

Unsupervised classification of eclipsing binary light curves through k-medoids clustering

Soumita Modak¹, Tanuka Chattopadhyay²
and
Asis Kumar Chattopadhyay¹

¹Department of Statistics, Calcutta University, Kolkata
35 Ballygunge Circular Road, Kolkata-700019, India
email: soumitamodak2013@gmail.com
email: akcstat@caluniv.ac.in

²Department of Applied Mathematics, Calcutta University, Kolkata, India
92 A.P.C. Road, Kolkata -700009
email: tanuka@iucaa.ernet.in

Abstract

An automatic unsupervised classification of 1318 light curves of variable stars, including eclipsing binaries along with some possible pulsating stars, has been performed using k-medoids clustering method. This separates the stars according to their geometrical configuration in a more scientific way compared to the subjective traditional classification scheme. The light curves in the Galaxy, subjectively grouped in four categories (EA, EB, EW, PUL) in Miller et al. (2010), have been found to consist of two optimum groups containing primarily eclipsing binaries corresponding to bright, massive systems and fainter, less massive systems. Our technique has been assessed in terms of clustering accuracy measure the Average Silhouette Width, which shows the resulting clustering pattern is quite good.

Keywords: Light curve of variable star; Clustering; k-medoids method; CID.

1 Introduction

Eclipsing binaries (Es) can be treated as fundamental probe for studying stellar structure and stellar evolution. The joint analysis of their light curves (LCs) as well as velocity curves (1 – 2%) (Bradstreet and Steelman 2002;

Chattopadhyay et al. 2016) determines their masses, radii, luminosity and temperature (Chattopadhyay et al. 2016). Study of Es in external galaxies (Akerlof et al. 2000; Street et al. 2004; Graczyk et al. 2011) has made it possible to explore stellar evolution and to establish various formation theories for galaxies with varying evolutionary and chemical history compared to our own Galaxy (e.g. LMC, SMC). Moreover, they play an important role in distance indicators to many galaxies. The distance moduli for early type nearby galaxies have shown that the accuracy is as close as ± 0.1 mag (Giménez et al. 1994), a precision comparable to that obtained for individual Cepheid variables. Hence an improved and concise database of Es may lead to an improvement in the extragalactic distance scale estimations.

In recent studies, several photometric surveys (Wyithe and Wilson 2002; López-Morales and Christopher 2004; Hełminiak et al. 2012) have been carried out to produce a wealth of LC of variable stars out of which a large amount of E systems can be found. For example, ESA astrometric satellite, Hipparcos, found 70% new variables. The Global Astrometric Interferometer for Astrophysics, a large scale photometric survey, also collected information on several variable stars of which many were Es. It is speculated that about 1 million Es, those with $V \leq 16$ mag, will be discovered (Niarchos 2006). Even if only 1% of the observed Es have derived physical parameters, that will have great contribution to stellar astrophysics compared to what has been obtained so far from ground-based observations. During the past decades many photometric surveys (Wyithe and Wilson 2002; López-Morales and Christopher 2004; Hełminiak et al. 2012) have aimed at the detection of extrasolar planets via transits or microlensing. A byproduct of these programmes is the detection of a large number of variable stars (Albrow et al. 2001; Bayne et al. 2002; Wozniak et al. 2002; Wyrzykowski et al. 2004; Weldrake et al. 2004, 2007; Soszyński et al. 2008a, 2008b).

The above observations are generally related to Galactic Bulge, Magellanic clouds or globular cluster surveys. Miller et al. (2010) have carried out observations of Es (along with some possible pulsating stars) covering 0.25 square degree region of the Galactic plane centered on Galactic coordinates (l, b) of $(330.94, -2.28)$ deg. The majority of stars in the above region are thought to be associated with the Normal Spiral Arm. A large catalogue of Es (viz. 7179 LCs) has been compiled by Malkov and Avvakumova (2013) along with update on some previous catalogues (Malkov et al. 2006; Avvakumova et al. 2013). However, in the present paper we decided to focus on the LCs of Es of Miller et al. (2010), as we are interested in the classification of

binary stars in our Galaxy associated with the spiral arms.

LCs of Es are time series data and can be clustered to find out the possible sources of homogeneous groups in the data set, e.g. Kochoska et al. 2017; Mowlavi et al. 2017; Süveges et al. 2017. In those model based classification schemes (e.g. polynomial or two-Gaussian model fitting, etc.), the error is around 10% which is mainly due to the similarity of LCs originating from different physical systems. In particular, Kochoska et al. 2017 have found four groups of which the first two and the last two have similar Kepler polyfit primary depths indicating merely two significant groups. This is somewhat similar to the present case, but unlike the previous methods, the present one is a nonparametric partitioning approach. The k-medoids algorithm (Kaufman and Rousseeuw 1990) is one of the most well-known partitioning based clustering method, which is very classical in statistical analysis. This method is robust against noise, outliers, extreme values and sparsely distributed data (Singh et al. 2011), which often arise in real data sets. This method can be used to partition the data where no prior classification information is available. It uses some distance measure computed between the data sets and allows to adopt any distance measure according to the nature of the given data. It is very important to notice that distance measures, which have been evidenced to analyze cross-sectional data well, may not be appropriate for time series data, as not all distance measures can extract the temporal information in the data. So a proper distance measure for computing distance between time series at hand is to be carefully chosen. Depending upon the nature of time series, we have selected complexity invariance distance (CID) measure (Prati and Batista 2012; Batista et al. 2014; Wei 2014) and used it in combination with k-medoids clustering method to explore the clusters in our data. As competitor we consider another popular time series distance, Dynamic Time Warping distance (hereafter denoted by DTW) (Rabiner and Juang 1993; Keogh and Ratanamahatana 2005; Giorgino 2009; Cassisi et al. 2012), which is computed by Dynamic Time Warping to find an optimal alignment between two given times series under certain restrictions, where the optimal alignment is reached by minimizing the sum of the distances between the aligned elements. We have applied k-medoids clustering method with CID, Euclidean distance (ED) and DTW to the time-dependent observations over phase, i.e. the folded LCs which are mono-periodic time series. We have compared the clustering results in terms of the Average Silhouette Width (ASW) (Rousseeuw et al. 1987), which stands for a measure of how appropriately the data has been clustered, and shown that CID has outper-

formed the other two, resulting in quite good clustering pattern consisting primarily of two groups of Es.

The paper is organized as follows. In Section 2, we have described the data and the transformations used on the data. Section 3 describes the clustering method with the distance measures and the accuracy measure used in our work. Result and discussion are given in Section 4. Section 5 concludes.

2 Data

The present data set is taken from Miller et al. (2010) where for each of the 1318 variable stars in our Galaxy, there is a LC file together with R-band magnitude, colours (B-R, R-I) and period (P). The variable stars mainly consist of Es along with some possible pulsating stars (201 uncertain and 118 confirmed pulsating stars according to the subjective classification). For each LC, relative flux variation in R-band is given on a continuous time scale in Heliocentric Julian Date (HJD) within the range from 2452450.62250 HJD to 2453607.61580 HJD. Each LC is unequally spaced of different length (here, length of a LC refers to the number of observations on that LC) having values at different time points. For these variable stars, the length of the LCs varies from 130 to 264 (except one LC having length 5) and period ranges from several hours to several weeks.

2.1 Phase computation, Interpolation and Binning

As there are different lengths of the LCs having values at different time points, comparison of the LCs are only possible in terms of observations over each cycle. Hence we transform the given LCs of the stars, with known and constant period, into phase series (see, Appendix A1).

Now, for clustering these time series, using the distance measure under consideration (discussed in Section 3.2), we need to get a full cycle over phase 0 to phase 1 for each LC having values at the same and equidistant phase points. Hence we use Linear Spline or Piecewise Linear Interpolation (Cassisi et al. 2012), and obtain observations at 272 equidistant phase points over the phase interval $[0,1]$ for each LC (for details, see, Appendix A1).

3 Classification scheme

A classification scheme aims to separate eclipsing binary systems according to the geometrical configuration. The subjective classification separates the LCs into four groups, viz. Algol type (EA), Beta Lyrae (EB), W Ursae Majoris (EW) and un-categorized Pulsating stars (PUL). But subjectivity sometimes includes degeneracy, i.e. it includes systems with different physical properties in the same group. This traditional scheme is based on the appearance of the LCs, which is not only almost obsolete but also rather misleading. Most importantly, many of the stars were categorized with uncertainty or ambiguity. The present classification tries to minimize the above limitations so that it can significantly solve the class heterogeneity and the subjectivity of the traditional LC classification. Our classification relates the groups established to the geometry, in the sense that the systems with the same geometrical configuration are classified in the same group. We found that a simple two-group clustering is sufficient to separate the systems into homogeneous classes.

Clustering is one of the most important research areas in the field of data mining. In simple words, clustering is a division of data into different groups, called clusters. Data are grouped into clusters in such a way that within a group observations are similar and between the groups observations are dissimilar. Because it aims at minimizing within-group distance while maximizing between-group distance. It is useful to obtain interesting patterns and structures from large data sets.

Clustering of time series (Liao 2005), either in time domain or frequency domain (Dargahi-Noubary 1992; Caiado et al. 2006), can be performed on time series of equal length or unequal length (Lomb et al. 1976; Caiado et al. 2009; Stefan et al. 2013), evenly spaced or unevenly spaced (Scargle 1989; Moller-Levet et al. 2003; Eckner 2014, 2017).

There are different clustering methods for cross-sectional and time series data in the literature (Chattopadhyay et al. 2007a, 2007b, 2009, 2010, 2012, 2013; Fraix-Burnet et al. 2010, 2012; Velmurugan and Santhanam 2011) such as partitioning based methods, hierarchical methods, fuzzy clustering (Moller-Levet et al. 2003), grid-based methods, density-based methods, model-based methods (Kalpakis et al. 2001; Fröhwrth-Schnatter et al. 2008), methods for higher dimensional data, constraint-based clustering, soft computing technologies including neural network clustering (e.g. Competitive learning, self-organizing feature maps) and genetic clustering (e.g.

