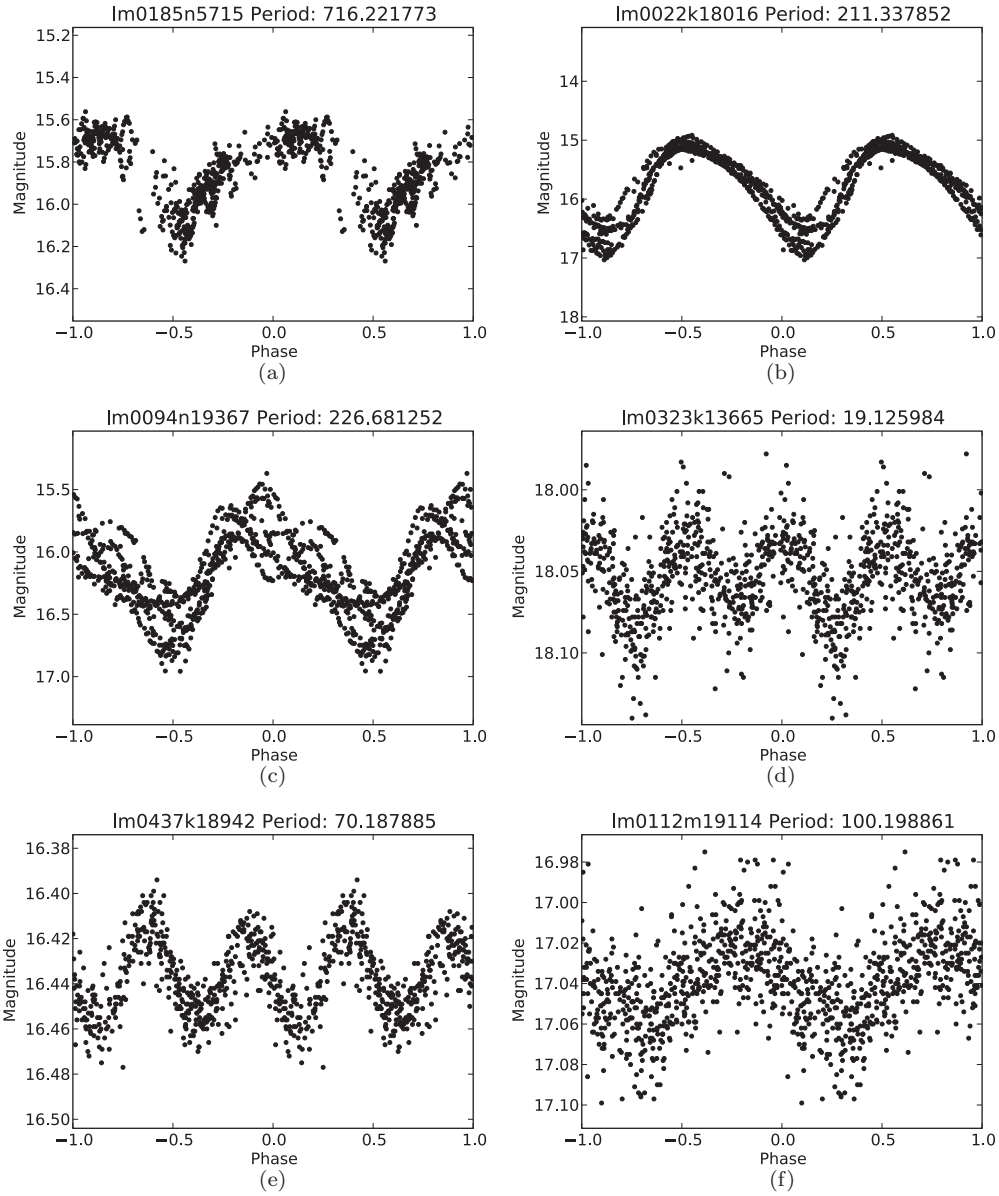


**Figure 31.** Examples of EROS-2 periodic light curves folded with their estimated period. (a)–(c) correspond to long-period variables found in the LPV cluster (see Figure 21(a)). (d)–(f) correspond to periodic variable stars found in the lower red giant branch.

