Exploring the Variable Sky with LINEAR. III. Classification of Periodic Light Curves

Lovro Palaversa[1],

lovro.palaversa@unige.ch

Željko Ivezić[2,3,4] Sarah Loebman[2], Domagoj Ruždjak[4], Davor Sudar[4], Mario Galin[5], Andrea Kroflin[3], Martina Mesarić[3], Petra Munk[3], Dijana Vrbanec[3], Hrvoje Božić[4], Nicholas Hunt-Walker[2], Jacob VanderPlas[2], David Westman[2], Laurent Eyer[1], J. Scott Stuart[6], Branimir Sesar[7], Andrew C. Becker[2], Przemyslaw Wozniak[8], Hakeem Oluseyi[9],

Received _____; accepted _____

[1]Observatoire astronomique de l'Université de Genève, 51 chemin des Maillettes, CH-1290 Sauverny, Switzerland

[2]University of Washington, Department of Astronomy, P.O. Box 351580, Seattle, WA 98195-1580

[3]Department of Physics, Faculty of Science, University of Zagreb, Bijenička cesta 32, 10000 Zagreb, Croatia

[4]Hvar Observatory, Faculty of Geodesy, Kačićeva 26, 10000 Zagreb, Croatia

[5]Faculty of Geodesy, Kačićeva 26, 10000 Zagreb, Croatia

[6]Lincoln Laboratory, Massachusetts Institute of Technology, 244 Wood Street, Lexington, MA 02420-9108

[7]Division of Physics, Mathematics and Astronomy, Caltech, Pasadena, CA 91125

[8]Los Alamos National Laboratory, 30 Bikini Atoll Rd., Los Alamos, NM 87545-0001

[9]Florida Institute of Technology, Melbourne, FL 32901

ABSTRACT

We describe the construction of a highly reliable sample of ∼7,000 optically faint periodic variable stars with light curves obtained by the asteroid survey LINEAR across 10,000 deg^2 of sky. The sample flux limit is several magnitudes fainter than for most other wide-angle surveys; the photometric errors range from ∼0.03 mag at $r = 15$ to ∼0.20 mag at $r = 18$. Light curves include on average 250 data points, collected over about a decade. Using SDSS-based photometric recalibration of the LINEAR data for about 25 million objects, we selected ∼200,000 most probable candidate variables with $r < 17$ and visually confirmed and classified ∼7,000 periodic variables using phased light curves. The reliability and uniformity of visual classification across eight human classifiers was calibrated and tested using a catalog of variable stars from the SDSS Stripe 82 region, and verified using unsupervised machine learning approach. The resulting sample of periodic LINEAR variables is dominated by 3,900 RR Lyrae stars and 2,700 eclipsing binary stars of all subtypes, and includes small fractions of relatively rare populations such as asymptotic giant branch stars and SX Phoenicis stars. A likely candidate for a heartbeat star with currently longest known period was also found. We discuss the distribution of these mostly uncataloged variables in various diagrams constructed with optical-to-infrared SDSS, 2MASS and WISE photometry, and with LINEAR light curve features. We find that combination of light curve features and colors enables classification schemes much more powerful than when colors or light curves are each used separately. An interesting side result is a robust and precise quantitative description of a strong correlation between the light-curve period and color/spectral type for close and contact eclipsing binary stars (β Lyrae and W UMa): as the color-based spectral type varies from K4 to F5, the median period increases from 5.9 hours to 8.8 hours.

These large samples of robustly classified variable stars will enable detailed statistical studies of Galactic structure and physics of binary and other stars, and we make them publicly available.

Subject headings: variable stars: general — pulsating: RR Lyrae, δ Scuti, SX Phoenicis, Mira, long period, semi-regular — binaries: eclipsing — astronomical databases: catalogs, surveys, classification

1. Introduction

Variability is an important phenomenon in astrophysical studies of structure and evolution, both stellar, Galactic and extragalactic. Its importance will only increase with the advent of massive time domain surveys, such as Gaia (Eyer et al. 2012) and LSST (Ivezić et al. 2008a), where the expected number of identified variable stars will reach hundreds of millions – roughly the same as the number of all the stars detected by the Sloan Digital Sky Survey (SDSS; York et al. 2000). Such a large number of light curves can be fully analyzed only using automated machine learning methods (e.g., Debosscher et al. 2007; Richards et al. 2011). Most such methods require reliable training samples; in addition to astrophysical motivation for improved understanding of the optical variability of faint sources, a goal of analysis presented here is to construct a large training sample of periodic variable stars that probes both a large sky area and faint magnitude range.

This paper is the third one in a series based on light curve data collected by the asteroid LINEAR survey in the period roughly from 1998 to 2009. In the first paper Sesar et al. (2011) described the LINEAR survey and photometric recalibration based on SDSS stars acting as a dense grid of standard stars. In the overlapping ∼10,000 deg^2 of sky between LINEAR and SDSS, Sesar et al. obtained photometric errors ranging from ∼0.03 mag for sources not limited by photon statistics to ∼0.20 mag at $r = 18$ (here r is the SDSS r band magnitude). LINEAR data provide time domain information for the brightest 4 magnitudes of SDSS survey, with 250 unfiltered photometric observations per object on average (rising to ∼500 along the Ecliptic). The public access to the recalibrated LINEAR data, including over 5 billion photometric measurements for about 25 million objects (about three quarters are stars; ∼5 million objects have $r < 17$ and photometric errors below about 0.1 mag) is provided through the SkyDOT Web site (https://astroweb.lanl.gov/lineardb/). Positional matches to SDSS and 2MASS (Skrutskie et al. 2006) catalog entries are also available for

the entire sample. In this work we also provide positional matches to WISE catalog entries (Wright et al. 2012) for confirmed periodic variables.

Sesar et al. (2011) compared LINEAR dataset to other prominent contemporary wide-area variability surveys in terms of depth and cadence. LINEAR extends the deepest similar wide-area variability survey, the Northern Sky Variability Survey (Woźniak et al. 2004), by 3 mag. This improvement in depth is significant; for example, it can be used to extend distance limit for Galactic structure studies based on RR Lyrae stars by a factor of 4 (to about ∼30 kpc; for details see the second paper in this series, Sesar et al. 2013). Thanks to the improved faint limit, the sample includes over a thousand quasars (for $r < 17$; for detailed analysis see Ruan et al. 2012). The large sky area, with resulting increase in sample sizes, enables robust statistical studies of samples such as eclipsing binary stars, and searches for rare objects (e.g., field SX Phe stars, asymptotic giant branch stars). In addition to these specific programs, the depth improvement of 3 mag will help quantify the variation of the composition of the variable source population with depth. For example, Eyer & Blake (2005) determined that 83% of variable objects with $V < 14$ are red giants, while in contrast Sesar et al. (2007) found that two thirds of variable objects with $14 < V < 20$ are RR Lyrae and quasars).

In order to make scientific use of the LINEAR dataset, the completeness and purity for samples of selected variable objects need to be understood and quantified. There are a number of automated methods for selecting variable objects and classifying their light curves proposed in the literature (e.g., Eyer & Blake 2005; Debosscher et al. 2007; Richards et al. 2011, and references therein). Measuring the performance of these methods on LINEAR dataset requires reliable training sample and full understanding of the photometric error distribution. It would be difficult to quantify the performance of these methods on LINEAR dataset because there are no reliable training samples, and the photometric

error distribution is not fully understood yet. The LINEAR survey was not designed as a photometric survey, and more importantly, it accepted data obtained in non-photometric conditions. Although the LINEAR photometric error distribution obtained by Sesar et al. (2011) is close to Gaussian, various tests show that of the order 1% of measurements can have anomalous errors (defined here as errors at least three times larger than reported errors) that are hard to recognize using available metadata (such as photometric zeropoint information and the photometric scatter for calibration stars). A part of the problem may be the fact that a large fraction of observations are obtained along the Ecliptic where contamination by blended main belt asteroids is not negligible.

Despite the fraction of measurements with anomalous errors as small as 1%, the resulting sample contamination can be substantial. According to Sesar et al. (2007), about 2% of objects with $14 < V < 20$ are variable at the 0.05 mag level (root-mean-square scatter, rms). Given that practical cutoff on rms is about 0.1 mag for the LINEAR dataset, and excluding quasars which are not numerous at magnitudes probed by LINEAR (fewer than 0.1% of objects in the LINEAR sample with $r < 18$ are quasars), robustly detectable variability is thus expected for much less than 1% of the sample. Hence, even if only 1% of the LINEAR sample is spuriously selected as variable star candidates, the resulting false positives would dominate the sample.

In order to better understand the behavior of photometric errors in the LINEAR sample, and to ultimately enable deployment of automated methods for selecting variable objects and classifying their light curves, we have undertaken an extensive program of visual classification of about 200,000 light curves by eight human classifiers. Further details about visual classification and the construction of the resulting sample of about 7,000 robust periodic variables are described in §2. The distribution of periodic variables, dominated by roughly equal fractions of RR Lyrae stars and eclipsing binary stars, in various color-color

and other diagrams is discussed in §3. We compare our results to existing variable star catalogs §4, and to supervised and unsupervised machine learning classification methods in §5. Our main results are discussed and summarized in §6.

2. Visual Classification of LINEAR Light Curves

The main goal of our analysis is the selection of a large robust sample of periodic variable stars, with a high purity (i.e., low contamination) within adopted flux, amplitude and period limits. To improve the sample robustness and light curve classification, we undertook three successive selection and classification steps. After the initial sample selection, period estimation and construction of phased light curves, eight human classifiers extracted about 7,000 likely periodic variables from a starting set of about 200,000 candidate variables, and also obtained initial light curve classification. In the following two steps, a single expert refined selection and classification of the smaller sample of 7,000 likely periodic variables, first by repeating visual classification, and then aided by the parameters measured from light curves and other information, such as photometry. In this section we first describe the initial sample selection and period estimation, and then discuss the visual classification procedures. A preliminary analysis of the resulting sample of robust periodic variables is presented in the next section.

2.1. Sample selection

We start by selecting candidate variables from the public LINEAR database[1] using the following criteria:

[1]Available at https://astroweb.lanl.gov/lineardb/.

- Brightness limit: $14.5 < \langle m_{LINEAR} \rangle < 17$, where $\langle m_{LINEAR} \rangle$ is the median value of the white-light LINEAR magnitude.

- Likely variability: $\chi^2_{dof} > 3$, where χ^2 per degree of freedom is computed using the unweighted mean magnitude and photometric errors reported in the database.

- Variability amplitude: $\sigma > 0.1$ mag, where σ is the rms scatter (standard deviation) of recalibrated LINEAR magnitudes.

The majority of about 200,000 selected objects are found in the region bounded by $125° <$ R.A. $< 268°$ and $-13° <$ Dec $< 69°$ (corresponding to the North Galactic Cap scanned by SDSS). Additional ∼8,000 objects are found in the SDSS Stripe 82 region ($-50° <$ R.A. $< 60°$ and $|\mathrm{Dec}| < 1.266°$).

The selected objects contain both true variable objects and spurious candidates. We limit our classification to objects exhibiting periodic variability (light curves $m(t)$ that satisfy $m(t + P) = m(t)$, where P is the period; assuming no noise), and use phased light curves for visual inspection. Phased light curves are constructed by plotting $m(t)$ as function of phase

$$\phi = \frac{t}{P} - \mathrm{int}\left(\frac{t}{P}\right), \tag{1}$$

where the function int(x) returns the integer part of x. The likely periods were determined as described next.

2.2. Period finding methods

For each selected object, the three most likely periods were found using the Supersmoother algorithm (Friedman 1984). This non-parametric method smooths the light curve using a variable smoothing length and uses cross-validation method to pick a best-fit

period with the smallest phased light curve dispersion. The Supersmoother algorithm was extensively used by the MACHO survey and should be robust for a large variety of variable stars because it makes no explicit assumptions about the shape of the light curve.

During the classification it soon became apparent that the Supersmoother algorithm often had problems with finding the correct period; for eclipsing binaries in particular a large fraction of best-fit periods were twice as short as the true period (we will return to this discussion in §2.3.6). For this reason, we also included two additional algorithms for estimating periods: the Lomb-Scargle (LS) and Generalized Lomb-Scargle (GLS) parametric methods (Lomb 1976; Scargle 1982; Zechmeister & Kürster 2009). We used the code implemented in Gaia's Coordination Unit 7 pipeline (Eyer et al. 2013).

The LS method essentially fits a single sine wave to the light curve, and is capable of using heteroscedastic errors. It assumes that the true light curve mean is equal to the mean of sampled data points. In practice, the data often do not sample all the phases equally, the dataset may be small, or it may not extend over the whole duration of a cycle: the resulting error in the estimated light curve mean can cause problems such as aliasing. A simple remedy implemented in the GLS algorithm is to add a constant offset term to the single sinusoid model (Zechmeister & Kürster 2009).

We note that when the light curve shape significantly differs from a single sinusoid, the LS and GLS methods may easily fail. Possible remedies in such cases are to fit pre-defined light curve templates, or to use multiple harmonics in the Fourier expansion, which we have not considered here.