Genetic k-means, Genetic k-medoids algorithms) (Liao et al. 2006).

3.1 k-Medoids: Partitioning Around Medoids (PAM)

k-medoids is an unsupervised learning method among the partitioning based clustering methods which can be applied to time series data. We use a fast and efficient algorithm for k-medoids clustering method (Kaufman and Rousseeuw 1990). It is based on the search for k medoids among the observations of the data set. A medoid is the representative object of the cluster it belongs to. These k medoids represent the various structural aspects of the data set being investigated. This method partitions the whole data set of size, say, N into k mutually exclusive and exhaustive clusters around k medoids, where a medoid is that object of the cluster for which the sum of distances to all the other objects of the cluster is minimal.

In the presence of noise or outliers, this method is more robust compared to k-means clustering using ED. Because it minimizes a sum of distances (any arbitrary distance which may not be ED) instead of a sum of squared EDs and it uses medoids, instead of means, which are less influenced by outliers or other extreme values than means (Singh et al. 2011). k-medoids method can be applied using any arbitrary distance measure depending upon the nature of the given data, and this algorithm enables us to input user-defined distance measure. The detailed algorithm is given in Appendix A2, where the number of clusters (k) is chosen by ASW (discussed in Section 3.3).

3.2 Distance measure

We consider the complexity invariance for time series and use CID (Prati and Batista 2012; Batista et al. 2014; Wei 2014) to measure distances between the LCs. Complex time series are those having a large number of peaks, in different quantities, amplitudes and durations, so are the LCs under our consideration. Pairs of complex time series which apparently look similar can reveal further distance when measured under complexity invariance. There exist many complexity measures for time series, like Fractal dimension, Absolute difference, Compression, Pseudo-pairs, Edges, Zero-crossings, Spectrum, Fourier entropy, Permutation entropy, etc. Batista et al. (2014) empirically showed that CID results better on an average over all other complexity measures mentioned before and CID is effective in clustering complex time series.

We adopted CID to compute the distances between time series, because the observed LCs over phase show considerable complexity in terms of the Complexity Estimate defined in (2). Moreover, CID can be computed in linear time, has no parameters, is easily interpretable, considers the relative complexity of time series, and hence improves the clustering accuracy.

For our data we have compared the clustering results obtained through CID, ED and DTW in combination with k-medoids clustering method, in terms of ASW (see, Section 3.3), and shown that CID outperformed the other two (see, Fig. 1).

Here CID between two time series X with values x_1, x_2, \dots, x_n and Y with values y_1, y_2, \dots, y_n corresponding to time points $t = 1, 2, \dots, n$, is defined as

$$\text{CID}(X, Y) = \text{ED}(X, Y) \times \text{CF}(X, Y), \quad (1)$$

where CF is a Complexity Correction factor given by

$$\text{CF}(X, Y) = \frac{\max(\text{CE}(X), \text{CE}(Y))}{\min(\text{CE}(X), \text{CE}(Y))},$$

with the following Complexity Estimate of time series X

$$\text{CE}(X) = \sqrt{\sum_{t=1}^{n-1} (x_t - x_{t+1})^2}; \quad (2)$$

and

$$\text{ED}(X, Y) = \sqrt{\sum_{t=1}^n (x_t - y_t)^2}.$$

There are various versions of DTW in the literature (Rabiner and Juang 1993; Keogh and Ratanamahatana 2005; Giorgino 2009; Cassisi et al. 2012), ours is defined as follows. Consider two time series X with values x_1, x_2, \dots, x_m and Y with values y_1, y_2, \dots, y_n corresponding to time points $t = 1, 2, \dots, m$ and $t = 1, 2, \dots, n$ respectively. Dynamic Time Warping finds the warping path $W = w_1, w_2, \dots, w_l, \dots, w_L$ of contiguous elements on the local distance matrix, whose $(i, j)^{th}$ element is $d(X_i, Y_j) = |X_i - Y_j|, i = 1, 2, \dots, m, j = 1, 2, \dots, n$, such that $w_l = (i_l, j_l) \in \{1, 2, \dots, m\} \times \{1, 2, \dots, n\}, l = 1, 2, \dots, L, \max(m, n) \leq L < m + n - 1$, satisfies the following conditions,

1) Boundary conditions: $w_1 = (1, 1), w_L = (m, n)$,

- 2) Continuity: For $w_{l+1} = (i_{l+1}, j_{l+1})$ and $w_l = (i_l, j_l)$, $i_{l+1} - i_l \leq 1$ and $j_{l+1} - j_l \leq 1$ for all $l = 1, 2, \dots, L - 1$, and
- 3) Monotonicity: For $w_{l+1} = (i_{l+1}, j_{l+1})$ and $w_l = (i_l, j_l)$, $i_{l+1} - i_l \geq 0$ and $j_{l+1} - j_l \geq 0$ for all $l = 1, 2, \dots, L - 1$.
- Now, DTW which is an optimal path between X and Y under the stated restrictions is defined as,

$$\text{DTW}(X, Y) = \min \left(\sqrt{\sum_{l=1}^L w_l} \right).$$

Dynamic programming can be used to effectively find this path by evaluating the recursive function given below (for details, see, Giorgino 2009 and references therein),

$$g[i, j] = \min \left(g[i, j - 1] + d(X_i, Y_j), g[i - 1, j - 1] + 2 d(X_i, Y_j), g[i - 1, j] + d(X_i, Y_j) \right), \quad i = 1, 2, \dots, m, j = 1, 2, \dots, n.$$

In our study, $m = n$ and ED is a particular form of DTW with $w_l = (i_l, j_l), i = j = l$.

3.3 Silhouette Width

Partitioning methods like k-medoids require that the number of clusters (i.e. k) be given by the user. Here the optimum value of k is chosen from ASW (Rousseeuw et al. 1987), shown in Table 1, which accounts for the efficacy of the cluster analysis and hence is used as an accuracy measure. For each observation i , the Silhouette Width (SW) $s(i)$ (see, Appendix A3) lies from -1 to 1. Observation with a large positive $s(i)$ (close to 1) is very well clustered, $s(i)$ around 0 means that the observation lies between two clusters, and observation with a small negative $s(i)$ (near -1) is probably placed in the wrong cluster. ASW for the data set is the average SW over all i . ASW is calculated for a range of values of $k = 2, 3, 4, \dots$, etc and the value of k is chosen for which ASW is maximum. For the present situation, ASW for various values of k is documented in Table 1, which indicates optimum value of k is 2.

The Silhouette plot (Rousseeuw et al. 1987) gives a graphical representation of SW of each of the members belonging to individual clusters. The grey shade indicates the SW of an observation, i.e. here an interpolated LC over phase $[0, 1]$ using which the clustering is performed, arranged in descending (from top to bottom) order for individual clusters (see, Fig. 2 corresponds to k-medoids clustering method through CID for $k = 2$). From the figure it is clear that all the LCs in cluster 1 have positive SWs with most LCs having significantly high SW values (i.e. very well clustered) and all LCs except a few in cluster 2 have positive SWs with most of them having quite high positive SWs (i.e. well clustered), indicate that the tightness of individual clusters and separation between the two clusters are significant. ASW for cluster 1, cluster 2 and the whole data set are computed as 0.77, 0.51, 0.68, respectively, show that the data is quite well clustered.

4 Result and discussion

We compare k-medoids clustering method through CID with other existing methods in the literature, which have been successfully used for unsupervised classification of Es. We apply very popular DBSCAN method (Kochoska et al. 2017), with distance measure CID and parameters $\epsilon = 0.5$ and $MinPts = 5$, to the interpolated phase series. But it fails to identify the clustering nature in the data set, and results in only one group of size 253 with 1065 variable stars assigned as noise points. If we consider the noise points as a separate group (Kochoska et al. 2017), then also it gives a poor discrimination with ASW equals to 0.24. Again we transform the interpolated series into two-dimensional data through t -SNE technique (Kirk et al. 2016; Kochoska et al. 2017), where the distance between the LCs is measured by CID, and apply DBSCAN ($\epsilon = 0.5, MinPts = 5$) with ED to this transformed data. This method also fails to reveal the inherent groups, showing five scattered clusters of sizes 5 to 7 and 1291 noise points. While our method, robust against noise and outliers, exposes physically interpretable clusters from the linearly interpolated series with significant accuracy.

Two clusters, denoted by k1 and k2, consisting of 838 and 480 variable stars respectively, are found through k-medoids clustering with CID, irrespective of their subjective classification (see, Table 2). Template LCs of the clusters are displayed in Fig. 3 and Fig. 4. Cluster-wise representative LCs, i.e. two sets of observed LCs, shown in Fig. 5 and Fig. 6, indicate their

similarity with the template LCs of the clusters they are classified to. The average properties of the clusters are reported in Table 3. We also check the superiority (see, Table 1) and the robustness (see, Table 4) of our method by comparing it with k-means clustering applied to linear features (Modak et al. 2017), extracted from the interpolated series in terms of the first ten principal components describing more than 80% variation in the series. It can be seen that k-means clustering also hints at two optimal groups in the variable stars (see, Table 1), namely c1 and c2, whose template LCs are shown in Fig. 7. Figs 3 and 7 indicate the similarity between c1 and k1, and c2 and k2. Also Tables 3 and 4 show that the average parameter values, except for B-I, are comparable for the two groups obtained from two different methods, whereas Table 1 suggests our method performs the clustering with significantly higher accuracy in terms of ASW (i.e. with much larger value of ASW). Hence further astrophysical analyses of the groups are carried out based on the results obtained from k-medoids clustering.

Figures 3 – 6 and Table 3 indicate that in k1 (red LCs), the average time period is larger and the variation of LC between the two minima is less compared to those in k2 (black LCs). These indicate that k1 system consists of stars which form a more or less detached or semidetached system. Also the depths of the two minima of LCs for k1 are smaller compared to those for k2. This indicates k1 systems have a less massive secondary, whereas the masses are comparable for k2 systems. The colour-magnitude diagram (Fig. 8), the colour histograms (Fig. 9 and Fig. 10) and Table 3 show that k1 systems are bluer, i.e. have temperature higher than those of k2 systems. So k1 systems belong to early and k2 systems belong to late spectral types.