Figure 28 shows a Petersen diagram of the 1165 light curves selected as dual-mode candidates. The triangles in the plot mark the 116 light curves in which a third mode was found. The periods are sorted so that $P_0 > P_1$ in all cases. The triple-mode candidates occupy two horizontal lines at period ratios of 0.72 and 0.8. These values are close to the known ratios associated with the first and second overtones (Moskalik 2013). A prominent horizontal line appears at $P_1/P_0 \sim 2/3$ for fundamental periods above 10 days. According to Smolec et al. (2012), this ratio is associated with the period doubling phenomenon. Another interesting feature, shown in the lower left part of the diagram, are two curves that follow an inversely proportional relationship between the period ratio and fundamental period.

### 7.2. Odd Periodic Stars

The method presented here is a not a classification method, and therefore the method does not distinguish between types of periodic variables. Most of the periodic objects found in this work can be classified to known classes, as is clearly shown in Figures 29–32. It is also expected that there should or could be stars with periodic behavior that does not fall in one of the known categories. It is the scope of a different paper to identify those rare or novel phenomena. Right here we only present a number of objects that we could not obviously attribute to any known classes or combination of classes. Figure 33 shows two such cases.

## 8. COMPUTATIONAL ISSUES

The proposed periodicity discrimination pipeline has been programmed for computational architectures based on graphical processing units (GPUs). The implementation is programmed in CUDA NVIDIA (2012), which is a variation of C developed by GPU manufacturer NVIDIA.

To evaluate the CKP metric (Equation (10)), one requires the $N(N-1)/2$ interactions between the $N$ samples of the time