2.3. Visual classification methodology

Visual classification was performed on a per-object basis. There were three classification/validation runs; the first run pruned the list of candidates by more than a factor of 20, and the subsequent two runs further improved the sample purity and light curve classification precision. In the first run, 200,000 variable star candidates were divided roughly equally among eight human classifiers, using right ascension boundaries, and each classifier processed approximately 30,000 light curves. Overlaps of 2,500 light curves between the samples of the "adjacent" classifiers were used to verify classification consistency (which was assessed as described in §2.3.2 and 2.3.4).

2.3.1. Initial visual classification

The initial visual classification was performed using user interface shown in Figure 1. The automated classification tool displayed three phased light curves, folded with the periods found by the Supersmoother period finding algorithm, as well as five templates of folded (phased) light curves spanning predicted classes of variable objects. Classifiers answered three questions with fixed possible answers.

The first question was whether the displayed phased light curves have "reasonably small" dispersion around some imaginary smooth shape, following the Phase Dispersion Minimization idea of Stellingwerf (1978). There were four possible answers to this question (coded by numerical values in parentheses): "definitely no" (0), "probably no, but not sure" (1), "probably yes, but not sure" (2), "definitely yes" (3). Unless the answer to the first question is "definitely no", classifier proceeds to the second question related to the the light curve shape. Possible answers are: "does not look like any template" (0), "RR Lyr ab" (1), "RR Lyr c" (2), "single minimum on top of a flat light curve" (3), "two minima on top of

a flat light curve with some flat part" (4), "two minima without the flat light curve part" (5). The third question asks the user to choose which of the three folded light curves of the given object shows the smallest dispersion (the intention was to determine which of the three periods is the best). In addition, there was an option to add comments if necessary (e.g., about period aliasing, or any problems with the data), or to go back and repeat the classification for the object if an error was made. By design, *only the light curve shape was used in this first classification stage.*

After a brief training period, it takes about 5 seconds on average to answer all three questions, for a throughput of ∼700 objects per hour (about a week worth of full-time work per classifier, or about 2 Full-Time-Equivalent person months for the whole effort, assumming an unrealistic efficiency of 100%).

2.3.2. Tests of the initial classification uniformity and repeatability

In order to assess the uniformity and repeatability of the visual classification, a subsample of 8,044 light curves was classified by all eight classifiers. These objects were selected from the SDSS Stripe 82 region so that a comparison with an SDSS-based variable object catalog can also be performed (described further below).

For each light curve, we averaged the eight answers to question 1 (ranging from 0 for "definitely not variable" to 3 for "definitely variable") to obtain its $A1$ "grade". We also computed its standard deviation among the eight classifiers, σ_{A1}, to quantify dispersion in classification grades. Based on the morphology of the $A1$ distribution, we divided the sample into four subsamples using $A1$, as summarized in Table 1. The 317 light curves with $A1 > 1.8$ have the smallest $\sigma_{A1} = 0.15$: that is, most classifiers agree that these 3.9% objects are "definitely variable". The classification robustness of other light curves is lower,

as seen from the increased dispersion among the classifiers.

After sorting light curves by $A1$, two coauthors have re-inspected all 438 light curves with $A1 > 1.1$ (classes 1-3), as well as 1000 light curves from class 0 with highest $A1$ values. No spurious classifications were found in class 3. Objects in class 2 seem definitely variable, but many appear to have incorrect periods. Class 1 is similar to class 2, except for a larger fraction of unconvicing periodic cases. Therefore, there are between 317 and 438 definite periodic variables in this sample, depending on how conservative a selection cut is adopted, implying an upper limit for the sample contamination of 28%. Our main conclusion is that human classifiers are mutually consistent when their answer to the first classification question is 2 or 3, that is, when they are highly confident about detected variability.

2.3.3. Robust χ^2 Selection

The LINEAR light curve database contains two values of χ^2: the standard value and the so-called robust χ^2, $R\chi^2$, where 5% of the most outlying points are excluded from the computation (note that despite its name, the measured χ^2 does not follow the statistical χ^2 distribution expected for Gaussian photometric errors). The robust χ^2 might be efficient at minimizing the impact of photometric outliers, but at the same time it may decrease the sample completeness for light curves where variability is not always present (e.g., bursts and Algol-like light curves).

We have investigated whether $R\chi^2$ can be used to significantly prune the initial sample without a large decrease in the final sample completenesss (that is, whether $R\chi^2$-based selection could be used instead of visual pruning of the candidate sample). If $R\chi^2 > 3$ selection is adopted (instead of $\chi^2 > 3$), the size of the initial sample decreases from $\sim$200,000 to $\sim$80,000. Of all the light curves with $A1 > 1.2$ (classes 2 and 3 above, see

Table 1), 86% have $R\chi^2 > 3$. Therefore, the initial sample could be made smaller by a factor of 2.5, while losing 10-20% of true variables. This tradeoff reflects both the properties of faint variable stars and the behavior of LINEAR photometry.

About 14% of light curves with $A1 > 1.2$ (robust variables, as suggested by visual classification) have $R\chi^2 < 3$ (no strong evidence for variability). We have re-inspected these puzzling cases and found that they all are indeed real variables. In other words, visual classification is correct but $R\chi^2 < 3$ is too conservative a cut – these objects mostly have small amplitudes, short-duration peaks, or are faint (and thus photometric errors are large). Therefore, it should be possible to extract additional variable stars from the LINEAR database because our initial sample of 200,000 candidates had to satisfy $\chi^2 > 3$.

We have also re-inspected a random sample of light curves with $A1 < 1.2$ and $R\chi^2 > 3$, that is, light curves that show significant variability according to $R\chi^2$ but were not visually classified as periodic variables. About a half of these light curves show significant variability which appears aperiodic. A subset of a few hundred light curves with periods exceeding 1000 days and $R\chi^2 > 10$ seem consistent with being semi-regular variable asymptotic giant branch stars. Therefore, their rejection from the periodic light curve sample during visual classification is justified.

In summary, $R\chi^2$ parameter cannot be used to replace the visual classification step by automated selection without a significant drop in the sample completeness.

2.3.4. Comparison to the variable star sample from the SDSS Stripe 82

SDSS has obtained multiple observations (about 50 on average) in the 300 sq. deg. large so-called Stripe 82 region. These data were used to select 67,507 candidate variable

point sources[2] (for details see Ivezić et al. 2007a; Sesar et al. 2007; and references therein). There are many more candidate variables per unit sky area in the SDSS Stripe 82 catalog than in LINEAR sample because the former is much deeper ($g < 20.5$ vs. $r < 17.5$) and has a more inclusive cutoff for variability rms (0.05 vs. 0.1 mag). We have used this SDSS catalog to assess the reliability and completeness of candidate variables visually selected from LINEAR database.

Out of 8,044 LINEAR objects found the Stripe 82 region, 543 have positional matches within 2 arcsec to candidate SDSS variables that show periodic behavior. Of those, 301 have $A1 > 1.2$, that is, 83% of 363 robust LINEAR variables are confirmed by SDSS data. Therefore, there are 62 robust LINEAR variables that are not in SDSS variable sample, representing an 11% addition to the SDSS sample. These 62 LINEAR variables are dominated by detached eclipsing binaries with most SDSS observations falling along the flat part of light curve. An example is shown in Figure 2. Therefore, the implied purity of $A1 > 1.2$ LINEAR variables must be higher than 83%, and is consistent with 100% (that is, we did not find a single questionable case among these 62 variables). Figure 2 also demonstrates synergy between the SDSS and LINEAR datasets: while LINEAR provides much better time-resolved photometry for studying variable objects, SDSS provides very informative 5-band photometry.

About 45% of SDSS variables which are sufficiently bright to be in LINEAR sample are not selected from LINEAR database using criteria listed in §2.1 and $A1 > 1.2$ based on visual classification. About one third of those could be recovered by relaxing the $A1$ limit. The remaining two thirds ($\sim$30% of all SDSS variables) typically have sparse LINEAR data and/or small variability amplitudes, and thus were justifiably rejected in visual

[2]Light curves are publicly available from
http://www.astro.washington.edu/users/ivezic/sdss/catalogs/S82variables.html

classification. Therefore, relative to the SDSS subsample limited to a similar depth, the completeness of the LINEAR sample is in the range 55-70%, depending on the adopted $A1$ cut (most of the LINEAR incompleteness is due to larger adopted minimum rms variability, 0.1 mag vs. 0.05 mag).

Finally, out of 301 stars that are recognized as periodic variables by both SDSS and LINEAR, 184 have LINEAR and SDSS periods that agree within 2%. Additional 57 objects have periods aliased by a factor of 2 in either SDSS or LINEAR (for one third of those, the SDSS periods are larger); they include a large fraction of eclipsing binary systems with similar depths of primary and secondary minima.

2.3.5. Iterative improvements to visual classification

The first classification step, which pruned the initial list of 200,000 candidate variables by more than a factor of 20, was performed by eight different classifiers which must have introduced some non-uniformity in the resulting classification. In addition, the resulting sample contamination could be as high as 17%, as discussed in §2.3.2 and §2.3.4. To improve sample purity and classification uniformity, all the objects tagged as plausibly variable in the first round were re-examined in the second round by the first author. Only a few percent of objects had their classification changed as a result of this re-examination. Generally, no significant variations among the eights subsamples were noticed, in agreement with the conclusions from the previous section.

When the available source attributes (period, amplitude, and skewness of light curves, and optical and infrared colors) were analyzed for the sample obtained in the second classification round, it became apparent that different types of variable star cluster in different regions of the multi-dimensional attribute space. Using selection boundaries based

on color, period, amplitude and light curve skewness listed in Table 2, and discussed in more detail in §4.1, an additional sample of about 750 objects was selected from the initial candidate sample of 200,000 objects. That is, about 10% more potential variables than extracted in the first classification round were selected for further inspection.

Visual inspection of these 750 candidates (by the first author) in the third classification round revealed that only about 10% represented convincing cases of periodic variability. They were added to the initial list to produce the final sample of 7,194 visually selected and classified periodic variables. Among those, 6,876 light curves (96%) have been assigned a definite type, while the remainder are classified as "Other" (i.e. definitely variable, but the exact variability type could not be reliably determined).

The six main light curve types are listed in Table 2, and a few supplemental ones in Table 3, and discussed in more detail in the next Section. Hereafter, we refer to this sample as "visually confirmed sample of periodic LINEAR variables", or simply "PLV" sample. The resulting catalog is made publicly available[3].

Table 3 quantitatively summarizes the results of visual classification. The first column "translates" our numerical codes used during visual classification to the adopted variability types. We hypothesize that the class "3" ("a single minimum on top of a flat light curve) mostly consists of EA type binaries (Algols) for which our data did not show a discernible secondary minimum (i.e. either too shallow to be detected, or too similar in depth to the primary minimum, recall §2.3). For that class of objects correct periods could be twice longer than listed in the catalog. The light curves classified as "5" include two types of eclipsing binaries: EB (or β Lyrae) and EW (W Ursae Majoris), which are grouped together because they are hard to distinguish using only LINEAR light curves. Classes "6" (SX

[3]Available from XXX add site here

Phoenicis and δ Scuti candidates) and "7" (long-period variables; defined here as variables with periods longer than 50 days, and as semi-regular variables).

2.3.6. Comparison of period finding methods

As we already indicated earlier, period finding algorithms often had problems with choosing the correct period. For example, for eclipsing binaries a large fraction of best-fit periods were twice as short as the true period. In this particular case, such behavior is easy to understand: primary and secondary minima are often of similar depth and are therefore often misidentified as the same feature in the phased light curve. This error, however, is not systematic: not all of the objects with similar depths of minima have periods that are too short by a factor of two.

Given the final sample of 6,876 reliably classified light curves, we tested period finding methods for each of the six main light curve types separately. Our results are summarized in Figure 3. We left out the "single minima on top of a flat light curve" class out of the analysis, as the sample is small (20 objects) and the correct period for those objects could not be identified with certainty. We speculate that those objects could correspond to eclipsing binaries of EA (Algol) type with similar depths of minima, but with periods that are too short by a factor of two. Another explanation would be that secondary minima for these objects are too shallow to be detected in LINEAR data.

Our results show that the Lomb-Scargle and generalized Lomb-Scargle methods typically outperform the Supersmoother algorithm for all variability types. For c type RR Lyrae, long-period variables, and SX Phe/δ Scu type light curves, Supersmoother has a much larger fraction of overestimated periods (typically by a factor of two, but sometimes more) than the other two methods. In addition, when the period is approximately correct,

the uncertainty is typically larger for Supersmoother values (that is, the width of the central peak in histograms shown in Figure 3 is larger).

The performance of the period finding algorithms for eclipsing binaries is rather different: while the Lomb-Scargle and generalized Lomb-Scargle methods produce narrower histogram peaks than Supersmoother, their periods are consistently (at >90% level) too short by a factor of two! After an overall correction of periods for eclipsing binaries by this factor, the Lomb-Scargle and generalized Lomb-Scargle methods display better performance than Supersmoother.

The reason for this consistent bias in period estimation by the Lomb-Scargle and generalized Lomb-Scargle methods is their fundamental assumption that the shape of the underlying light curve can be described by a single sinusoid. A remedy is to fit a Fourier series with many terms (but more computationally expensive). As illustrated in Figure 4, a Fourier series model with six terms correctly recognizes two minima in the light curve of an eclipsing binary star. For additional discussion, please see Hoffman et al. (2009).