In this respect it is worthwhile to mention that classification of Es has been performed by many authors, e.g. Sarro et al. (2006), Malkov et al. (2007), Prša et al. (2008), Matijevič et al. (2012), Kirk et al. (2016), etc. Sarro et al. (2006) classified 81 Es using neural networks, which fall in Bayesian ensemble and hence are model-based, and compared their various properties, e.g. mass, period, separation, etc. The groups found have a high degree of superposition with respect to the above-mentioned parameters (e.g. total mass versus orbital separation). But the present work, first of all, is a nonparametric classification scheme, and there is a well-defined distinction between the LCs of the two groups of Es, both from the aspects of their average LCs and representative LCs. Also the parameters regarding several magnitudes, colours and period are distinctly different for the two groups (e.g. the period is almost double in k1 compared to k2). In addition, the

size of our data set is much larger. Finally Sarro et al. (2006) concluded that LC classification, like we perform, is always better than the classification with respect to the parameters derived from the LCs. Malkov et al. (2007) classified a new catalogue of 6330 Es on the basis of several observable parameters depending on the LCs. Though the size of the data set is larger than the present one, but the classification is subjective and restricted by several assumptions unlike our method. Prša et al. (2008) classified the parameters of Es, derived from 10,000 synthetic LCs and 50 real LCs of Es, by artificial neural network (ANN) method. In contradistinction to them, our nonparametric method is solely based on the observed LCs rather than the parameters derived from the synthetic LCs. In our case, the number of classes of Es is found scientifically, whereas they made use of subjectively presumed five classes of Es and the properties of each class are predicted through ANN by artificially best reproducing parameters. Matijevič et al. (2012) used dimension reduction technique based on LLE algorithm, and found that the projection onto a two-dimensional space can preserve the local geometry. This is somewhat consistent with our findings as we also objectively obtain two groups of Es, but finally their groups reduce to a single parameter equivalent to “detachedness” of the binaries and their method is a dimension reduction technique rather than a classification scheme. Kirk et al. (2016) classified about 2,00,000 observed LCs of Es by LIE method, but they also discussed the properties of the classes through dimension reduction technique and assuming the number of classes of Es as prerequisite. Hence, in the present work, we use k-medoids algorithm in a multidimensional space, which is a classifier (for given the number of classes, i.e. k) rather than a dimension reduction technique, and does not adopt any model assumptions unlike the previous works and is completely nonparametric. We also show statistically that k-medoids is better compared to other classifiers. Overall, our k-medoids method has the following merits over the possible competitors.

- Real LCs are used.
- Parameters derived from the LCs are not used.
- Nonparametric approach (i.e. not model-based) is adopted.
- Number of groups of Es is unknown at the initial point (i.e. the algorithm can be used for both clustering or unsupervised classification, and classification).

- The algorithm is a classifier in a multidimensional space, in our case the relative R-band magnitude at each phase point corresponds to a dimension.

5 Conclusion

We have classified 1318 LCs of eclipsing binaries along with some possible pulsating stars in the Galaxy which lie primarily along the spiral arms. k-medoids method in combination with CID has been employed for cross-verification of the subjectively classified variable stars (Miller et al. 2010), gives rise to two clusters k1 and k2, where the resulting clustering has been assessed as quite good in terms of ASW. We observe that k1 consists of all three categories EB, EW and PUL in very close number, whereas the number of EA is comparatively small. In contrast, k2 is significantly dominated by EW over EA, EB and PUL (see, Table 2). So our objective method results in classification which is independent of the so-called subjective classification, consisting primarily of two types of eclipsing binaries. It indicates there is probably no pulsating stars in the present data set, otherwise the method must have resulted in a separate group of pulsating stars. In k1, the systems are bluer and consist of stars with unequal mass, whereas the systems in k2 are redder and consist of stars with comparable mass.

6 Appendix

6.1 A1: Phase computation, Interpolation and Binning

As there are different lengths of the LCs having values at different time points, comparison of the LCs are only possible in terms of observations over each cycle. To tell exactly what the shape of a cycle is, all the cycles could be superimposed on top of each other. Hence each data point can be plotted, but instead of plotting the time, we would like to plot “how far it is into the cycle”. That way, all the cycles will be “folded” on top of each other, and we may have enough data to give us an accurate picture of what the cycle looks like. For a LC “how far it is into the cycle” is termed as its phase. So if the period of a star is known, and constant, it is possible to define phase

(Percy 2007)- the fraction of the star’s variability cycle which has elapsed. The phase is defined as,

$$\text{decimal portion of } [(t - t_0)/P], \quad (3)$$

where t is the time of the measurement of the star, here in HJD; t_0 is an arbitrary epoch - usually a time of maximum or minimum brightness, here the time of the first observed maximum; and P is the period of the star, here in days.

A graph of observed magnitude versus corresponding phase is called a phase diagram. Phase diagram for phase ranging from 0.0 to 1.0 shows one complete cycle of the corresponding star. A phase of 0 is the same as a phase of 1, -1 or 2. The standard phase (as given in equation (3)), always lies between 0 and 1, subtracting 1 gives the previous-cycle phase within -1 and 0, or adding +1 gives the next-cycle phase within 0 and 2.

Linear Spline or Piecewise Linear Interpolation is a famous time series segmentation method (Cassisi et al. 2012), which can be simply implemented to approximate astronomical time series. We also considered different cubic splines (Fritsch and Carlson 1980; Hyman 1983; Dougherty et al. 1989). In our situation, linear spline is preferred over cubic splines as cubic model fitting turned out to be inappropriate to our data set.

For clustering these time series, using the distance measure under consideration (discussed in Section 3.2), we need to get a full cycle over phase 0 to phase 1 for each LC having values at the same and equidistant phase points. For this purpose we performed the following steps:

- i) Using equation (3), the time points are converted into phases. It results in 1318 LCs of different lengths, at different phases, over different phase ranges. For the i^{th} LC, we get non-equidistant phase points from phase 0 to phase p_i (close to 1 but <1), where p_i = maximum of standard phases, available for the i^{th} LC, $i = 1, 2, \dots, 1318$.
- ii) For all the LCs, we have observations over phase interval $[0, p_i] \in [0, 1)$, provided values of p_i are different for all i . Here we do not use any extrapolation techniques as it involve larger uncertainty compared to interpolation, so to interpolate the series over phase $[0,1]$, we extend the phase interval of the i^{th} LC to $[0, p_i + 1] \in [0, 2)$, $i = 1, 2, \dots, 1318$, by adding +1 to the standard phases (computed in step (i)) and apply linear interpolation technique (described in step (iv)) to the observations over this extended phase interval.
- iii) In our data set, length of the given LCs varies from 130 to 264, except

one LC of length 5. For the analysis, we need LCs of equal length (say, l_*), which are equidistant over the phase interval $[0,1]$. Now, we fix this $l_* = 272$ empirically, so that there is no loss of information (as $l_* > 264$) and a length of 272 is not too large to increase the computational burden. Also, the error in linear interpolation decreases with the increase in l_* , provided l_* should not be large enough to bring about considerable error for the under-sampled series. Here all the LCs (except one) have lengths at least equal to 130 with most of them having lengths close to 272, so $l_* = 272$ does not cause significant error in approximating the series of lower lengths.

iv) Finally, piecewise linear interpolation is applied to the observations at the phases (computed in step (ii)) to obtain observations at 272 equidistant phase points over the phase interval $[0,1]$, using the following Linear Spline (Press et al. 1992):

Given a tabulated function $y_i = y(x_i)$, $i = 1, 2, \dots, N$ with $x_i < x_{i+1}$ for $i = 1, 2, \dots, N-1$, a linear spline is successive linear interpolations through the data points, i.e. it is a continuous piecewise degree-1 polynomial. The interpolating function joins $N-1$ linear functions of the form $f_i(x) = a_i y_i + b_i y_{i+1}$, $x \in [x_i, x_{i+1}]$, Here a_i and b_i are constants, and (a) $f_i(x_i) = y_i$, (b) $f_i(x_{i+1}) = y_{i+1}$. So, $a_i = \frac{x_{i+1}-x}{x_{i+1}-x_i}$ and $b_i = 1 - a_i = \frac{x-x_i}{x_{i+1}-x_i}$, $i = 1, 2, \dots, N-1$.

6.2 A2: k-Medoids: Partitioning Around Medoids (PAM)

We use the ‘‘PAM’’ algorithm for k-medoids clustering method (Kaufman and Rousseeuw 1990). The initial choice of k-medoids is quite important in the success of the output of the method. So it is vital not to choose them arbitrarily or subjectively. This algorithm takes care of this aspect. The algorithm consists of two phases. In the first phase, called BUILD, initial k-medoids are chosen such that the sum of the distances of all the objects in the data set from the first medoid is as small as possible, the sum of the distances of all the previously non-selected (i.e. non-medoid) objects in the data set from the second medoid is as small as possible, and so on. The algorithm is given below:

Firstly, choose the object of the data set for which the sum of the distances to all the other objects of the data set is minimal. Subsequently, at each step another object is selected such that the objective function is decreased as much as possible. To get this object, the following steps are carried out (Kaufman and Rousseeuw 1990):

1. Consider an object i which has not yet been selected.
2. Consider a non-selected object j and calculate the difference between its distance D_j with the most similar previously selected object, and its distance $d(j, i)$ with object i .
3. If this difference is positive, object j will contribute to the decision to select object i . Therefore we calculate

$$C_{ij} = \max(D_j - d(i, j), 0).$$

4. Object i is selected for which the total gain $\sum_j C_{ij}$ is the maximum.

This process stops when k objects have been selected as initial medoids.