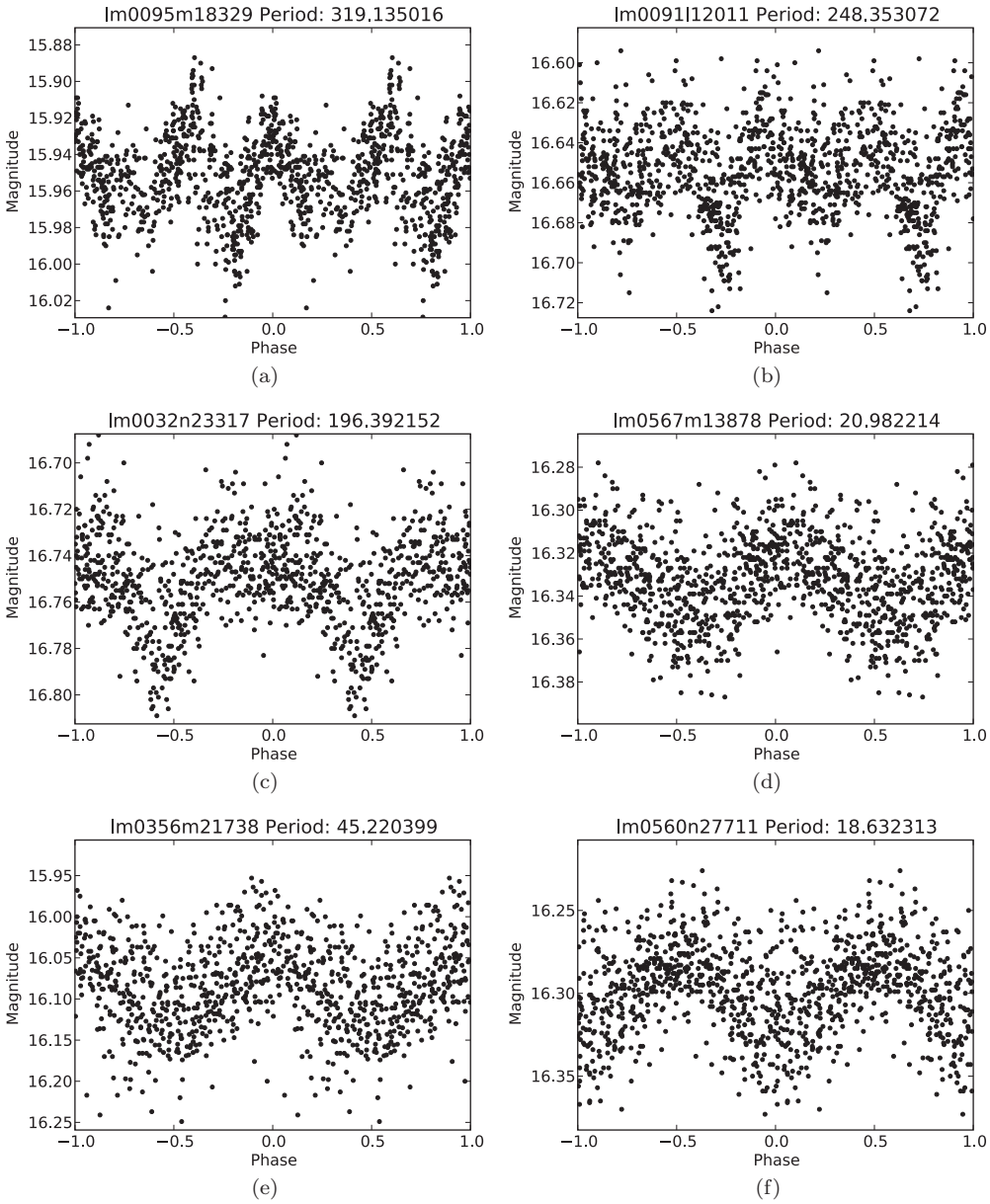


**Figure 32.** Examples of EROS-2 periodic light curves folded with their estimated period. (a)–(c) correspond to long-period variables found in the LPV-2 cluster (see Figure 21(a). (d)–(f) correspond to ellipsoidal variables found in the ELL cluster.

series.[20] The CKP can be computed efficiently by mapping each of these interactions to a single GPU thread. The final value of the CKP is obtained through a $\log(N)$-step sum reduction performed on the GPU. The computational time required to analyze one light curve using our periodic discrimination pipeline is shown in Figure 34. These times include the importation and transferring of the light curves to the GPU device. Times were measured on an NVIDIA Tesla C2070 GPU.

The 32.8 million light curves from the EROS-2 survey are processed on the NSCA Dell/NVIDIA cluster Forge. Forge is part of the Extreme Science and Engineering Discovery Environment (XSEDE). Forge has a total of 288 NVIDIA Tesla C2070 accelerators distributed on 44 nodes; however, the maximum number of nodes that can be used at a time is 26. Each GPU processes one chip from EROS-2. Table 8 shows the total computational time required to process the 32.8 million light curves from the LMC and SMC. These times does not include the time required to transfer the data set to the cluster or the time a job is waiting on the queue.

**Table 8**
Total Computational Time Required to Process the 32.8 Million EROS-2 Light Curves (LMC plus SMC) on XSEDE Forge Cluster

| Hardware | Computational Time |
|---|---|
| Using 1 GPU | 52.2 days |
| Using 6 GPUs (1 node) | 8.71 days |
| Using 12 nodes (6 GPUs/node) | 17.41 hr |
| Using all available nodes | 7.28 hr |

**Note.** GPUs in all nodes are NVIDIA Tesla C2070.

[20] The kernel matrices given by Equations (4) and (8) are symmetric; thus, only the upper triangular part needs to be computed. The diagonal of the kernel matrices is constant and is omitted from the computations.