During the visual inspection it was relatively easy, albeit time consuming, to apply this correction factor to the periods. In a fully automated classification scheme that has only single band light curves and no color information this might be more difficult since values of period, amplitude and skewness are in large part similar for c type RR Lyrae and EB and EW binaries. Addition of appropriate color information (e.g. $g - i$) easily breaks this degeneracy (see §3.1 and §3.2). Ultimately, the performance of period finding algorithms based on a single sinusoid can be significantly improved by including more Fourier terms.

3. Preliminary Analysis of Periodic LINEAR Variables

The remainder of our analysis is performed using the public version of the PLV catalog. We show in this section that the distribution of selected periodic variables displays distinctive features in the multi-dimensional attribute space spanned by the light-curve parameters (period, amplitude, shape) and optical/infrared colors. This behavior enables robust and efficient classification of objects into various classes of variable population. These features are not seen for the full sample of 200,000 candidate variable objects, and thus strongly suggest that visual classification successfully extracted true variables.

We first discuss the distribution of classified variables in diagrams constructed with the three light curve parameters, and then investigate the correlation of light curve parameters with optical and infrared colors. We quantify a strong correlation between the period and optical color for contact eclipsing binaries, provide evidence that the sample contains a large number, compared to the known objects, of likely Population II field SX Phe stars, and demonstrate that the infrared colors from the WISE survey provide further support that long-period variables are correctly classified.

3.1. Analysis of Light Curve Properties

The light-curve amplitude is estimated non-parametrically from the cumulative magnitude distribution as the range between the 5% and 95% points. The light-curve skewness is computed as described in Sesar et al. (2007). Therefore, light curves are quantitatively described using three parameters: period, amplitude and skewness. This choice is of course not unique. For example, in addition to, or instead of, amplitude, other estimators of the width of the observed magnitude distribution could be used, such as standard deviation (which is not robust to outliers) and the inter-quartile range (which

is not sensitive to single minima in otherwise flat light curves). Similarly, the light-curve shape could be further quantified using higher moments (such as kurtosis, but they quickly become very noisy), Fourier coefficients (which help greatly to classify eclipsing binary subtypes, see Pojmański 2002), or even non-parametrically using the principal component analysis (e.g. Deb & Singh 2009). In this preliminary analysis, we find that even our simple approach based on period, amplitude and skewness provides informative description of the light curve behavior. Nevertheless, exploring these other options would be a worthwhile analysis to undertake.

The distribution of variables in the period–amplitude–skewness space is illustrated separately for each of the six main variability classes in Figure 5. The period distribution of the PLV sample is multi-modal, as further quantified in Figure 6. Even the period alone enables remarkable, although not perfect, classification of periodic variables: SX Phe/δ Scu candidates clearly stand out ($P < 0.1$ day), and ab type and c type RR Lyrae are fairly well separated by $P = 0.4$ days. Nevertheless, eclipsing binaries overlap with the period range of RR Lyrae stars (especially EW/EB type eclipsing binaries and c type RR Lyrae). In addition, the light-curve amplitude distributions are similar for c type RR Lyrae and EB/EW eclipsing binaries. This degeneracy can be readily lifted using the light curve skewness (and object color, see below). Indeed, all six classes can be readily defined when all three light-curve parameters are considered (e.g. EB/EW class has much larger skewness than c type RR Lyrae; compare the symbol color in the top right and bottom left panels in Figure 5). In other words, the visual classification of light curves in essence reflects the distribution of these three parameters (and also of the light curve smoothness). We analyze the performance of automated classification methods based on this behavior in §4.

It is possible to further separate ab type RR Lyrae into Oosterhoff type I and Oosterhoff type II stars (Sesar et al. 2010), as shown in the top right inset in the "RRAB" panel of

Figure 5 (note also the strong correlation between the amplitude, skewness and period for ab RR Lyrae). Average periods of Oosterhoff type I and type II ab RR Lyrae for the PLV sample are $\langle P_{ab}^{I} \rangle = 0.56$ days and $\langle P_{ab}^{II} \rangle = 0.65$ days. This result is in good agreement with Oosterhoff's conclusion that period of RR Lyrae ab in Oosterhoff type I clusters is 0.1 day shorter than that of those in Oosterhoff type II clusters (Oosterhoff 1944). For a more detailed analysis of the Oosterhoff's dichotomy for field RR Lyrae stars based on this sample, see Sesar et al. (2013).

3.2. Correlations between Colors and Light Curve Properties

The addition of the color information to light-curve parameters significantly improves the separation of visually defined classes and ultimately enables better performance of automated classification methods. For a detailed discussion of the distribution of stars in various color-color diagrams constructed with SDSS and 2MASS photometry, see Covey et al. (2007), and references therein. The most useful SDSS-2MASS colors are $u-g$, $g-r$ (or $g-i$), $i-K$ and $J-K$, which are sensitive to various combinations of effective temperature, metallicity, and surface gravity. Therefore, the minimal useful dimensionality (the number of measured attributes that are independent for at least some subsamples) of this dataset is at least five (the three light curve attributes and at least two color attributes).

Figure 7 demonstrates that the addition of just one color to the period, here the SDSS $g-i$ color which is a good measure of the effective temperature (Ivezić et al. 2008b), helps to clearly separate c type RR Lyrae from EB/EW binaries. A more detailed illustration of the correlations between the $g-i$ color and light curve properties is shown in Figure 8. Note in particular how EA and EB/EW are well separated in this diagram. The EB/EW subsample displays a good correlation between the period and color, discussed in more detail in §3.3.

3.2.1. The $g-r$ vs. $u-g$ diagram

In addition to the three-dimensional $g-i$ color–period–amplitude projection of the full multi-dimensional attribute space discussed above, the three-dimensional projection spanned by the SDSS $u-g$ and $g-r$ colors and light curve skewness is also rich in content. The $u-g$ vs. $g-r$ diagram is one of the most informative SDSS color-color diagrams; it clearly distinguishes quasars from stars, main sequence stars from binary stars and white dwarfs, and it contains information about effective temperature and even metallicity for blue main sequence stars (Smolčić et al. 2004; Ivezić et al. 2007a; Ivezić et al. 2008b).

The distribution of variables in the $u-g$ vs. $g-r$ vs. skewness space is shown separately for each of the six main variability classes in Figure 9. As known from previous work based on SDSS data, RR Lyrae color distribution is localized to the region populated by spectral types A and early F (Sesar et al. 2010, and references therein). Only about 1-2% of light curves classified as RR Lyrae fall outside the expected small color regions discernible in Figure 9.

Based on the $u-g$ vs. $g-r$ color-color diagram and the skewness distributions, we identified **Lovro, how many?** suspected mis-classifications between c type RR Lyrae and EB/EW eclipsing binaries and visually re-inspected their light curves. We found that **Lovro, how many?** light curves were indeed likely mis-classifications and were subsequently revised. The cross-contamination of these two subsamples is easy to understand; a light curve of an eclipsing binary with similar depths of minima can easily be misidentified as a nearly symmetric (sinusoidal) c type RR Lyrae light curve. This ambiguity is particularly problematic in case of faint objects, or objects with sparsely sampled light curves. We note that the color distribution of c type RR Lyrae has a well defined red edge – it is thus easy to prevent the contamination of EB/EW subsample by c type RR Lyrae but the converse is not true because EB/EW stars can have colors as blue

as RR Lyrae colors.

We have also explored a few other three-dimensional projections of the seven-dimensional attribute space (there are 35 possible independent attribute combinations) and did not find diagrams as revealing as the $g-i$ color vs. period vs. amplitude diagram and the $g-r$ vs. $u-g$ vs. skewness diagram. A noteworthy color is the 2MASS $J-K$ color which is capable of separating main sequence stars from quasars and late-type giants (including the long-period asymptotic giant branch stars); for main sequence stars the 2MASS $J-K$ color and the SDSS $g-i$ color are highly correlated (both are by and large driven by the effective temperature), while for those other populations the measured $J-K$ color is redder than the $J-K$ color of main sequence stars of the same $g-i$ color (for more details, see Covey et al. 2007).

3.3. Period-color correlation for contact eclipsing binaries

The distribution of EB (β Lyrae) and EW (W Ursae Majoris) eclipsing binary stars is remarkably well outlined in the period vs. $g-i$ color diagram (see the bottom left panel in Figure 8, and a zoomed-in version in Figure 10). Since the sample selection is primarily driven by the light-curve shapes, and substantial selection effects in the $g-i$ color and period in the relevant ranges are not expected, the discernible strong correlation is likely of astrophysical origin. A similar result was reported for a much smaller sample of contact binary systems by Eggen (1967), (see also Rucinski & Duerbeck 1997, and references therein). The range of observed $g-i$ colors correspond to spectral types F5 ($g-i=0.3$) to K4 ($g-i=1.4$) (see Table 3 in Covey et al. 2007). Rucinski & Duerbeck (1997) used Hipparcos distance estimates for 40 W UMa stars to derive a relationship between the absolute V band magnitude, period and color. According to their results, our sample includes stars with $1 < M_V < 6$.

We compute the median $log(P)$ in bins of the $g-i$ color for stars with $0.2 < g-i < 1.6$ and $-0.4 < log(P) < -0.67$, and fit a parabola to the resulting points,

$$log(P/day) = 0.05\,(g-i)^2 - 0.24\,(g-i) - 0.37. \quad (2)$$

Due to the large sample size, the random errors for the fitted data points are sufficiently small to rule out a linear relationship. This best-fit relation implies that the median period for EB/EW eclipsing binaries increases from 5.9 hours to 8.8 hours as the color-based spectral type varies from K4 to F5. An alternative form based on the Johnson $B-V$ color, derived using using transformations between the SDSS and Johnson systems from Ivezić et al. (2007b), is

$$log(P/day) = 0.038\,(B-V)^2 - 0.29\,(B-V) - 0.33, \quad (3)$$

and valid in the range $0.3 < B-V < 1.1$. This relation agrees well with a similar relation obtained by Rucinski (1997) for ∼400 W UMa stars observed by the OGLE project in Baade's window (note that we fit the median relation and Rucinski obtained the short-period limit as a function of color; the two sequences are offset by about 0.1-0.15 mag at a given period).

These findings are related to the fact that the period distribution for contact binary star systems appears to have a well-defined lower limit at 0.22 days (Rucinski 1992). More recent data show that this limit may be a bit smaller (∼0.20 days, see Dimitrov & Kjurkchieva 2010; Davenport et al. 2013), but the existance of a well-defined boundary is not disputed. Indeed, the falloff of the distribution at small periods for M dwarf systems (see Figure 6 in Becker et al. 2011) is very similar to the falloff for EB/EW systems in our Figure 6). If we extrapolate our best-fit to $g-i = 2.0$ corresponding to the spectral type M0, we obtain a period of 0.22 days in good agreement with other studies. Since no consensus about the origin of the short-period boundary for contact binaries is reached yet, the improvement in observational constraints enabled by LINEAR data will be valuable for

future studies of stellar evolution.

3.4. Candidate SX Phe stars

The PLV sample presented here includes a class of 112 blue stars ($-0.3 < g - i < 0.2$, bluer than thick disk and halo turn-off stars and corresponding to $-0.2 < B - V < 0.3$ using transformations between the SDSS and Johnson systems from Ivezić et al. 2007b), with very short periods (1–2.5 hours), and with asymmetric light curves (see bottom right panel in Figures 5 and 8). These stars can be identified as a mixture of δ Scuti and SX Phoenicis stars (e.g. see Figure 8 in Eyer & Mowlavi 2007). Both types of stars are usually considered as variable counterparts of blue straggler stars (main sequence stars in open or globular clusters that appear younger than they should be given the cluster age), with δ Scu subsample belonging to Population I disk stars and SX Phe subsample to Population II halo stars (see e.g. Jeon et al. 2004).

In a recent study based on the largest catalog of SX Phe stars assembled to date (about 250 stars identified in globular clusters), Cohen & Sarajedini (2012) demonstrate that this population appears to occupy a narrow region at the bottom of the instability strip with $1.5 < M_V < 3.5$, and are all likely radial model pulsators. Given the apparent magnitude limits of our sample, the implied distances span the range 2–10 kpc, that is, many disk scale heights away, and thus SX Phe probably dominate because they are Population II (halo) objects. We note that the $B - V$ color distribution of our sample extends to bluer colors than the range displayed by the Cohen & Sarajedini (2012) sample (their range is approximately $0.1 < B - V < 0.4$, corresponding to $0.0 < g - i < 0.3$; about 20% of our candidates have $g - i < -0.1$).

A much higher fraction of SX Phe stars than δ Scu stars in this sample is supported

by SDSS spectra that are available for 34 stars in the candidate sample. All the spectra appear very uniform and characteristic for A stars; an example is shown in Figure 11. Using the default SDSS metallicity and radial velocity estimates (see Figure 12), we find that the sample is dominated by stars with [Fe/H]< -1, low metallicities characteristic of halo stars, with a large velocity dispersion (134 km/s) that is also consistent with presumed halo population (for a review of recent observational constraints on the differences between the metallicity and kinematics distributions of disk and halo stars, see e.g. Ivezić, Beers & Jurić 2012).