In the second phase of the algorithm, called SWAP, we try to improve the clustering by improving the set of k medoids. We consider all pairs of objects (i, h) for which object i has been selected as medoid and object h has not, and we consider the swap when object i is no longer selected as medoid but object h is. Here we improve medoids so as to minimize the sum of the distances of the objects in the cluster from the cluster medoid. To calculate the effect of a swap between i and h on the value of the clustering, the following calculations are carried out (steps 1 and 2):

1. Consider a non-selected object j and calculate its contribution C_{jih} to the swap:
 - a. If j is more distant from both i and h than from one of the other medoids, C_{jih} is zero.
 - b. If j is not further from i than from any other selected medoid ($d(j, i) = D_j$), two situations must be considered:
 - b1. j is closer to h than to the second closest medoid

$$d(j, h) < E_j,$$

where E_j is the distance between j and the second most similar medoid. In this case the contribution of object j to the swap between objects i and h is

$$C_{jih} = d(j, h) - D_j.$$

- b2. j is at least as distant from h as from the second closest medoid

$$d(j, h) \geq E_j.$$

In this case the contribution of object j to the swap is

$$C_{jih} = E_j - D_j.$$

It should be observed that in situation b1 the contribution C_{jih} can be either positive or negative depending on the relative position of objects j , h and i . Only if object j is closer to i than to h , the contribution is positive, which indicates that the swap is not favorable from the point of view of object j . On the other hand, in situation b2 the contribution is always positive because it cannot be advantageous to replace i by an object h further away from j than from the second closest medoid.

c. j is more distant from object i than from at least one of the other medoids but closer to h than to any medoid. In this case the contribution of j to the swap is

$$C_{jih} = d(j, h) - D_j.$$

2. Calculate the total result of a swap by adding the contributions C_{jih} :

$$T_{ih} = \sum_j C_{jih}.$$

In the next steps it is decided whether to carry out a swap.

3. Select the pair (i, h) which minimizes T_{ih} .

4. If the minimum T_{ih} is negative, the swap is carried out and the algorithm returns to step 1. If the minimum T_{ih} is positive or 0, the value of the objective cannot be decreased by carrying out a swap and the algorithm stops.

It is noteworthy that as all potential swaps are considered, the results of the algorithm do not depend on the order of the objects in the input file (except in case some of the distances between objects are tied). The number of desired clusters (i.e. k) is required to be specified in advance. This algorithm instantly computes ASW (discussed in Section 3.3) for different values of k and helps us choose the number of clusters. It provides a novel graphical display, the Silhouette plot (Rousseeuw 1987) (discussed in Section 3.3).

6.3 A3: Silhouette Width

For each observation i , the Silhouette Width (SW) $s(i)$ is defined as follows (Rousseeuw et al. 1987):

$a(i)$ = the average distance between i and all other points of the cluster to which i belongs.

If i is the only observation in its cluster,

$$s(i) = 0.$$

For all other clusters C ,

$d(i; C)$ = the average distance of i to all observations of C .

The smallest of these $d(i; C)$ is,

$$b(i) = \min_C d(i; C).$$

This can be seen as the distance between i and its “neighbor” cluster, i.e. the nearest one to which it does not belong. Now, the number $s(i)$ is obtained by combining $a(i)$ and $b(i)$ as follows:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i), \\ 0 & \text{if } a(i) = b(i), \\ \frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i). \end{cases}$$

It is possible to write this in one formula:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

It is clear that

$$-1 \leq s(i) \leq 1.$$

Table 1: The Average Silhouette Width ($ASW \times 10^2$) for different number of clusters (k)

k	$ASW_{k-medoids}$	$ASW_{k-means}$
2	67.610	52.426
3	49.309	47.398
4	37.666	35.575
5	37.608	31.107
6	28.659	23.574

Table 2: Membership of subjective types in two groups k1 and k2

Type	k1	k2
EA	43	44
EB	84	24
EW	234	260
PUL	99	19
EA:	15	31
EB:	165	50
EW:	7	14
PUL:	145	23
CV:	1	0
EA/EB	6	4
EW/EA	2	4
EW/EB	8	4
EB/PUL	11	1
DCEP/PUL	9	0
CV/PUL	9	2
Total	838	480

Note: An uncertain type is followed by a colon and an ambiguous type is given with a slash.

Table 3: Average values of the parameters for two groups k1 and k2 from k-medoids clustering

Name of cluster	No. of members	P (day)	R (mag)	B (mag)	I (mag)	B-I (mag)	R-I (mag)
k1	838	2.816 ± 0.045	17.783 ± 0.036	19.917 ± 0.045	18.597 ± 0.045	1.320 ± 0.025	0.814 ± 0.020
k2	480	1.400 ± 0.125	19.636 ± 0.062	22.046 ± 0.075	20.614 ± 0.075	1.432 ± 0.040	0.978 ± 0.028

Table 4: Average values of the parameters for two groups c1 and c2 obtained through k-means clustering

Name of cluster	No. of members	P (day)	R (mag)	B (mag)	I (mag)	B-I (mag)	R-I (mag)
c1	1173	2.427 ± 0.102	18.371 ± 0.042	16.133 ± 0.038	17.521 ± 0.041	1.388 ± 0.022	0.850 ± 0.017
c2	145	1.274 ± 0.223	19.155 ± 0.117	16.945 ± 0.108	18.086 ± 0.118	1.141 ± 0.068	1.069 ± 0.049

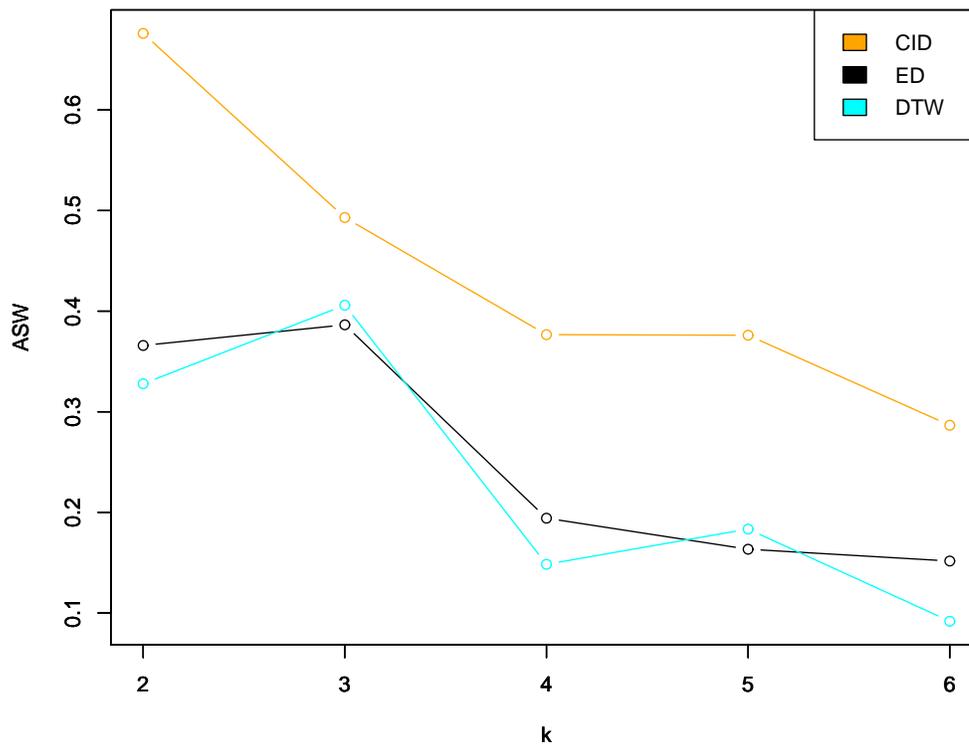


Figure 1: The Average Silhouette Width (ASW) for different number of clusters (k) corresponding to three different distance measures in combination with k -medoids clustering method. The circles indicate values of ASW corresponding to a value of k .

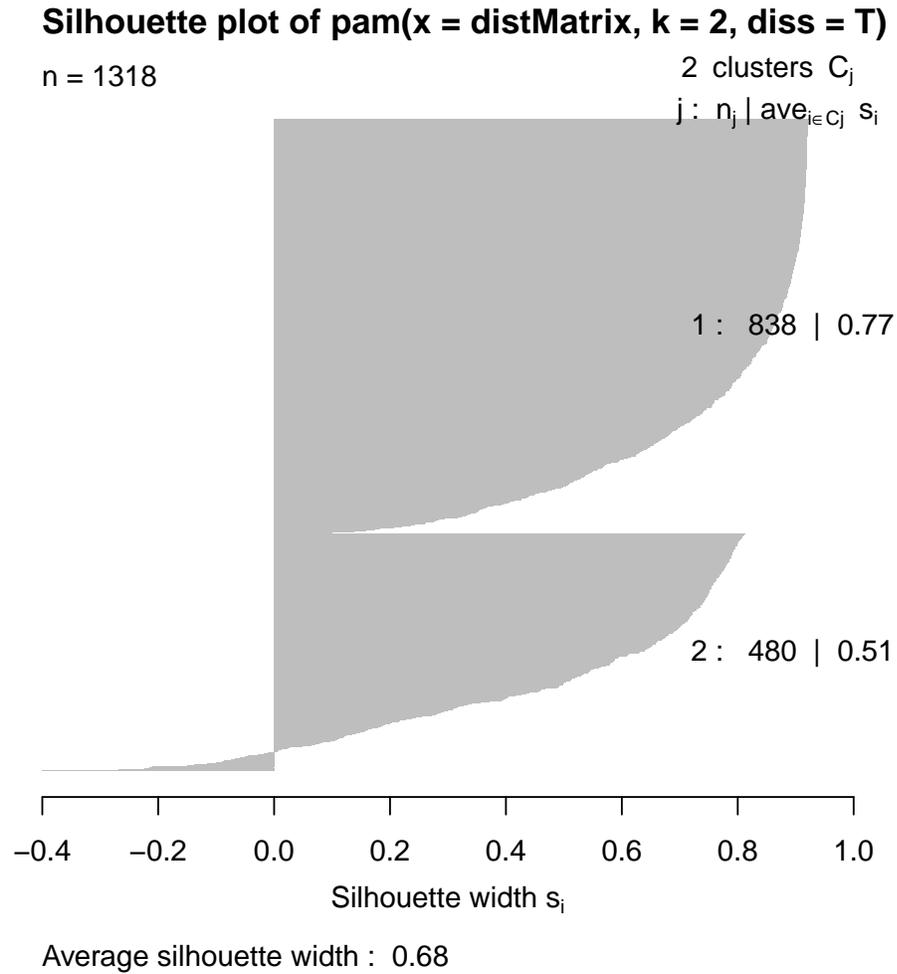


Figure 2: The Silhouette plot gives a graphical representation of Silhouette Width of each of the light curves belonging to individual clusters, resulted in the k-medoids clustering method through CID for $k = 2$. The grey shade indicates the Silhouette Width of a light curve, arranged in descending (from top to bottom) order for individual clusters. The Average Silhouette Width for cluster 1, cluster 2 and the whole data set of respective sizes of 838, 480 and 1318 are computed as 0.77, 0.51 and 0.68, respectively.