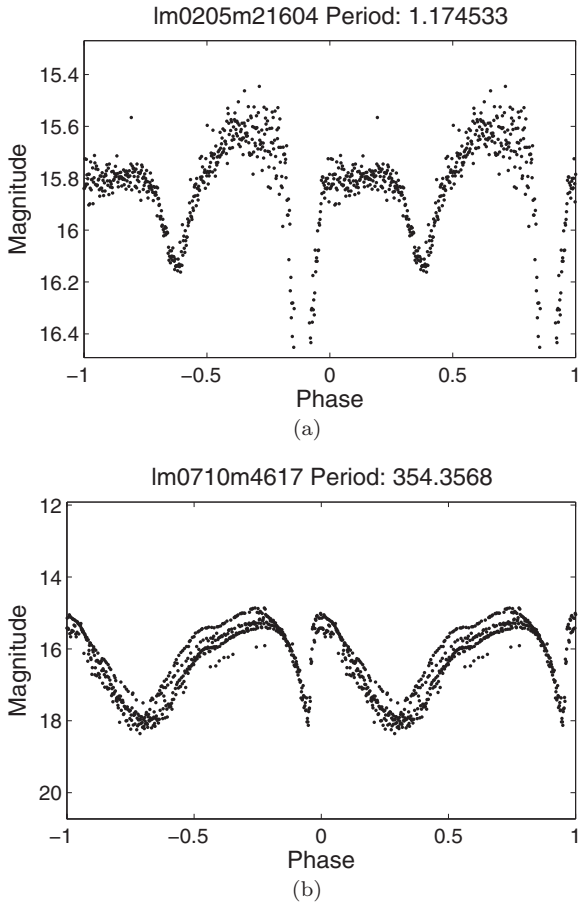


**Figure 33.** Examples of EROS-2 periodic light curves folded with their estimated period. A priori these objects cannot be attributed to any known class.

## 9. CONCLUSIONS

We presented and described a fully automated pipeline for periodic light curve discrimination. The method is based on the CKP, a robust information theoretic metric that discriminates periodic behavior by analyzing the similarities between light-curve samples. The method is computational efficient; the pipeline takes 0.16 s to discriminate whether a light curve is periodic or not. The 32.8 million light curves were processed using a GPU cluster in less than 24 hr. This suggests that with a few additional optimizations and up-to-date hardware the methods may scale well for modern and larger light curve databases.

The periodicity discrimination pipeline was tested on light curves from the EROS-2 survey. The methods were calibrated using synthetic time series that preserve the characteristics of EROS-2 light curves. The calibration procedure is general and could be applied to other astronomical time series databases easily. In total, 32.8 million light curves from the LMC and SMC were processed, finding a grand total of 121,147 and 24,855 periodic variables in the LMC and SMC, respectively. The results obtained are consistent in terms of period distribution and localization of the periodic variables in the color–magnitude diagram. The observed results suggest that the periodic variable catalogues generated by our method could be used to find multimode variables and periodic variables that do not fall in any known category. It is also hinted that higher-order analysis, such as stellar classification and clustering, may be carried out straightforwardly using the provided periods.

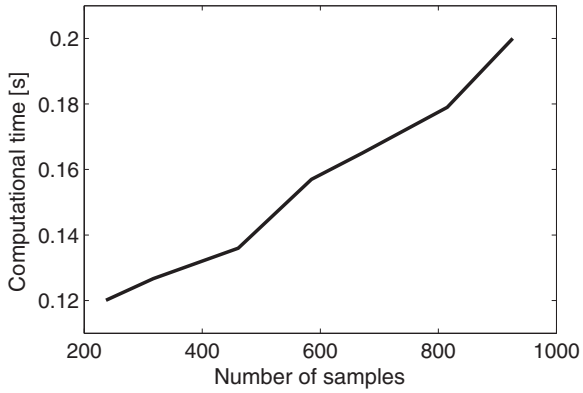


**Figure 34.** Computational time per light curve as a function of the number of samples.

Using the synthetic data set and visually inspecting a small subset of the data set, we were able to characterize the completeness and efficiency of the pipeline. We infer that 0.5% of the light curves with $S/N > 0.5$ are periodic.

Future work involves quasi-periodic and semiregular behavior discrimination, more in-depth analysis of nonstationarities (trends), and developing more general kernel size selection schemes.

This work was funded by CONICYT-CHILE under grant FONDECYT 1110701 and 1140816 and its Doctorate Scholarship program. P.E. acknowledges support from the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC12009, awarded to the Millennium Institute of Astrophysics, MAS.

The authors thank the Harvard Institute for Applied Computational Science for providing research space and computing facilities.

The help received from the SEAS academic computing support staff and the time on the Harvard SEAS "Resonance" GPU cluster are greatly acknowledged.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

The EROS-2 project was funded by the CEA and the CNRS through the IN2P3 and INSU institutes. J.-B.M. acknowledges financial support from "Programme National de Physique Stellaire" (PNPS) of CNRS/INSU, France.