Assuming that our conclusion about the sample being dominated by halo stars is correct, these 112 candidates likely represent a major addition to the total number of known SX Phe stars (according to Cohen & Sarajedini 2012, fewer than 300 SX Phe stars are known). Our sample would also increase the number of known *field* SX Phe stars by as much as a factor of six (according to Rodríguez et al. 2001, there are only 17 known field SX Phoenicis known). This large increase in the sample size of field SX Phe stars is due to the fact that the LINEAR dataset is among the first ones to explore sufficiently faint flux levels, over a large sky area, and with appropriate cadence. We are currently undertaking photometric and spectroscopic followup efforts to better characterize this sample.

3.5. Candidate AGB stars and WISE color distribution

The PLV sample includes 77 light curves classified as "long-period variables", defined here as variables with periods longer than 50 days, and as semi-regular variables. These stars are expected to be dominated by asymptotic giant branch (AGB) stars which often display infrared excess emission due to their dusty envelopes (see e.g. Ivezić & Elitzur 1995, and references therein). The correctness of their classification can thus be tested by inspecting their infrared colors.

The best available infrared sky survey was obtained by the recent Wide-field Infrared Survey Explorer (WISE, launched in 2010); its all-sky catalog includes about 560 million objects (Wright et al. 2012). WISE mapped the sky at 3.4, 4.6, 12, and 22 μm with 5-σ point source sensitivities better than 0.08, 0.11, 1 and 6 mJy (corresponding to Vega-based magnitudes 16.5, 15.5, 11.2 and 7.9, respectively) in unconfused regions on the Ecliptic. The astrometric precision for high signal-to-noise sources is better than $0\farcs15$. WISE is photometrically calibrated to Vega system and thus objects with infrared excess should have colors greater than zero (not accounting for the measurement noise).

We have positionally matched the PLV and WISE catalogs with a matching radius of XXX arcsec and obtained XXX WISE matches for objects listed in the PLV catalog. **Lovro, please add these numbers; also, how many LPVs are in that figure?**. Our analysis of this sample is shown in Figure 13. The distribution of WISE colors for objects classified as "long-period variables" is consistent with the majority of them being genuine AGB stars (Tu & Wang 2012; Tisserand 2012). Indeed, the brightest and most famous carbon-rich AGB star, CW Leo (IRC+10216) is recovered in our sample (LINEAR ID=17154286; P=632.511 days based on 475 LINEAR measurements; see also §3.6).

The top panel in Figure 13 shows the period-color relation for long-period variables. Although there is some correlation between the quantities, the scatter is substantial. The observed scatter in $logP$ at a fixed color of about 0.2 dex is in good agreement with earlier work (e.g. see Whitelock et al. 2006, and references therein).

There are nine objects with light curves classified as "Other" that show infrared colors consistent with quasars ($W1 - W2 > 0.7$, see e.g. Yun et al. 2012). **Lovro, can you send me ra/dec list for these 9 objects? I want to check whether they have SDSS spectra; it would be great if you could also make a quick'n dirty g-r vs. u-g diagram for them**. In addition, there are 14 objects with $W2 - W3 > 2.0$, implying

strong infrared excess that is likely inconsistent with AGB stars (Nikutta et al., in prep.), but also with blue $W1 - W2 < 0.5$ colors inconsistent with quasars. A few but not all of them could be chance positional coincidences with background quasars which would mostly affect $W3$ and $W4$ measurements (based on a quasar surface density of several hundred per square degree and a matching radius of 3 arcsec).

3.5.1. Noteworthy objects

There are five interesting sources that deserve direct mention by name. There is one case of a likely supernova (LINEAR ID=7682813) which increased in brightness by 0.8 mag over about 10 days, and then gradually returned to the initial brightness over about 90 days. The corresponding SDSS image clearly shows a positionally coincident blue emission-line galaxy at a redshift of 0.028. The object with LINEAR ID=17655724 steadily increased in brightness by 0.5 mag over about 5 years. If this trend continues, in 400 years it would outshine the Sun; nevertheless, this is unlikely because its SDSS spectrum confirms that this object is a quasar at a redshift of 0.531 (we note that this variability behavior is a bit unusual when compared to typical quasar variability properties, see e.g. MacLeod et al. 2012). Given its light curve that shows large variations (e.g. a decrease in brightness of $\sim$3 mag over $\sim$600 days), and its WISE colors, the object with LINEAR ID=3766947 is a good candidate for an R Coronae Borealis star, a supergiant carbon-rich star with episodic mass loss (Tisserand 2012). The object LINEAR ID=7455728 is classified as an Algol (EA); it displays a flat-bottom primary minimum and frequent faint outliers. While these outliers could be due to the effects of a nearby (6 arcsec) star, it is not obvious what is the origin of its very red WISE colors ($W2 - W3 = 2.58$). Possibly the most curious case is an optically resolved (see the next section) and spectroscopically confirmed quasar at a redshift of 0.152, with quasar-like WISE colors, but with an apparently periodic light curve (LINEAR

ID=23417507, $P = 604$ d, amplitude ~ 0.4 mag). A periodic quasar light curve might have interesting astrophysical implications and searches for such objects have been reported in the literature. In the largest such search, MacLeod et al. (2010) found 66 candidates in a sample of ∼9,000 quasars from the SDSS Stripe 82 region with spectroscopic confirmation and SDSS light curves. They declared them all as unconvincing cases of periodicity because their best-fit periods are roughly the same as the span of observations – that is, only a single putative oscillation was detected. In contrast, our object displays three full oscillations in the LINEAR light curve and may be worthy of a followup study. Finally, we note that the ratio of EB/EW and RR Lyr stars in this subsample is the same as in the full PLV.

```
Lovro, can we add a mosaic with these LCs (just LCs,
no phased light curves, except for Algol that would only
have the phased curve, and for periodic quasar that would
have both - 6 panels; since we want details, perhaps 2x3 panels):
# supernova: 7682813
# steadily brightening quasar: 17655724
# R CorB: 3766947
# Algol: 7455728
# periodic quasar: 23417507

FYI: the period quasar has a weird QSO spectrum; I contacted a
few experts to get a second opinion. Early responses are enthusiastic!
This object also has the Catalina Sky Survey data which by itself
gives a period of 534 days and rules out systematics in LINEAR data!
We should run both datasets simultaneously and get an improved
period (though not for this paper!).
Joj kak sam bedast, pa to si ti vec napravio u fig. 15! Jesi li
```

mozda izvrtio period finder za kombinirane podatke? Btw, jesi li koristio kakvu korekciju zbog razlicitih LINEAR i CSS passband-ova?

3.6. Optically-resolved periodically-variable objects

Lovro, u zadnjem katalogu kojeg ja imam, sa 7195 objekata, ima 19, a ne 18, sa objtype = 6. Takodjer, ima ih 116 sa objtype = -99; znas li otkuda taj -99?.

Among the 7194 objects listed in the PLV catalog, 18 are optically resolved in the SDSS imaging data (and additional 116 objects have unreliable size measurements). Their SDSS image stamps are shown in Figure 14. As evident, eight objects are clearly galaxies and their variability may be at least to some extent due to photometric measurement difficulties when using LINEAR images. Nevertheless, three objects (LINEAR IDs=7682813, 8440571, 9183803) show spectroscopic evidence for AGN activity and their variability may be real (the last object is also listed in the X-ray ROSAT catalog).

The light curves for the ten objects that do not appear as well-resolved galaxies are shown in Figure 15. Object in the middle right panel (LINEAR ID=22993473, the fourth object in the third row in Figure 14) is beyond doubt a barely resolved binary system, with a light curve classified as EW/EB. A few sources show color gradients in their SDSS point spread function (including a known RR Lyrae star V368 Her, shown in the top left panel); such gradients can be a sign of their binary nature, or possibly of fast changes in the point spread function that led to their misclassification as resolved objects by the SDSS image processing pipeline (Lupton et al. 2002). The objects shown in the bottom row in Figure 15 have already been discussed: carbon-rich AGB star CW Leo and a quasar with nearly-periodic light curve. For the latter, we have added data from the Catalina Sky

Survey; during the overlap with the LINEAR data, the two light curves are consistent. These additional data provide further support for quasi-periodic light variations displayed by this quasar.

4. Comparison of Visual Classification and Machine Learning Classification Methods

We have demonstrated in the preceding section that the distribution of visually-selected periodic variables displays distinctive features in the multi-dimensional attribute space spanned by the light-curve parameters (period, amplitude, skewness) and optical/infrared colors. In this section we explore to what extent can this behavior enable robust and efficient automated classification of objects into various classes of variable population. We consider two classification methods.

First, we perform the so-called supervised classification where a training sample is used to define selection boundaries. We choose to use simple rectangular (linear) boundaries in the four-dimensional space spanned by the three light curve parameters and the SDSS $g-i$ color, which serves as a proxy for the effective temperature. The optimization of the classification boundaries is performed manually using the training (PLV) sample. **Lovro, kako si ono to radio? Koju velicinu si optimizirao (problem je uvijek u tradeoffu izmedju kompletnosti i kontaminacije)** This selection method is then applied to the sample of 200,000 potential variables that were subjected to visual classification. The main goal of this method is to quantify whether visual classification could be improved, or perhaps entirely bypassed.

We also explore the performance of an unsupervised classification algorithm that attempts to recognize existing variability classes in the PLV catalog using their clustering

in the multi-dimensional attribute space. While this algorithm represents an improvement to the above simple classification based on rectangular boundaries, it can only produce meaningful results if applied to a sample of confirmed variables.

4.1. Classification based on supervised attribute cuts

We define selection boundaries using simple, rectangular cuts in the four-dimensional attribute space (period, amplitude, skewness, $g - i$ color). The adopted boundaries are listed in Table 2. We limit quantitative analysis of the performance of this classification scheme to ab and c type RR Lyrae, EB/EW eclipsing binaries and SX Phoenicis/δ Scuti candidates. We do not include classes not exceeding 1% of the full sample, nor Algols (EA eclipsing binaries) and objects classified as "Other". We do not include Algols because their distribution does not have well-defined boundaries (not too surprising since in the case of detached binaries we have an ensemble of paired objects with presumably little or no common physical characteristics). An analogous diversity is expected among long-period variables which include both Miras and semi-regular variables, and possibly other classes of variable star. Indeed, even the definition of Mira stars suffers from quantitative ambiguity ("red long-period variables with visual amplitudes exceeding 2.5 mag"), although it has been shown that they are actually fundamental mode pulsators — a physical characteristic that differentiates them from other long period variables (e.g. Wood & Sebo 1996; Spano et al. 2011).

In order to maintain analysis uniformity, we use best-fit periods found by the Lomb-Scargle method. Objects with unreliably measured SDSS colors, and Lomb-Scargle periods close **Lovro, how close is close?** to one day and half a day were excluded from the analysis. The performance of this supervised classification is statistically compared to our visual classification results in Figure 16. We have visually re-examined all 3,270 light

curves with differing visual and automated classifications.

The automated method selected 74% of PLV objects from the four analyzed types. This result does not imply a 26% contamination in the PLV catalog but rather an incompleteness of the automated selection method; the majority of missing objects had unreliable SDSS colors, or the best-fit Lomb-Scargle period was too close to one day, or half a day. **Lovro, how many in each group? 26% sounds like too many for these two causes...** This selection fraction varies little among the four types (see the bottom row in Figure 16).

The automated selection method selected 835 objects that are not included in the PLV catalog (a 12% addition, varying from 4% for c type RR Lyrae to 23% for EB/EW). Of the 246 objects selected by cuts corresponding to ab type RR Lyrae, but not found in PLV, the majority are located very close to the red cutoff for the $g - i$ color. Approximately 15% of these 246 objects have light curves hinting at ab type RR Lyrae, but not of sufficient quality to enable reliable visual confirmation. Therefore, at most about 40 ab type RR Lyrae included in the initial sample of 200,000 candidates are missing from the PLV catalog (1.4% effect). In case of c type RR Lyrae candidates selected by cuts, 44 objects not found in PLV are uniformly distributed throughout the selection volume. About 30% of these objects have light curves that might be classified as c type RR Lyrae, though not reliably. Similar behavior is displayed in EB/EW case, with only about 10% of 545 objects not found in PLV potentially classifiable as reliably periodic. Therefore, the PLV catalog is only slightly incomplete relative to the initial sample of 200,000 candidates (by about 1-2% at most).

The automated classification is correct for a high fraction of objects, and an even higher fraction when only objects found in PLV are considered: 88% for ab type RR Lyrae (97% for the PLV subset), 87% for c type RR Lyrae (91%), 74% for EB/EW (97%) and 100% for SX Phe.

In summary, this analysis provides further support that the PLV catalog is highly complete, has exceedingly low contamination, and a high rate of correct light curve classification.

4.2. Unsupervised classification based on a Gaussian Mixture Model

The strong clustering of objects, visually classified in six different types using their light curves, in the multi-dimensional attribute space suggests that an automated unsupervised classification scheme might be at least as successful as visual classification (and definitely easier!). To investigate this possibility, we used a machine learning algorithm based on a Gaussian mixture model to describe the observed distribution of objects. We note that the only attribute describing light curve shape is skewness. More sophisticated schemes, such as those based on best-fit parameters for a multi-harmonic Fourier series fit to light curve, are also possible (e.g., Debosscher et al. 2007; Richards et al. 2011; and references therein).