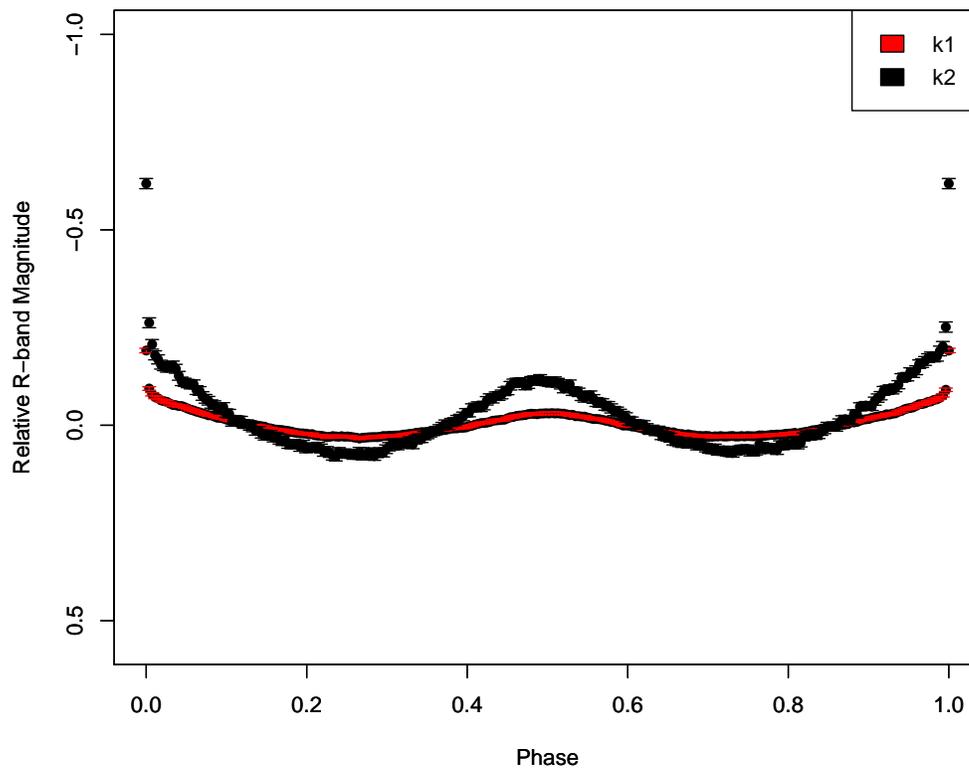


Figure 3: Template average light curves of two clusters k1 and k2, with standard error, obtained from k-medoids clustering with CID.

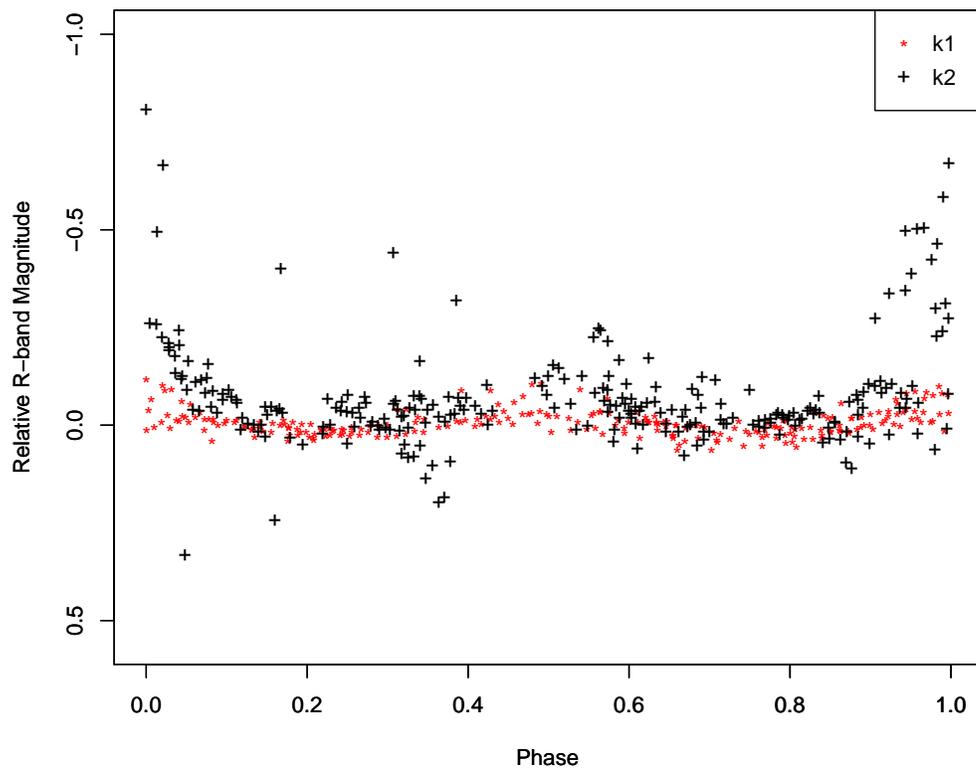


Figure 4: Template Medoid light curves, with ID V-1221 and V-1138 (see, Miller et al. (2010)), for two clusters k1 and k2 respectively obtained from k-medoids clustering with CID.

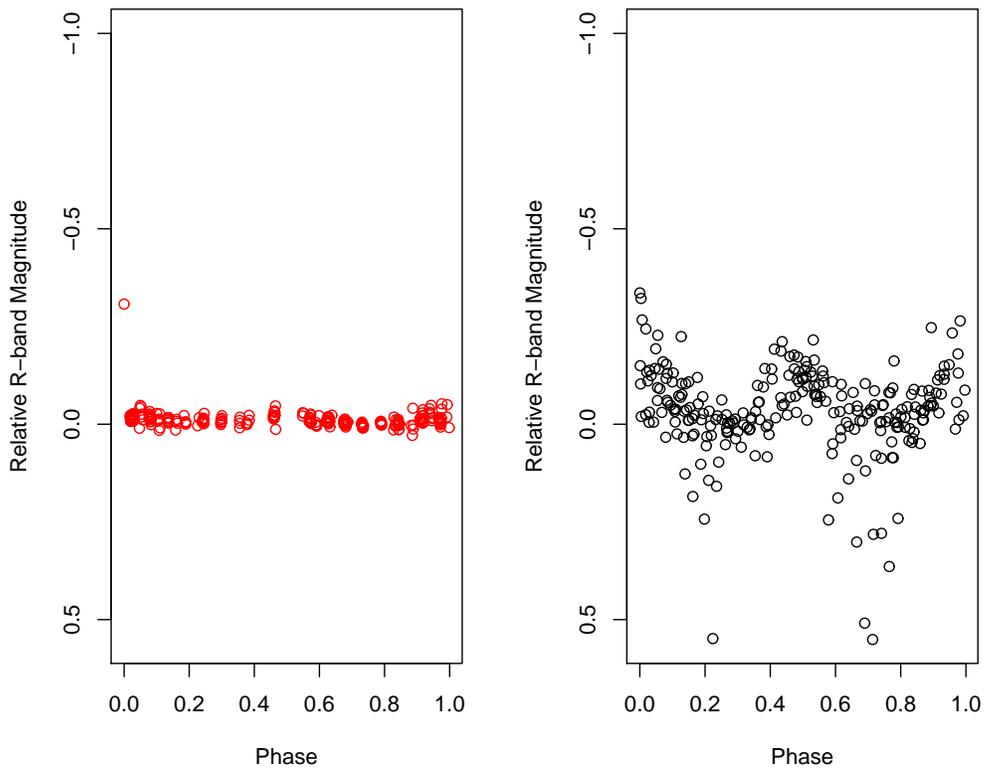


Figure 5: A pair of representative light curves, which are the observed light curves with ID V-94 and V-384 (see, Miller et al. (2010)), of k1 (left) and k2 (right) respectively obtained from k-medoids clustering with CID.

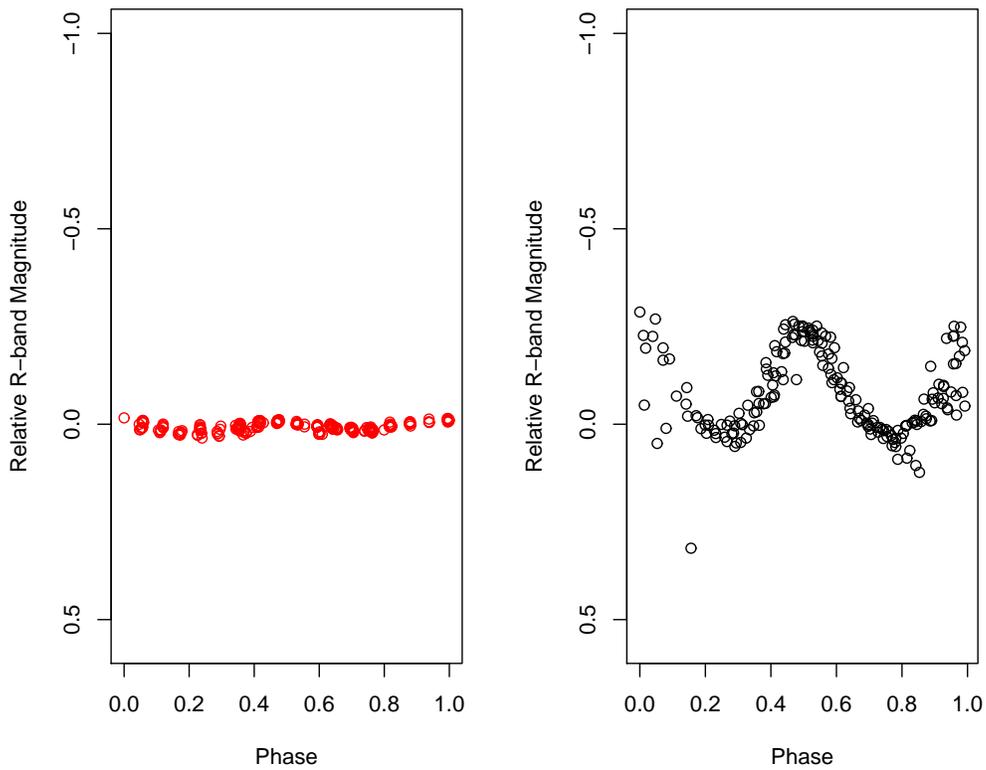


Figure 6: Another pair of representative light curves, which are the observed light curves with ID V-334 and V-817 (see, Miller et al. (2010)), of k1 (left) and k2 (right) respectively obtained from k-medoids clustering with CID.