## REFERENCES

Alcock, C., Allsman, R. A., Alves, D. R., et al. 2000, ApJ, 542, 281
Buhlmann, P. 1999, StaSc, 17, 52
Cook, K. H., Alcock, C., Allsman, H. A., et al. 1995, in ASP Conf. Ser. 83, IAU Colloq. 155: Astrophysical Applications of Stellar Pulsation, ed. R. S. Stobie & P. A. Whitelock (San Francisco, CA: ASP), 221
Edelson, R. A., & Krolik, J. 1988, ApJ, 333, 646
Eyer, L. 1999, BaltA, 8, 321
Hodapp, K. W., Kaiser, N., Aussel, H., et al. 2004, AN, 325, 636
Huijse, P., Estevez, P. A., Protopapas, P., Zegers, P., & Principe, J. C. 2012, ITSP, 60, 5135
Ivezic, Z., Tyson, J. A., Acosta, E., et al. 2011, arXiv:0805.2366
Jenkins, G. M., & Watts, D. G. 1968, Spectral Analysis and Its Applications (San Francisco, CA: Holden-day)
Larson, S., Beshore, E., Hill, R., et al. 2003, BAAS, 35, 982
Law, N. M., Kulkarni, S. R., Dekany, R. G., et al. 2009, PASP, 121, 1395
Mackay, D. 1998, Introduction to Gaussian Processes, Vol. 168 (Berlin: Springer), 133
Marquardt, D., & Acuff, S. 1984, Direct Quadratic Spectrum Estimation with Irregularly Spaced Data (New York: Springer), 211

Marquette, J. B., Beaulieu, J. P., Buchler, J. R., et al. 2009, A&A, 495, 249
Michalak, M. 2010, in Computer Recognition Systems 4, ed. R. Burduk, M. Kurzyński, M. Woźniak, & A. Zołnierek (Berlin: Springer), 136
Moskalik, P. 2013, in Advances in Solid State Physics, Vol. 31, ed. J. C. Suárez, R. Garrido, L. A. Balona, & J. Christensen-Dalsgaard (Berlin: Springer), 103
NVIDIA. 2012, CUDA C Programming Guide version 4.2 (NVIDIA)
Principe, J. 2010, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives (New York: Springer)
Rahal, Y. R., Afonso, C., Albert, J.-N., et al. 2009, A&A, 500, 1027
Rasmussen, C. E., & Williams, C. K. I. 2006, Gaussian Processes for Machine Learning (Cambridge, MA: MIT Press)
Reimann, J. D. 1994, Frequency Estimation Using Unequally-Spaced Astronomical Data (Berkeley, CA: University of California)
Scargle, J. 1982, ApJ, 263, 835
Schmitz, A., & Schreiber, T. 1999, PhRvE, 59, 4044
Schölkopf, B., & Smola, A. 2002, Learning with Kernels (Cambridge, MA: MIT Press)
Schreiber, T., & Schmitz, A. 1999, PhyD, 142, 346
Smolec, R., Soszyński, I., Moskalik, P., et al. 2012, MNRAS, 419, 2407
Soszynski, I., Poleski, R., Udalski, A., et al. 2008a, AcA, 58, 163
Soszyński, I., Poleski, R., Udalski, A., et al. 2010a, AcA, 60, 17
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2008b, AcA, 58, 293
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2009a, AcA, 59, 239
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2009b, AcA, 59, 1
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2010b, AcA, 60, 165
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2010c, AcA, 60, 91
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2011, AcA, 61, 217
Spano, M., Mowlavi, N., Eyer, L., & Burki, G. 2009, in AIP Conf. Ser. 1170, Stellar Pulsation: Challenges for Theory and Observation, ed. J. A. Guzik & P. A. Bradley (Melville, NY: AIP), 324
Stellingwerf, R. 1978, ApJ, 224, 953
Taylor, J. S., & Cristianini, N. 2004, Kernel Methods for Pattern Analysis (Cambridge: Cambridge Univ. Press)
Tisserand, P., Le Guillou, L., Afonso, C., et al. 2007, A&A, 496, 387
Udalski, A., Kubiak, M., & Szymanski, M. 1997, AcA, 47, 319
Wang, Y., Khardon, R., & Protopapas, P. 2012, ApJ, 756, 67
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579