The Gaussian Mixture model (GMM) describes the density of data points using a sum of multi-variate Gaussians. Statistically significant clusters of points are assigned a Gaussian, and in case of complex cluster morphology, multiple Gaussians. This clustering method does not require a training sample and thus belongs to the class of unsupervised classification (clustering) methods. The number of required clusters and their best-fit parameters are typically obtained using the Expectation Maximization method (Dempster et al. 1977). We used a GMM implementation from *astroML*, a set of publicly available[4] data mining and machine learning tools implemented in *python*. Figures 17 and 18 show the GMM results for two cases.

The top panel in Figure 17 shows a 12-component Gaussian mixture model using only

[4]See http://astroml.github.com

two most discriminative data attributes, the $g-i$ color and $\log(P)$. Out of the 12 clusters, five are significant, while the rest seem to describe the background. These five clusters can be identified with RR Lyrae stars (ab type are described by two Gaussian components and c type by one), eclipsing binaries, and δ Scu/SX Phe class. Therefore, no new classes were revealed by this restricted automated analysis.

In another instance of GMM analysis, the clustering attributes included four photometric colors based on SDSS and 2MASS measurements ($u-g$, $g-i$, $i-K$, $J-K$) and three parameters determined from the LINEAR light curve data ($\log(P)$, amplitude, and light curve skewness). Fitting a 20-component Gaussian mixture to this seven-dimensional dataset yields the clusters shown in the bottom panels of Figure 17. The clusters derived from all seven features are remarkably similar to the clusters derived from just two features: this shows that the additional data adds very little new information (equivalently, this shows that the seven attributes are strongly correlated). Nevertheless, note that the fifth significant cluster is not δ Scu/SX Phe stars, but a horizontal feature that is likely due to unreliable periods. **ZI: this needs to be updated after I rerun GMM for Lovro's final sample!** Figure 18 shows the locations of these clusters in the space of other attributes. The means and widths of the distribution of points assigned to each cluster for the 7-attribute clustering are shown in Table 4.

As evident from the inspection of Figures 17 and 18, and reinforced by statistics listed in Table 4, the most discriminative attribute remains the light-curve period. Clusters #2 and #3, which have very similar period distributions, are separated by the $g-i$ and $i-K$ colors. The problematic cluster #5 has a very narrow distribution of periods, and also unusually large amplitudes. **ZI: this needs to be revisted too!** The automated discovery of these erroneous periods is a good example of the benefits of machine learning techniques.

5. Discussion and Conclusions

implement this SX Phe/δ Scu candidates are found in the region of the $u-g$–$g-r$ color diagram populated by RR Lyrae stars, with a number ratio of 1:40. Therefore they do not represent a major contaminant of RR Lyrae samples; our results confirm early estimates of the upper limit of their contamination fraction of 10% (Ivezić et al. 2000).

The preliminary work described in §3 is by no means a complete analysis of our sample. To point out but a single example, detailed analysis of light curves for eclipsing binaries using more sophisticated methods such as Fourier analysis, or full physical model fitting (Rucinski 1992; Devor 2004; Prša & Zwitter 2005), is capable of providing valuable further insight into the physics of such stellar systems.

Compared to e.g. 10,000 eclipsing binaries in the Bulge fields discovered by OGLE II and analyzed by Devor (2004), or to ∼2,000 eclipsing binaries discovered in the Kepler survey data (Prša & Zwitter 2005), our sample of ∼2,700 stars is in the sample ballpark. Its comparative advantage is in the large sky area which potentially enables studies of the variation of eclipsing binary star properties with location in the Galaxy (and by extension, with metallicity and possible other parameters).

This sample will be valuable for comparison to Gaia, for example, to search for period evolution (ref. to SDSS work, Jim or Becker?)

The period distribution is generally in agreement with previous work, e.g. for eclipsing binaries (Giuricin et al. 1983; Devor 2004; Prša & Zwitter 2005).

Where is text about the heartbeat star?

untouched: This work describes creation of a visually confirmed catalog of periodic variable stars created from data acquired by LINEAR asteroid survey, the "PLV" catalog. Catalog consists of 7,194 variable stars, with at least 6,876 stars that are periodic variables

(i.e. less than 5% are not periodic). Combined with large area coverage ($\approx$ 10,000 deg^2) and a flux limit several magnitudes fainter than most other wide angle surveys ($14 < r < 17$), our sample can be useful for a wide variety of research topics such as studies of Galactic halo structure and physics of pulsating stars and eclipsing binaries.

Folded light curves of all the objects in the catalog were visually inspected. Aditional attributes (SDSS, 2MASS and WISE colors) were used to better characterize each of the objects and thus improve classification purity. Furthermore, we compared our results to GCVS and VSX catalogs in order to ascertain effectiveness of our method. Moreover, we compared our database to DR13 (cf. §4), a similar catalog but consisting only of ab type RR Lyrae stars. Our database compares favourably to all of these catalogs, and we are able to claim that indeed our catalog has very high purity, exceeding 9?%. Based on a comparison with the SDSS Stripe 82 variable stars and DR13, we estimate that the completeness of the PLV catalog is 55–70%; most of the LINEAR incompleteness is due to larger adopted minimum rms variability, 0.1 mag vs. 0.05 mag.

In order to estimate the effectiveness of visual classification, we compared it to a supervised classification method based on simple cuts in period, amplitude, skewness and the $g - i$ color, as well as to an unsupervised classification method based on a Gaussian mixture model. In the former case we conclude that cuts-based selection can achieve high purity, but due to imperfections in the data the completeness of this simple approach will suffer (cf. §5.1, Fig. 16). **Zeljko please add some details about GMM**

move earlier? The PLV catalog is dominated by RR Lyrae (3,913 or 54%) and eclipsing binaries (2,762 or 38%). We also found 112 candidate SX Phoenicis/δ Scuti variables and 77 red variables with long regular or semi-regular periods (Mirae, LPV, SR). Table 3 describes the proportions in detail. Furthermore, we found a candidate "heartbeat" star wtih a period of 562.4 days (Fig. 19).

An exciting result of our effort is the discovery of 112 SX Phoenicis/δ Scuti candidates. It is not possible to differentiate the two on the basis of light curve attributes and color. However, our preliminary analysis based on SDSS spectra and radial velocities (available for 30 of these objects) shows that they are consistent with the Population II objects. Therefore we assume that the sample is dominated by SX Phoenicis (see §3.2, Figure 12). Until now SX Phoenicis have been found mostly in Galactic globular clusters ($\approx$ 250 objects in total) and only 17 field SX Phoenicis are currently known (cf. §3.2). Therefore, if our assumption is correct, sample of PLV SX Phoenicis could increase the number of currently known SX Phoenicis by 30% and the number of field SX Phoenicis by a factor of five. This increase in the sample size and "unusual" location of their discovery could play an important role in characterizing not only this type of variables but blue stragglers as well. Follow-up efforts are under way.

revise; this relation was known... Furthermore we found an interesting correlation between colors of EB and EW type binaries and their period (cf. §3.2.1). As the spectral type (determined from $g - i$ SDSS color) of these binaries changes from approximately K6 to F8,their periods decrease form 9.3 to 5.4 hours.

expand... We would like to conclude that processing this ammount of data is still manageable by human resources. However with the upcoming large surveys (e.g. Gaia and LSST), automated schemes will have to classify the vast ammounts of data. In order to do this, large, clean and relatively faint samples of objects are required to train the automated classification algorithms. Our catalog fulfills these requirements and is available online (**link here**) and as an electronic supplement to this article.

This work was fully or partially supported by the Gaia Research for European Astronomy Training (GREAT-ITN) Marie Curie network, funded through the European Union Seventh Framework Programme ([FP7/2007-2013] under grant agreement n° 264895.

Ž.I. acknowledges support by NSF grants AST-0707901 and AST-1008784 to the University of Washington, by NSF grant AST-0551161 to LSST for design and development activity, and by the Croatian National Science Foundation grant O-1548-2009. The LINEAR program is funded by the National Aeronautics and Space Administration at MIT Lincoln Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

A. Comparison to Extant Catalogs of Variable Stars

A.1. Comparison to General Catalog of Variable Stars and AAVSO International Variable Star Index

In order to estimate the number of previously unknown variable stars in the PLV catalog, we compared it to two online catalogs — the General Catalog of Variable Stars (GCVS, Samus et al. 2009) and the American Association of Variable Star Observers International Variable Star Index (VSX, Watson et al. 2012). The Topcat tool (Taylor 2005) was used to find positional matches within 3 arcsec radius (in early February 2013). Our results are summarized in Figures 20 and 21.

Approximately 60% of PLV objects could not be matched to an VSX catalog entry, and approximately 90% could not be matched to a GCVS entry. **Lovro, is this due to faint flux limits?** Majority of the unmatched objects in both catalogs are eclipsing binaries, followed by c type RR Lyrae, SX Phoenicis/δ Scuti candidates and long period variables. Classification of the matched objects shows good overall agreement between catalogs, and very good agreement for particular types of objects (e.g. ab type RR Lyrae). A full visual re-inspection of light curves for the objects matched in VSX and GCVS was performed, and

we stand by our classification in all cases. In Figure 22 we show several examples where the classification from GCVS and/or VSX did not match PLV classification.

Comparison to VSX and GCVS motivated us to introduce two more variable star classes: anomalous Cepheids and BL Herculis. Both can have light curves that are very similar to ab type RR Lyrae. Furthermore, their colors are also similar to ab type RR Lyrae colors. However, some of them depart slightly from locus populated by ab type RR Lyrae (in the color-period and other diagrams). Lacking data that would allow more precise determination of their class, VSX and GCVS classification in majority of the cases (less than 10 cases in total). **Lovro, tu kao da fali neki glagol?**

move this to main text when describing classes in PLV Matching PLV to GCVS and VSX showed us that many other interesting object types could be extracted from PLV. Many of these are not periodic and therefore we made no true attempt to classify them. We did, however, stumble upon some while examining the light curves. Some of these variables and transients (i.e. active galactic nuclei, AM Herculis, BL Lacertae, BY Draconis, cataclismic variables, quasars, RS Canum Venaticorum...) were registered in the "Other" PLV class.

A.2. Comparison to RR Lyrae Catalog from the Catalina and Mount Lemmon Surveys

We also compared our results with the combined RR Lyrae catalogs assembled by Drake et al. (2013a) and Drake et al. (2013b). Their catalogs include 15,000 ab type RR Lyrae selected from more than 200 million light curves obtained by Catalina Schmidt Survey (CSS) and Mount Lemmon Survey (MLS) over 20,000 deg^2 of sky down to a magnitude limit $V = 20$. In the following text we refer to this work as DR13. Approximately 6,460

DR13 objects are located inside area covered by both PLV and DR13 (approximately $125° < \text{R.A.} < 268°$ and $-13° < \text{Dec} < 65°$). A cut on the magnitude range that corresponds to brightness of objects potentially included in PLV ($14 < V < 17$) selects approximately 3,170 ab type RR Lyrae from DR13. In further analysis we use these area and magnitude cuts, where applicable.

A 3 arcsec radius match between the initial 200,000 object sample and DR13 selects a total of 2,612 objects (Figure 23). All but 3 are classified as variable and included in the PLV catalog. Only 86 ($\approx 3\%$) of the matched objects are not classified as ab type RR Lyrae in PLV. The variable group **what does "variable group" mean?** (66 objects in total) is dominated by objects that have poor LINEAR data and thus could not be reliably classified. Their median magnitude and coordinates are distributed roughly equaly within the PLV brightness range and observed area. These objects were identified as periodic and/or variable in PLV, but the variability type could not be determined (i.e. they were classified as "Other" in PLV). Thirteen of the remaining objects with better data were classified as c type RR Lyrae, one was classified as EB/EW eclipsing binary, one as a BL Herculis candidate and two as anomalous Cepheids (in VSX, these two objects were classified as ACEP and ACEP:). Therefore, the only true disagreement in classification between LINEAR and DR13 is for those **which?** 13 objects (0.5%). Several examples of objects where PLV and DR13 classification did not match can be found in Figure 24.

ZI: ostatak je fest predugacak... Lovro, mozemo li to skratiti na jedan paragraf?

A total of 620 DR13 objects (within the imposed limits on magnitude and area) could not be matched to our initial 200,000 object sample. Two thirds of those could be matched to the entire LINEAR data set available through SkyDOT. These objects are roughly isotropically distributed in the area observed by LINEAR. Their LINEAR light curves were

inspected, folded with DR13 periods. While we did not try to classify them, we do confirm that $\approx 85\%$ appear to be periodic and most of them have typical ab type RR Lyrae light curves. This indicates that our initial cuts (cf. 2.1) were perhaps too strong. **too strong for selecting ab type RR Lyrae?**

Thirteen objects matched in SkyDOT passed our initial cuts and it is a question why were they not included in our 200,000 sample. All of them have light curves typical for ab type RR Lyrae. (**kaj da radimo s ovim? Da dodam u PLV? Ne moze, jer nemaju GAIA periode; ZI: ne znam jos sto s tim, ne znam ni kako je tako nesto moguce...**)

Total of 159 objects that could not be matched to full (SkyDOT) LINEAR database. We assume that they do not exist as LINEAR objects and therefore exclude them from further analysis.

From the total of 2,923 ab type RR Lyrae in PLV, 2,888 are inside the area covered by both LINEAR and DR13. Approximately 365 of those could not be matched to DR13 catalog. These objects are also roughly equally dispersed in the overlapping area and magnitude range between PLV and DR13.