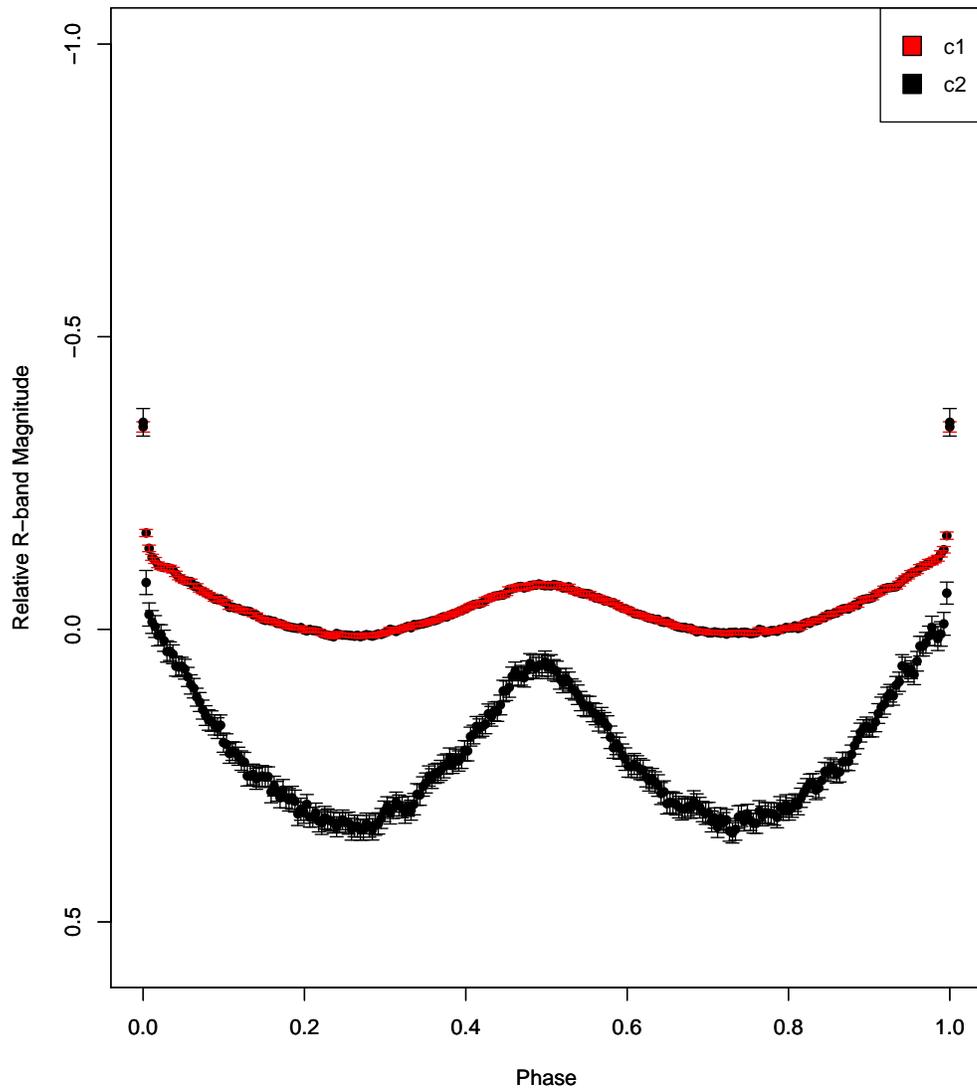


Figure 7: Template average light curves of two clusters c1 and c2, with standard error, obtained from k-means clustering.

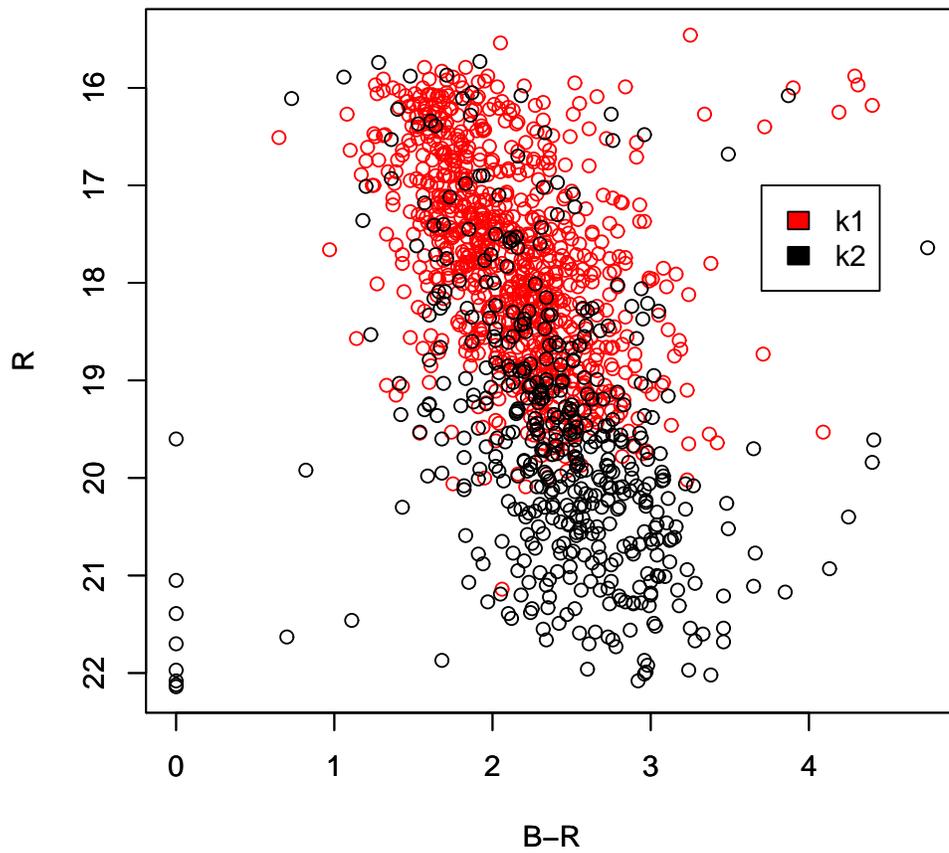


Figure 8: Colour-magnitude diagram of the variable stars clustered in two groups k1 and k2 through k-medoids clustering method with CID.

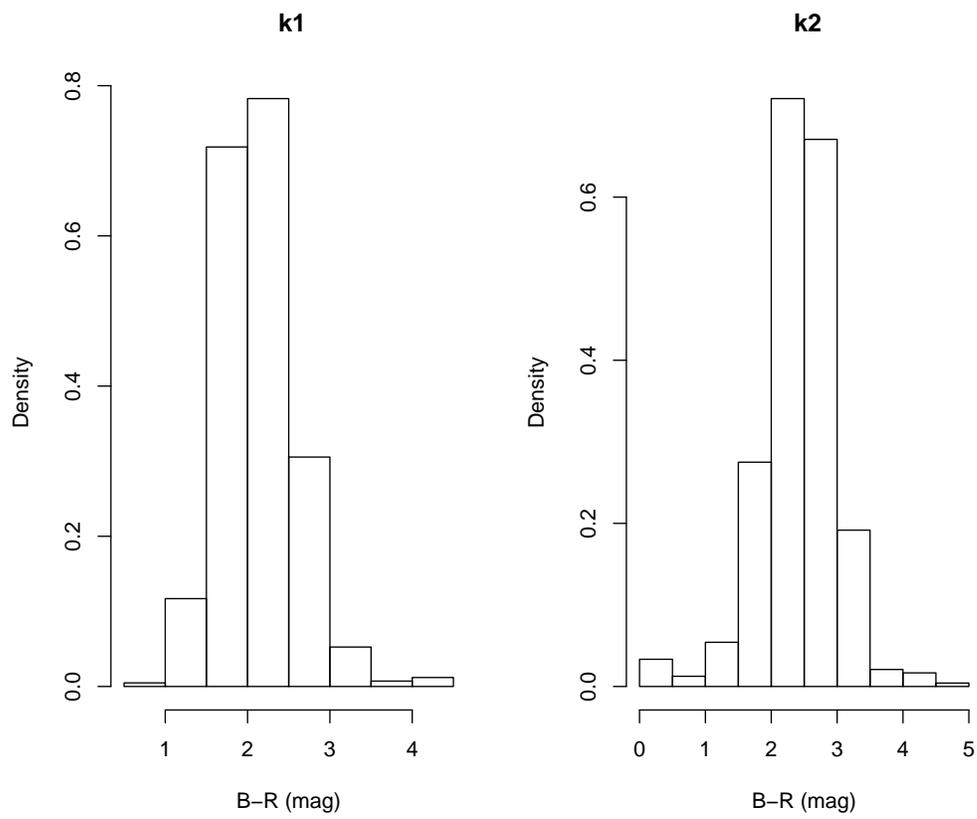


Figure 9: Histograms of B-R colour index of two clusters k1 and k2 obtained from k-medoids clustering with CID.