We selected random 15% subsample of those 365 objects missing from the DR13 for inspection. Subsample was matched to online Catalina Surveys Data Release 2 (CSDR2; Drake et al. 2009) publicly available data. Approximately 10% of the objects from the subsample could not be successfully matched to an CSDR2 object. Remaining (matched) objects were folded with PLV periods, and light curves were visually inspected. In 15% of the cases light curves could not be identified as ab type RR Lyrae, given the CSDR2 data.

We also inspected a random subsample of light curves from CSDR2 that were classified as periodic and variable in the PLV sample, but not as ab type RR Lyrae. Our conclusion is

that other types of periodic variable objects could easily be extracted from CSDR2 as well.

Comparison with DR13 can also give an additional constraint (upper limit) on the completeness of our sample. If we assume that in the area covered by LINEAR and DR13, and within the adopted magnitude range a total of ≈ 3600 (i.e. 2612 + 620 + 365) ab type RR Lyrae could be extracted, that gives us an upper limit on the completeness of PLV ab type RR Lyrae sample of $\approx 85\%$ (since the total number of ab type RR Lyrae in PLV is 2,913).

REFERENCES

Abazajian, K. et al. 2009, ApJS, 182, 543

Akerlof, C. et al. 2000, AJ, 119, 1901

Andersen, J. 1991, A&A Rev., 3, 91

Ankerst, M. et al. 1999, "Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data", 49

Barron, J. T. et al. 2008, AJ, 136, 1490

Becker, A. C. et al. 2004, ApJ, 611, 418

Becker, A. C. et al. 2011, ApJ, 731, 17

Bertin, E. & Arnouts, S. 1996, A&A Supplement, 117, 393

Bond, B. et al. 2010, ApJ, 716, 1.

Burke, B. et al. 1998, Experimental Astronomy, 8, 31-40

Covey, K. et al. 2007, AJ, 134, 2398

Cohen, R.E. & Sarajedini, A. 2012, MNRAS, 419, 342

Davenport, J.R.A. et al. 2013, ApJ, 764, 62

Deb, S. & Singh, H.P. 2009, A&A, 507, 1729

Debosscher, J. et al. 2007, A&A, 475, 1159

Devor, J. 2004, ApJ, 628, 411

Dempster, A.P., Laird, N.M. & Rubin, D. 1977, J. R. Stat. Soc. Ser. B 39, 1

Dimitrov, D.P. & Kjurkchieva, D.P. 2010, MNRAS, 406, 2559

Drake, A. J., Djorgovski, S. G., Mahabal, A., et al. 2009, ApJ, 696, 870

Drake, A. J., Catelan, M., Djorgovski, S. G., et al. 2013a, ApJ, 763, 32

Drake, A. J., Catelan, M., Djorgovski, S. G., et al. 2013b, ApJ, 765, 154

Eggen, O.J. 1967, MmRAS, 70, 111

Eyer, L. & Blake, C. 2005, MNRAS, 358, 30

Eyer, L. & Mowlavi, N. 2007, arXiv:0712.3797

Eyer, L. et al. 2012, arXiv:1201.4889v1

Eyer, L., Holl, B., Pourbaix, D., et al. 2013, arXiv:1303.0303

Friedman, J.H. 1984, A Variable Span Smoother. Technical Report No. 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University

Giuricin, G., Mardirossian, F. & Messetti, M. 1983, å, 119, 218

Guinan, E.F. et al. 1998, ApJ, 509, L21

Hoffman, D. I., Harrison, T.E., and McNamara, B. J. 2009, AJ, 138, 466

Ivezić, Ž. & Elitzur, M. 1995, ApJ, 445, 415.

Ivezić, Ž. et al. 2000, AJ, 120, 963

Ivezić, Ž. et al. 2007a, AJ, 134, 973

Ivezić, Ž. et al. 2007b, ASP Conference Series, 34, 165 (also arXiv:0701508)

Ivezić, Ž. et al. 2008, arXiv:0805.2366

Ivezić, Ž. et al. 2008b, ApJ, 684, 287.

Ivezić, Ž., Beers, T.C. & Jurić, M. 2012, ARA&A, 50, 251

Jeon, Y.-B., et al. 2004, AJ, 128, 287.

Kaiser, N. et al. 2002, Proc. SPIE, 4836, 154

Lang, D. et al. 2010, AJ, 139, 1782

Lomb, N. R. 1976, ApS&S, 39, 447

Lupton, R. H. et al. 2002, Proc. SPIE, 4836, 350

MacLeod, C.L. et al. 2010, ApJ, 721, 1014

MacLeod, C.L. et al. 2012, ApJ, 753, 106

Monet, D. G. et al. 2003, AJ, 125, 984

Oosterhoff, P. T. 1944, Bull. Astron. Inst. Neth., 10, 55

Pier, J. R. et al. 2003, AJ, 125, 1559

Pojmański, G.2002, Acta Astron., 52, 397

Prša, A. & Zwitter, T. 2005, ApJ, 628, 426

Richards, J. W. et al. 2011, ApJ, 733, 10

Rodríguez, E. et al. (2001), A&A366, 178

Ruan, J. J. et al. 2012, ApJ, 760, 51

Rucinski, S. M. 1974, Acta Astron., 24, 119

Rucinski, S. M. 1992, AJ, 103, 960

Rucinski, S. M. 1997, AJ, 113, 407

Rucinski, S. M. & Duerbeck, H.W. 1997, PASP, 109, 1340

Samus, N. N., Durlevich, O. V., & et al. 2009, VizieR Online Data Catalog, 1, 2025

Scargle, J. D. 1982, ApJ, 263, 835

Sesar, B. et al. 2007, AJ, 134, 2236

Sesar, B. et al. 2010, ApJ,708, 717

Sesar, B. et al. 2011, AJ, 142, 190

Sesar, B. et al. 2013, submitted to AJ.

Skrutskie, M. F. et al. 2006, AJ, 131, 1163

Smolčić, V. et al. 2004, ApJ, 615, L142

Spano, M., Mowlavi, N., Eyer, L., et al. 2011, A&A, 536, A60

Stellingwerf, R.F. 1978, ApJ, 224, 953

Taylor, M. B. 2005, Astronomical Data Analysis Software and Systems XIV, 347, 29

Tisserand, P. 2012, arXiv:1110.6579

Tu, X. & Wang, Z. 2012, arXiv:1207.0294

Woźniak, P. R. et al. 2004, AJ, 127, 2436

York, D. G. et al. 2000, AJ, 120, 1579

Yun, L., et al. 2012, arXiv:1209.2065

Watson, C., Henden, A. A., & Price, A. 2012, VizieR Online Data Catalog, 1, 2027

Whitelock, P.A., et al. 2006, MNRAS, 369, 751

Wood, P. R., & Sebo, K. M. 1996, MNRAS, 282, 958

Wright, E. L. et al. 2012, AJ, 140, 1868

Zechmeister, M. & Kürster, M. 2009, A&A 496, 577

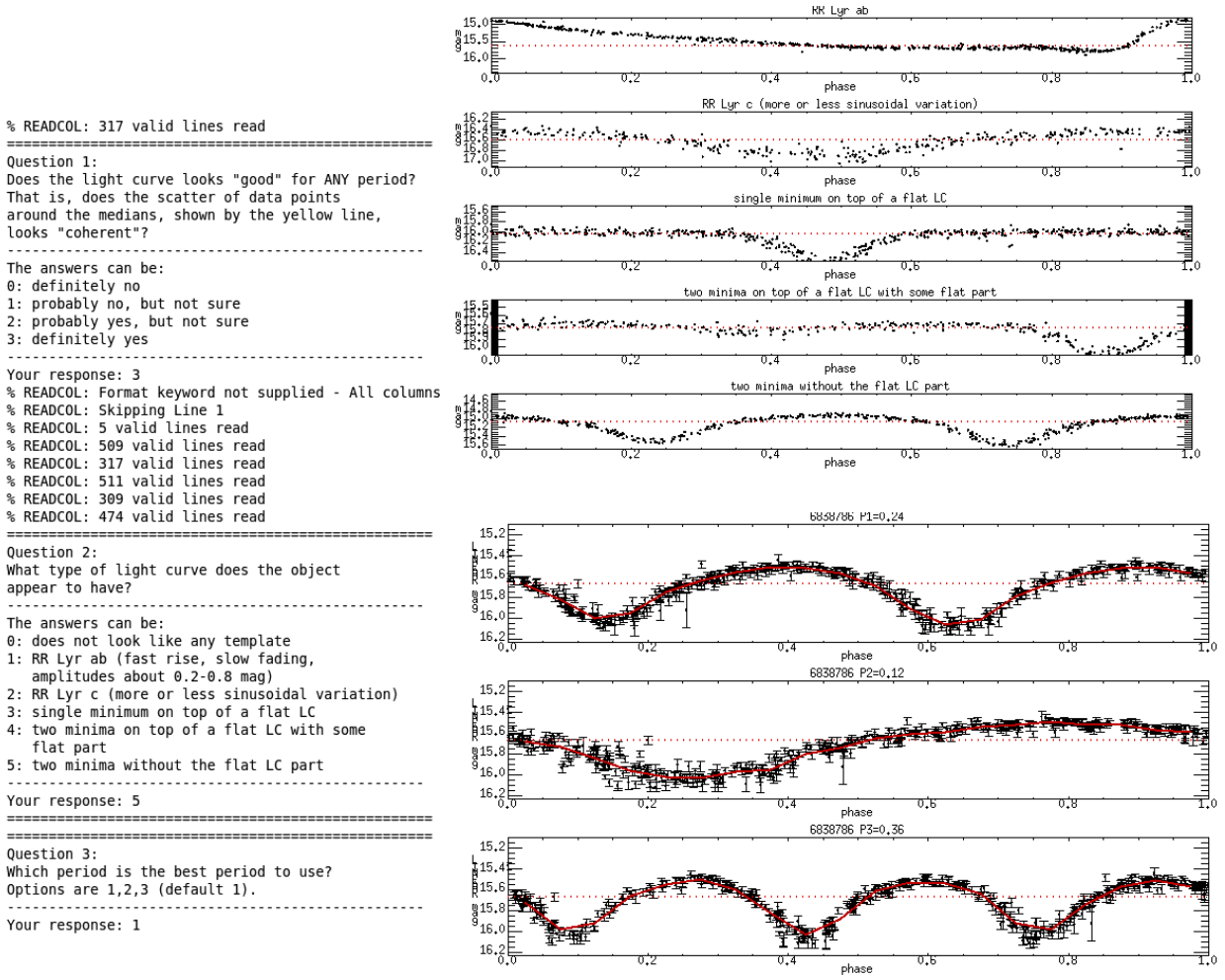

Fig. 1.— User interface for the classification tool. The three bottom right panels show phased LINEAR light curves of the given object for the three most probable periods calculated by the Supersmoother algorithm. The five top right panels represent light curve templates used in classification.

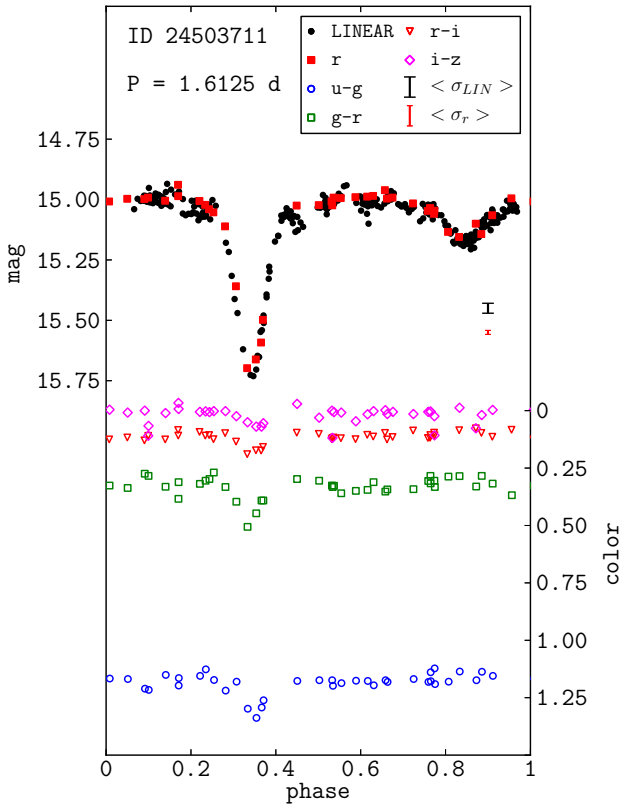

Fig. 2.— An example of LINEAR/SDSS synergy. Scale on the left corresponds to unfiltered LINEAR magnitudes and scale on the right to SDSS colors. Red and black bars show average LINEAR and SDSS errors, respectively. LINEAR provides a better cadence for studying variable objects, while SDSS provides multi-band photometry that encodes valuable additional information about the variable object.

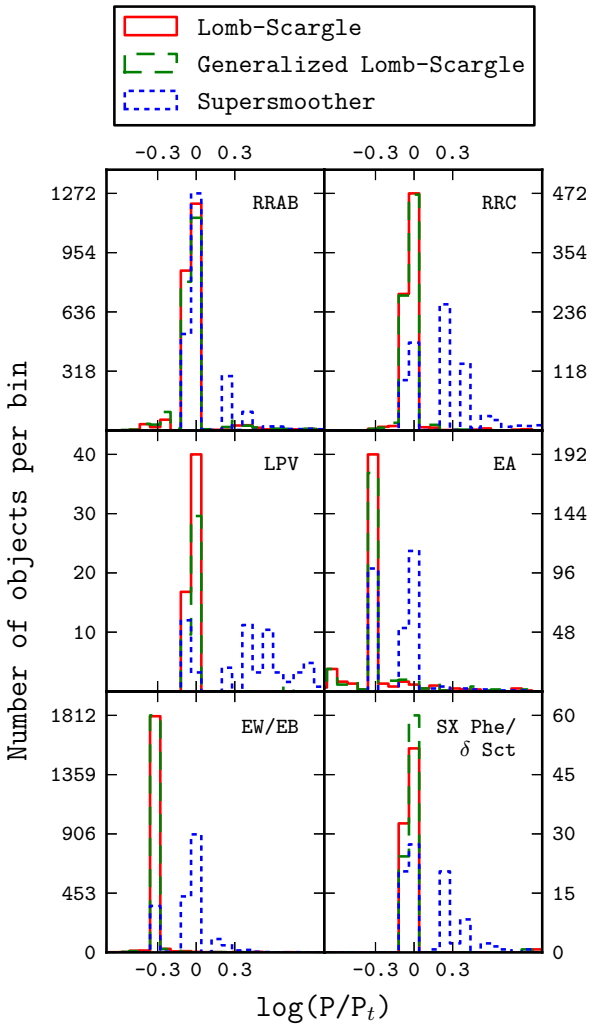

Fig. 3.— Comparison of the three period finding methods, separately for each of the six main light curve types (clockwise, from the top left: ab type RR Lyrae, c type RR Lyrae, EA eclipsing binaries (Algol), SX Phe/δ Scu variables, EW/EB eclipsing binaries (β Lyr and W UMa), and long-period variables (asymptotic giant branch stars). The abscissa shows the logarithm of the ratio of the period computed by each method and the visually confirmed true period (note that a factor of 2 bias corresponds to 0.30 on logarithmic scale). Note that the Lomb-Scargle methods consistently underestimate period of EA and EW/EB light curves by a factor of 2 (this systematic effect has been corrected in the public catalog).

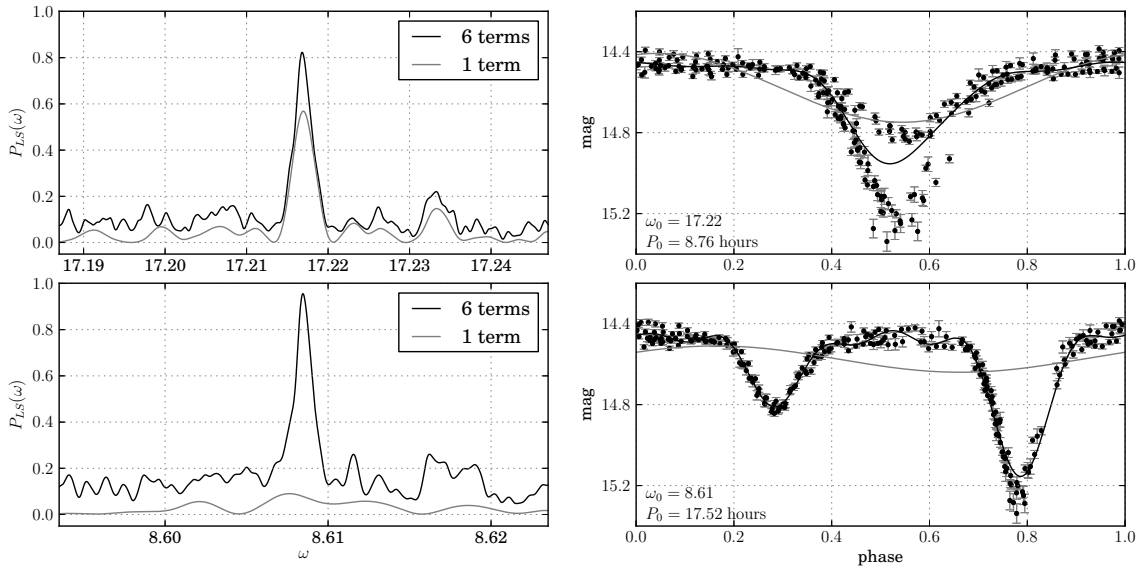

Fig. 4.— An illustration of the failure of the Lomb-Scargle method to find the correct period when the light curve shape significantly differs from a single sinusoid. The two top panels show the Lomb-Scargle periodogram (left) and phased light curves (right) for truncated Fourier series models with one and six terms. Symbols with error bars represent LINEAR data for star with ID=14752041 (the data and the python code to produce this figure, including period estimation, are publicly available from the *astroML* site, http://astroml.github.com). Phased light curves are computed using the aliased period favored by the single-term model, and the model light curves are shown by lines using the same line styles as in the top left panel. The correct period is favored by the six-term model but unrecognized by the single-term model, as illustrated in the bottom left panel. The phased light curve constructed with the correct period is shown in the bottom right panel.

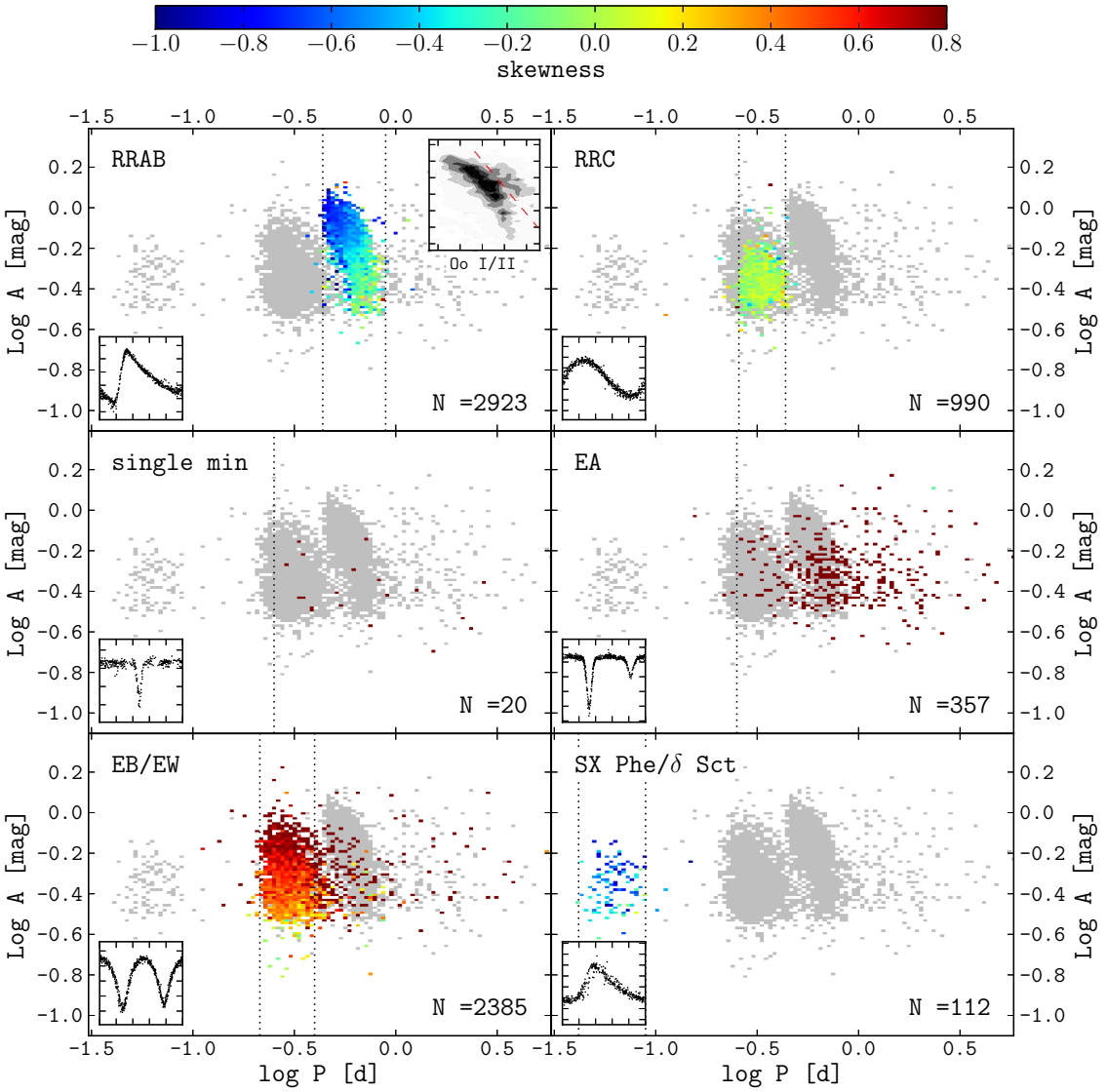

Fig. 5.— Period—Amplitude diagram for visually confirmed periodic LINEAR variables (PLV). Each panel represents a given class of variable stars confirmed by visual classification. Width of the bins is 0.03 in color and 0.02 in $log(P/day)$. Bins are color coded by the median value of skewness (per bin). Grey backgroud corresponds to all PLV sample variables. Insets in panels represent a typical light curve for the variability type in the given panel. N is the total number of objects of a given type. Top right inset in the "RRAB" panel shows separation between Oosterhoff type I and type II RR Lyrae ab, former being left and below dashed red line and the latter right and above. The gray map shows the density of ab type RR Lyrae per bin. Axes in the folded light curve diagrams correspond to phase and magnitude, and in the inset in the Oosterhoff diagram for RR Lyrae to $log(P)$ $[d]$ and $log(A)$ $[mag]$.

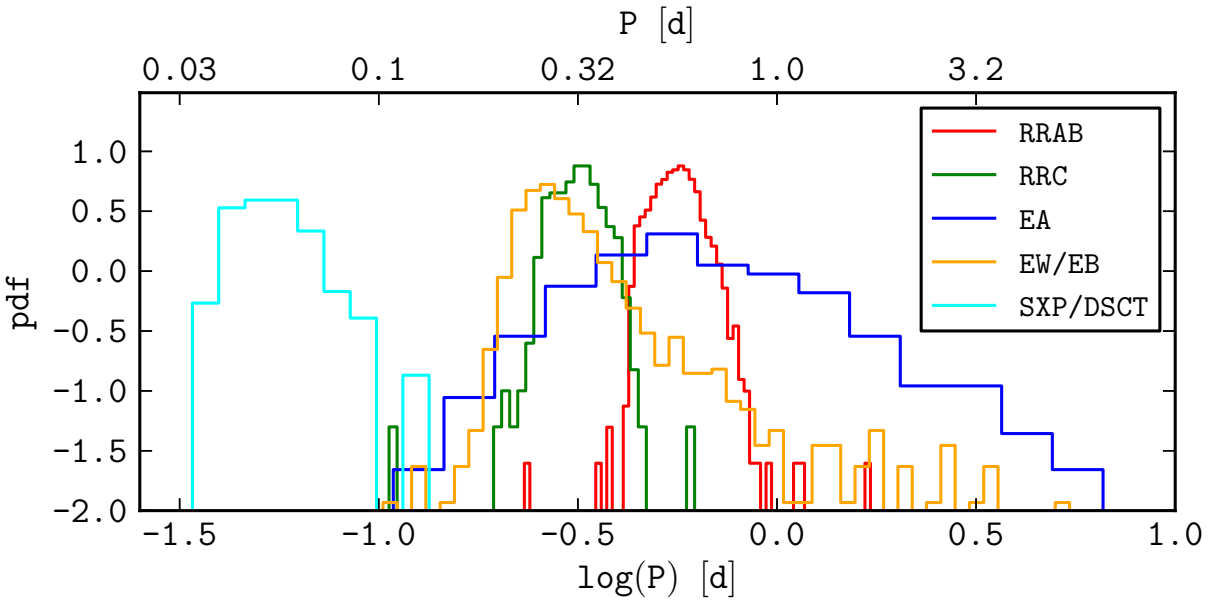

Fig. 6.— The period distribution for the five most populous variability classes. While SX Phe/δ Scu candidates clearly stand out ($P < 0.1$ day), and ab type and c type RR Lyrae are fairly well separated by P =0.4 days, eclipsing binaries overlap with the period range of RR Lyrae stars (especially EW/EB type eclipsing binaries and c type RR Lyrae).