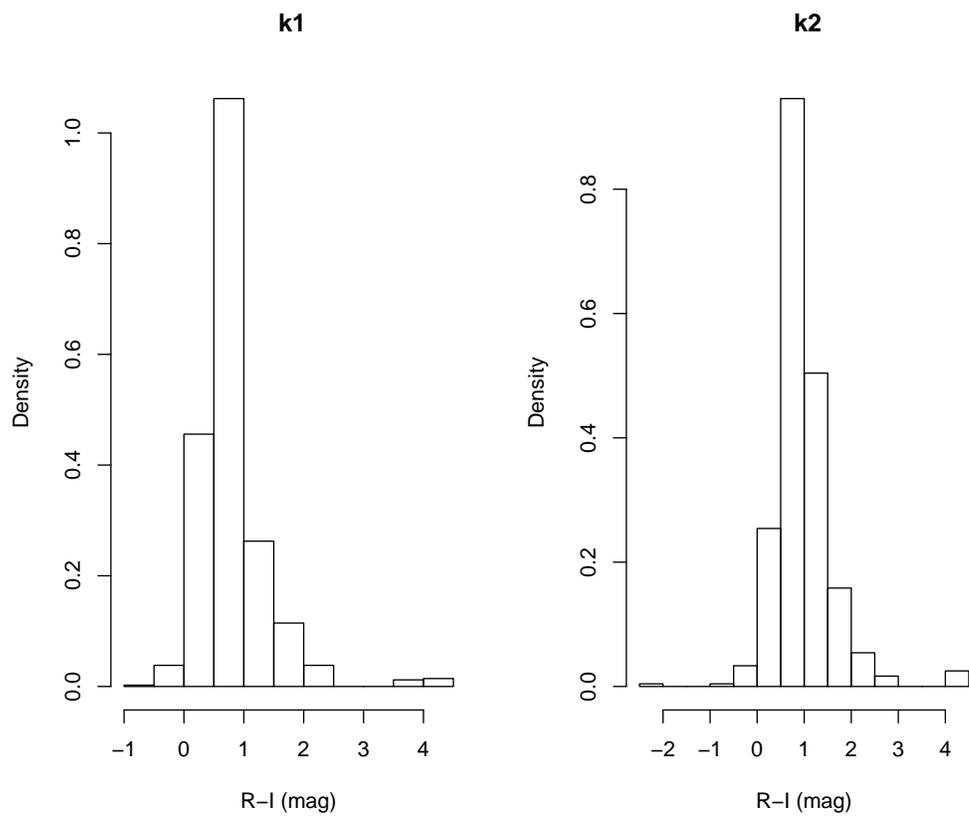


Figure 10: Histograms of R-I colour index of two clusters k1 and k2 obtained from k-medoids clustering with CID.

References

- [1] Akerlof, C., Amrose, S., Balsano, R., Bloch, J., Casperson, D., Fletcher, S., Gisler, G., Hills, J., Kehoe, R., Lee, B., Marshall, S., McKay, T., Pawl, A., Schaefer, J., Szymanski, J., Wren, J.: ROTSE All-Sky Surveys for Variable Stars. I. Test Fields. *The Astronomical Journal*. 119(4), 1901–1913 (2000)
- [2] Albrow, M., An, J., Beaulieu, J.-P., Caldwell, J. A. R., Dominik, M., Greenhill, J., Hill, K., Kane, S., Martin, R., Menzies, J., Pollard, K., Sackett, P. D., Sahu, K. C., Vermaak, P., Watson, R., Williams, A., PLANET Collaboration, Hauschildt, P. H.: $H\alpha$ Equivalent Width Variations across the Face of a Microlensed K Giant in the Galactic Bulge. *The Astrophysical Journal*. 550(2), L173–L177 (2001)
- [3] Avvakumova, E.A., Malkov, O.Yu. and Kniazev, A.Yu.: Eclipsing variables: Catalogue and classification. *Astronomische Nachrichten*. 334(8), 860-865 (2013)
- [4] Batista, G. E. A. P. A., Keogh, E. J., Tataw, O. M., de Souza, V. M. A.: CID: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*. 28(3), 634–669 (2014)
- [5] Bayne, G., Tobin, W., Pritchard, J. D., Bond, I., Pollard, K. R., Besier, S. C., Noda, S., Sumi, T., Yanagisawa, T., Sekiguchi, M., Honda, M., Muraki, Y., Takeuti, M., Hearnshaw, J. B., Kilmartin, P. M., Dodd, R. J., Sullivan, D. J., Yock, P. C. M.: The MOA catalogue of eclipsing binary stars in the Small Magellanic Cloud. *Monthly Notices of the Royal Astronomical Society*. 331(3), 609–614 (2002)
- [6] Bradstreet, D. H. and Steelman, D. P.: Binary Maker 3.0 - An Interactive Graphics-Based Light Curve Synthesis Program Written in Java. American Astronomical Society, 201st AAS Meeting, id.75.02; *Bulletin of the American Astronomical Society*. Vol. 34, p. 1224 (2002)
- [7] Caiado, J., Crato, N. and Pea, D.: A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*. 50, 2668–2684 (2006)

- [8] Caiado, J., Crato, N. and Pea, D.: Comparison of times series with unequal length in the frequency domain. *Communications in StatisticsSimulation and Computation*. Vol. 38, Issue 3, Pages 527–540 (2009)
- [9] Cassisi, C., Montalto, P., Aliotta, M., Cannata, A. and Pulvirenti, A.: Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining. *Advances in Data Mining Knowledge Discovery and Applications*, pp. 71–96. Intech. (2012)
- [10] Chattopadhyay, T., Misra, R., Chattopadhyay, A.K. and Naskar, M.: Statistical evidences of three classes of Gamma Ray Bursts. *Astrophysical Journal*. 667, 1017–1023 (2007a)
- [11] Chattopadhyay, T. and Chattopadhyay, A.K.: Globular clusters of the Local Group statistical classification. *Astronomy and Astrophysics*. 472(1), 131–140 (2007b)
- [12] Chattopadhyay, A.K., Chattopadhyay, T., Mondal, S., Sharina, M. and Davoust, E.: Study of NGC5128 globular clusters under multivariate statistical paradigm. *Astrophysical Journal*. 705, 1533–1547 (2009)
- [13] Chattopadhyay, A.K., Fraix-Burnet, D., Dugue, M., Chattopadhyay, T. and Davoust, E.: Structures in the fundamental plane of early-type galaxies. *Monthly Notices of the Royal Astronomical Society*. 407, 2207–2222 (2010)
- [14] Chattopadhyay, T., Sharina, M., Davoust, E., De, T. and Chattopadhyay, A.K.: Uncovering the formation of ultracompact dwarf galaxies by multivariate statistical analysis. *The Astrophysical Journal*. 750:91 (13pp), (2012)
- [15] Chattopadhyay, A.K., Mondal, S. and Chattopadhyay, T.: Independent Component Analysis for the objective classification of globular clusters of the galaxy NGC 5128. *Computational Statistics and Data Analysis*. 57, 17–32 (2013)
- [16] Chattopadhyay, T., Sinha, A., Chattopadhyay, A. K.: Influence of binary fraction on the fragmentation of young massive clusters– a Monte Carlo simulation. *Astrophysics and Space Science*. 361, 120 (2016)

- [17] Clifford, H., Wessely, F., Pendurthi, S. and Emes, R.D.: Comparison of clustering methods for investigation of genome-wide methylation array data. *Front Genet.* Volume 2, Article 88 (2011). doi: 10.3389/fgene.2011.00088
- [18] Dargahi-Noubary, G. R.: Discrimination between Gaussian time series based on their spectral differences. *Communications in Statistics: Theory and Methods.* 21, 2439–2458 (1992)
- [19] Dougherty, R. L., Edelman, A. and Hyman, J. M.: Positivity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation. *Mathematics of Computation.* 52, 471-494 (1989)
- [20] Eckner, A.: A Framework for the Analysis of Unevenly Spaced Time Series Data. Working Paper (2014). URL: http://eckner.com/papers/unevenly_spaced_time_series_analysis.pdf.
- [21] Eckner, A.: Algorithms for Unevenly-spaced time series: Moving averages and other rolling operators. Working Paper (2017). URL: <http://eckner.com/papers/Algorithms%20for%20Unevenly%20Spaced%20Time%20Series.pdf>.
- [22] Fraix-Burnet, D., Chattopadhyay, T., Chattopadhyay, A.K., Davoust, E., and Thuillard, M.: A six-parameter space to describe galaxy diversification. *A&A* 545, A80 (2012). DOI: 10.1051/0004-6361/201218769
- [23] Fraix-Burnet, D., Dogué, M., Chattopadhyay, T., Chattopadhyay, A.K. and Davoust, E.: Structures in the fundamental plane of early-type galaxies. *Monthly Notices. Royal Astro. Soc.* 407, 2207–2222 (2010)
- [24] Fritsch, F. N. and Carlson, R. E.: Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis* 17, 238-246 (1980)
- [25] Fröhwrth-Schnatter, S., Kaufmann, S.: Model-based clustering of multiple time series. *Journal of Business & Economic Statistics.* 26(1), 78–89 (2008). doi:10.1198/073500107000000106
- [26] Giménez, A., Clausen, J. V., Guinan, E. F., Maloney, F. P., Bradstreet, D. H., Storm, J., Tobin, W.: Eclipsing Binaries as Accurate Distance Indicators to Nearby Galaxies. *Experimental Astronomy.* Volume 5, Issue 1–2, pp. 181–183 (1994)

- [27] Giorgino, T.: Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7), 1–24 (2009). URL: www.jstatsoft.org/v31/i07/.
- [28] Graczyk, D., Soszyński, I., Poleski, R., Pietrzyński, G., Udalski, A., Szymański, M. K., Kubiak, M., Wyrzykowski, L., Ulaczyk, K.: The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XII. Eclipsing Binary Stars in the Large Magellanic Cloud. *Acta Astronomica*. 61, no. 2, pp. 103–122 (2011)
- [29] Helminiak, K. G., Konacki, M., Różyczka, M., Kałużny, J., Ratajczak, M., Borkowski, J., Sybilski, P., Muterspaugh, M. W., Reichart, D. E., Ivarsen, K. M., Haislip, J. B., Crain, J. A., Foster, A. C., Nysewander, M. C., LaCluyze, A. P.: Orbital and physical parameters of eclipsing binaries from the All-Sky Automated Survey catalogue - IV. A $0.61 + 0.45 M_{\odot}$ binary in a multiple system. *Monthly Notices of the Royal Astronomical Society*. 425(2), 1245–1256 (2012)
- [30] Hyman, J. M.: Accurate monotonicity preserving cubic interpolation. *SIAM J. Sci. Stat. Comput.* 4, 645–654 (1983)
- [31] Kalpakis, K., Gada, D., Puttagunta, V.: Distance measures for effective clustering of ARIMA time-series. *Proceedings of the 2001 IEEE International Conference on Data Mining*, San Jose, CA, pp. 273–280 (2001)
- [32] Kaufman, L. and Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. pp. 68–125, Wiley, New York (1990)
- [33] Keogh, E. and Ratanamahatana, C. A.: Exact indexing of dynamic time warping. *Knowledge and Information Systems*. 7(3), 358–386 (2005)
- [34] Kirk, B., Conroy, K., Prša, A. et al.: Kepler Eclipsing Binary Stars. VII. The Catalog of Eclipsing Binaries Found in the Entire Kepler Data Set. *The Astronomical Journal*. 151, 68–88 (2016)
- [35] Kochoska, A., Mowlavi, N., Prša, A., Lecoer-Taïbi, I., Holl, B., Rimoldini, L., Süveges, M., Eyer, L.: Gaia eclipsing binary and multiple systems. A study of detectability and classification of eclipsing binaries with Gaia. *Astronomy & Astrophysics*. 602, A110 (2017)

- [36] Liao, T.W.: Clustering of time series data-a survey. *Pattern Recognition*. 38(11), 1857-1874 (2005)
- [37] Liao, T.W., Ting, C. & Chang, P.-C.: An adaptive genetic clustering method for exploratory minning of feature vector any time series data. *International Journal of Production Research*. 44:14, 2731–2748 (2006). doi:10.1080 /00207540600 600130
- [38] Lomb, N. R.: Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*. 39, 447–462 (1976)
- [39] López-Morales, M. and Clemens, J. C.: The Pisgah Automated Survey: A Photometric Search for Low-Mass Detached Eclipsing Binaries and Other Variable Stars. *The Publications of the Astronomical Society of the Pacific*. 116(815), 22–37 (2004)
- [40] Malkov, O. Yu., Oblak, E., Snegireva, E. A., Torra, J.: A catalogue of eclipsing variables. *Astronomy & Astrophysics*. 446(2), 785–789 (2006)
- [41] Malkov, O. Yu., Oblak, E., Avvakumova, E. A. and Torra, J.: Classification of Eclipsing Binaries. *Solar and Stellar Physics Through Eclipses, ASP Conference Series*. Vol. 370 (2007). Ed: O. Demircan, S. O. Selam & B. Albayrak
- [42] Malkov, O. Yu., Avvakumova, E. A.: Classification of eclipsing binaries: attractive systems. *Central European Astrophysical Bulletin*. 37, 173-185 (2013)
- [43] Matijevič, G., Prša, A., Orosz, J. A., Welsh, W. F., Bloemen, S., Barclay, T.: Kepler Eclipsing Binary Stars. III. Classification of Kepler Eclipsing Binary Light Curves with Locally Linear Embedding. *The Astronomical Journal*. 143, 123–128 (2012)
- [44] Miller, V. R., Albrow, M. D., Afonso, C., Henning Th.: 1318 new variable stars in a 0.25 square degree region of the Galactic plane. *Astronomy & Astrophysics*. 519, A12 (2010)
- [45] Modak, S., Chattopadhyay, A. K. & Chattopadhyay, T.: Clustering of gamma-ray bursts through kernel principal component analysis. *Communications in Statistics - Simulation and Computation*. (2017). DOI: 10.1080/03610918.2017.1307393

- [46] Moller-Levet, C.S., Klawonn, F., Cho, K., Wolkenhauer, O.: Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points. *Advances in Intelligent Data Analysis V Lecture Notes in Computer Science*. 2810, 330–340 (2003)
- [47] Mowlavi, N., Lecoeur-Taïbi, I., Holl, B., Rimoldini, L., Barblan, F., Prsa, A., Kochoska, A., Süveges, M., Eyer, L., Nienartowicz, K., Jevardat, G., Charnas, J., Guy, L., Audard, M.: Gaia eclipsing binary and multiple systems. Two-Gaussian models applied to OGLE-III eclipsing binary light curves in the Large Magellanic Cloud. *Astronomy & Astrophysics*. 606, A92 (2017)
- [48] Niarchos, P. G.: On the Gaia Expected Harvest on Eclipsing Binaries. *Astrophysics and Space Science*. Volume 304, Issue 1–4, pp. 387–390 (2006)
- [49] Percy, J.R.: *Understanding Variable Stars*. Cambridge University Press, New York (2007)
- [50] Prati, R. C. and Batista, G. E. A. P. A.: A Complexity-Invariant Measure Based on Fractal Dimension for Time Series Classification. *International Journal of Natural Computing Research*. 3(3), 59–73 (2012)
- [51] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, W.T.: *Numerical Recipes in C. The Art of Scientific Computing*, Second Edition, pp. 105–128. Cambridge University Press, Cambridge (1992)
- [52] Prša, A., Guinan, E. F., Devinney, E. J., DeGeorge, M., Bradstreet, D. H., Giammarco, J. M., Alcock, C. R., Engle, S. G.: Artificial Intelligence Approach to the Determination of Physical Properties of Eclipsing Binaries. I. The EBAI Project. *The Astrophysical Journal*. 687, 542–565, (2008)
- [53] Rabiner, L. and Juang, B.-H: *Fundamentals of speech recognition*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA (1993)
- [54] Rousseeuw, P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 20, 53-65 (1987)

- [55] Sarro, L. M., Sánchez-Fernández, C., Giménez, Á.: Automatic classification of eclipsing binaries light curves using neural networks. *Astronomy & Astrophysics*. 446, 395-402 (2006)
- [56] Scargle, J. D.: Studies in astronomical time series analysis. III - Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. *Astrophysical Journal*. 343, 874–887 (1989)
- [57] Singh, S.S., Chauhan, N.C.: K-Means v/s K-Medoids: A Comparative Study. National Conference on recent trends in Engineering And Technology. (2011)
- [58] Soszyński, I., Udalski A., Szymański, M. K., Kubiak, M., Pietrzyński, G., Wyrzykowski, L., Szewczyk, O., Ulaczyk, K., Poleski, R.: The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. II. Type II Cepheids and Anomalous Cepheids in the Large Magellanic Cloud. *Acta Astronomica*. 58, 293 (2008a)
- [59] Soszyński, I., Poleski, R., Udalski, A., Szymański, M. K., Kubiak, M., Pietrzyński, G., Wyrzykowski, L., Szewczyk, O., Ulaczyk, K.: The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. I. Classical Cepheids in the Large Magellanic Cloud. *Acta Astronomica*. 58, 163–185 (2008b)
- [60] Stefan, A., Athitsos, V., Das G.: The Move-Split-Merge Metric for Time Series. *IEEE Transactions on Knowledge & Data Engineering*. 25(6), 1425–1438 (2013)
- [61] Street, R. A., Christian, D. J., Clarkson, W. I., Collier Cameron, A., Evans, N., Fitzsimmons, A., Haswell, C. A., Hellier, C., Hodgkin, S. T., Horne, K., Kane, S. R., Keenan, F. P., Lister, T. A., Norton, A. J., Pollacco, D., Ryans, R., Skillen, I., West, R. G., Wheatley, P. J.: Status of SuperWASP I (La Palma). *Astronomische Nachrichten*. 325(6), 565–567 (2004)
- [62] Süveges, M., Barblan, F., Lecoœur-Taïbi, I., Prša, A., Holl, B., Eyer, L., Kochoska, A., Mowlavi, N., Rimoldini, L.: Gaia eclipsing binary and multiple systems. Supervised classification and self-organizing maps. *Astronomy & Astrophysics*. 603, A117 (2017)

- [63] Velmurugan, T. and Santhanam, T.: A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach. *Information Technology Journal*. 10, 478–484 (2011)
- [64] Wei, Y.: Multi-dimensional time warping based on complexity invariance and its application in sports evaluation. 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE. 677–680 (2014)
- [65] Weldrake, David, T. F., Sackett, Penny, D., Bridges, Terry, J., Freeman, Kenneth, C.: A Comprehensive Catalog of Variable Stars in the Field of 47 Tucanae. *The Astronomical Journal*. 128(2), 736–760 (2004)
- [66] Weldrake, David, T. F., Sackett, Penny, D., Bridges, Terry, J.: A Deep Wide-Field Variable Star Catalog of ω Centauri. *The Astronomical Journal*. 133(4), 447–1469 (2007)
- [67] Wozniak, P. R., Udalski, A., Szymanski, M., Kubiak, M., Pietrzynski, G., Soszynski, I., Zebrun K.: Difference Image Analysis of the OGLE-II Bulge Data. III. Catalog of 200000 Candidate Variable Stars. *Acta Astronomica*. 52, 129–142 (2002)
- [68] Wyithe, J. S. B. and Wilson, R. E.: Photometric Solutions for Semidetached Eclipsing Binaries: Selection of Distance Indicators in the Small Magellanic Cloud. *The Astrophysical Journal*. 571(1), 293–319 (2002)
- [69] Wyrzykowski, L., Udalski, A., Kubiak, M., Szymanski, M. K., Zebrun, K., Soszynski, I., Wozniak, P. R., Pietrzynski, G., Szewczyk, O.: The Optical Gravitational Lensing Experiment. Eclipsing Binary Stars in the Small Magellanic Cloud. *Acta Astronomica*. 54, 1–17 (2004)