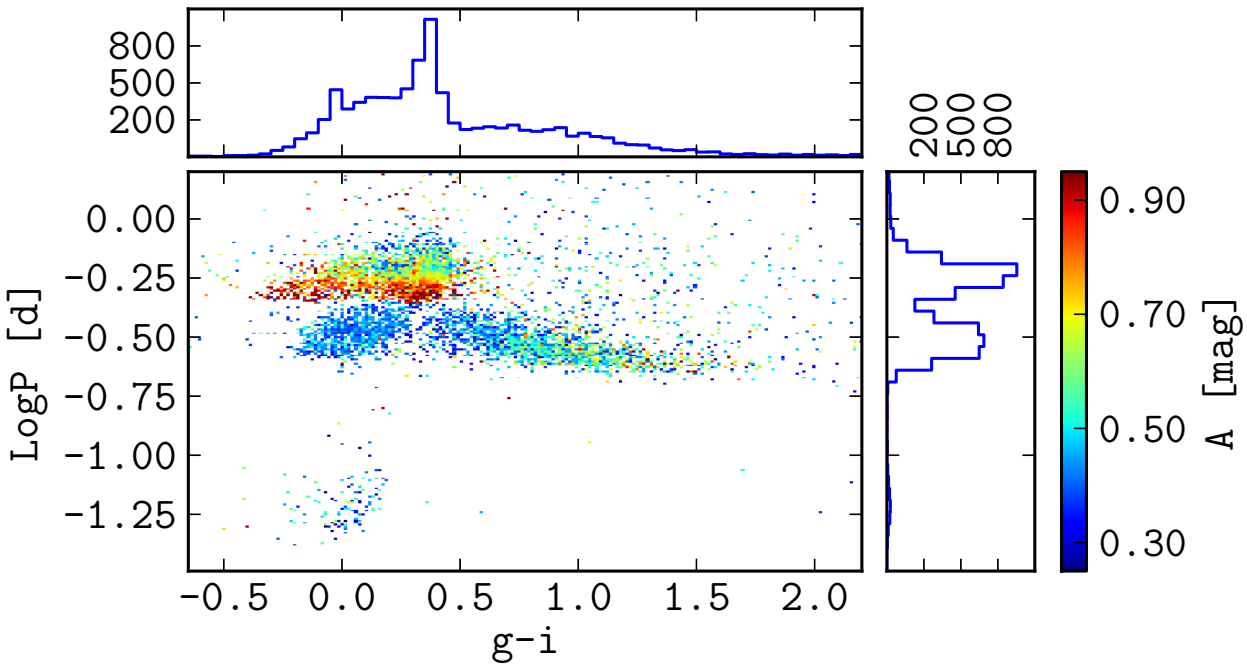

Fig. 7.— The distribution of periodic variables in the period-color diagram. Bins are color coded by the median value of light curve amplitude according to the legend on the right. The two histograms show marginal distributions of the period and the $g - i$ color.

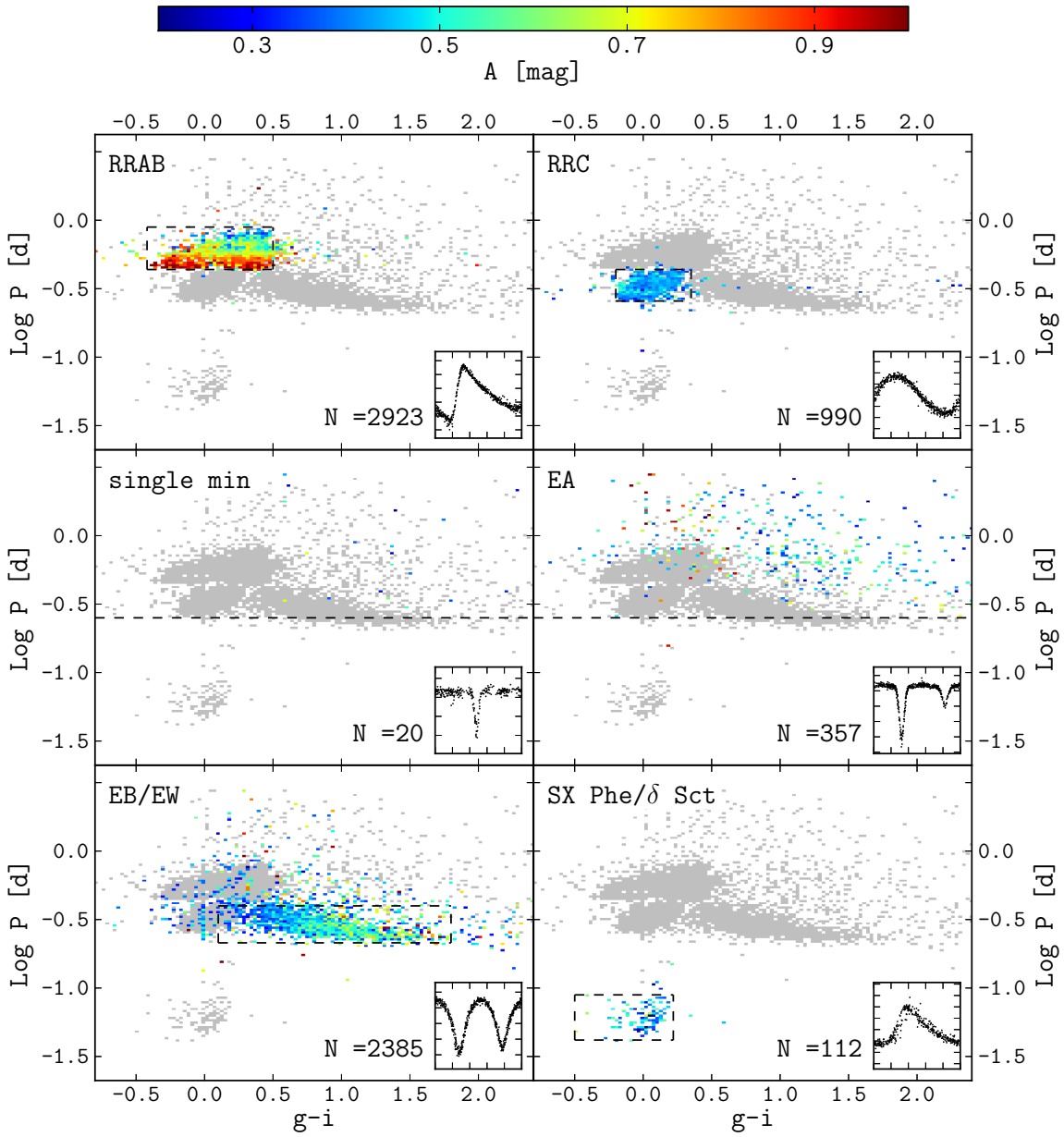

Fig. 8.— The distribution of visually confirmed periodic LINEAR variables (PLV) in the period-color diagram. Each bin has been color coded by the median amplitude of objects inside it, according to the color bar above. Width of the bins is 0.03 in color and 0.02 in $log(P)$ $[d]$. Grey backgroud represents all PLV sample variables. Insets in panels represent a typical light curve for the variability type in the given panel. N is the number of objects of a given type. Axes in the folded light curve diagrams correspond to phase and magnitude. The dashed lines outline the selection boundaries listed in Table 2 and discussed in §4.1.

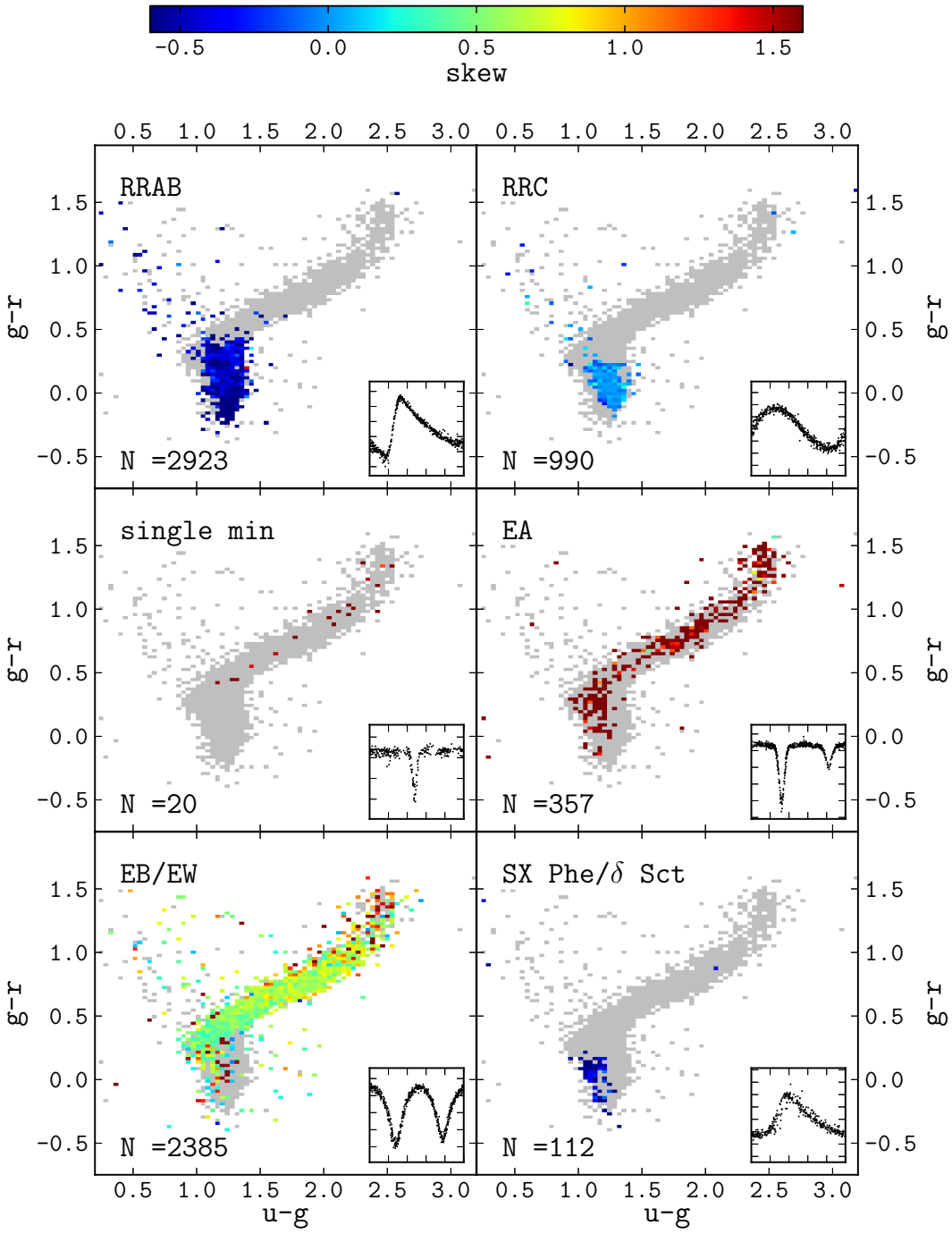

Fig. 9.— Analogous to Figure 8, except that the PLV sample distribution is shown in the SDSS $g-r$ vs. $u-g$ color-color diagram, and the color coding is based on the light curve skewness.

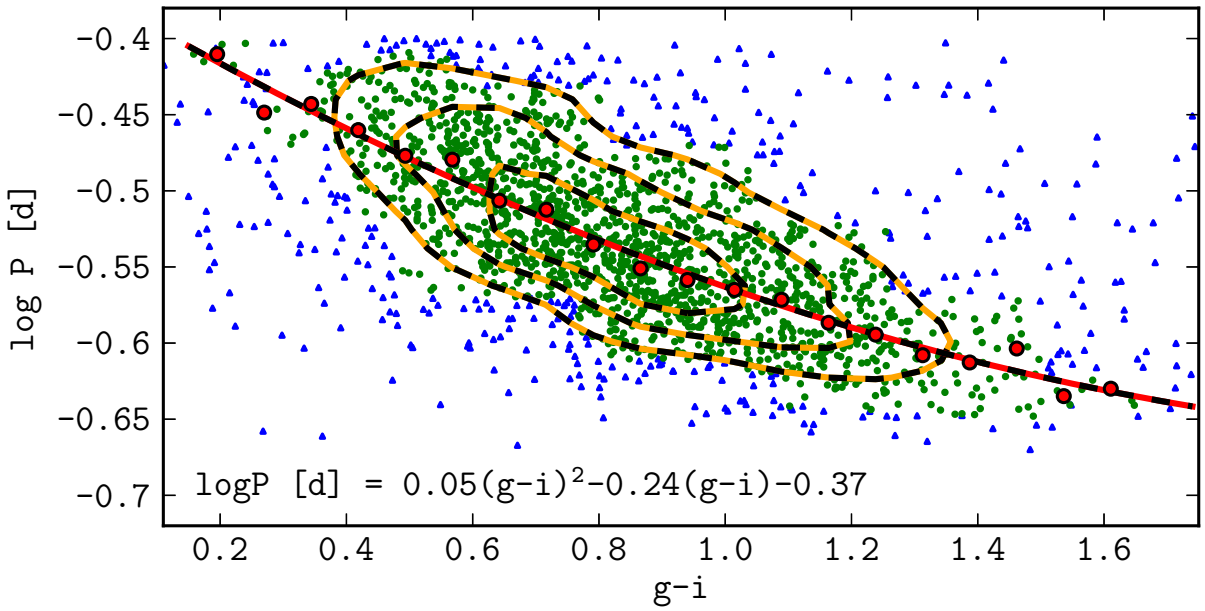

Fig. 10.— Quadratic fit for correlation between period and color of EB/EW binaries. Selected objects were visually classified as EB/EW binaries and satisfy criteria: $0.1 < g - i < 1.8$ and $-0.67 < log(P) < -0.4$. Data was binned in 0.1 wide bins in (g-i) and 0.05 wide bins in logP. Objects outside 5% and 95% points of the distribution in color and period per bin were removed from subsequent fitting procedure (blue triangles). Remaining objects (green points) were again binned in 0.1 wide bins in (g-i), and median of logarithm of period in days per bin was calculated (red points). Quadratic function was fit to these red points. Contours designate areas of 5, 10, 15 points per bin, i.e. the density of points.

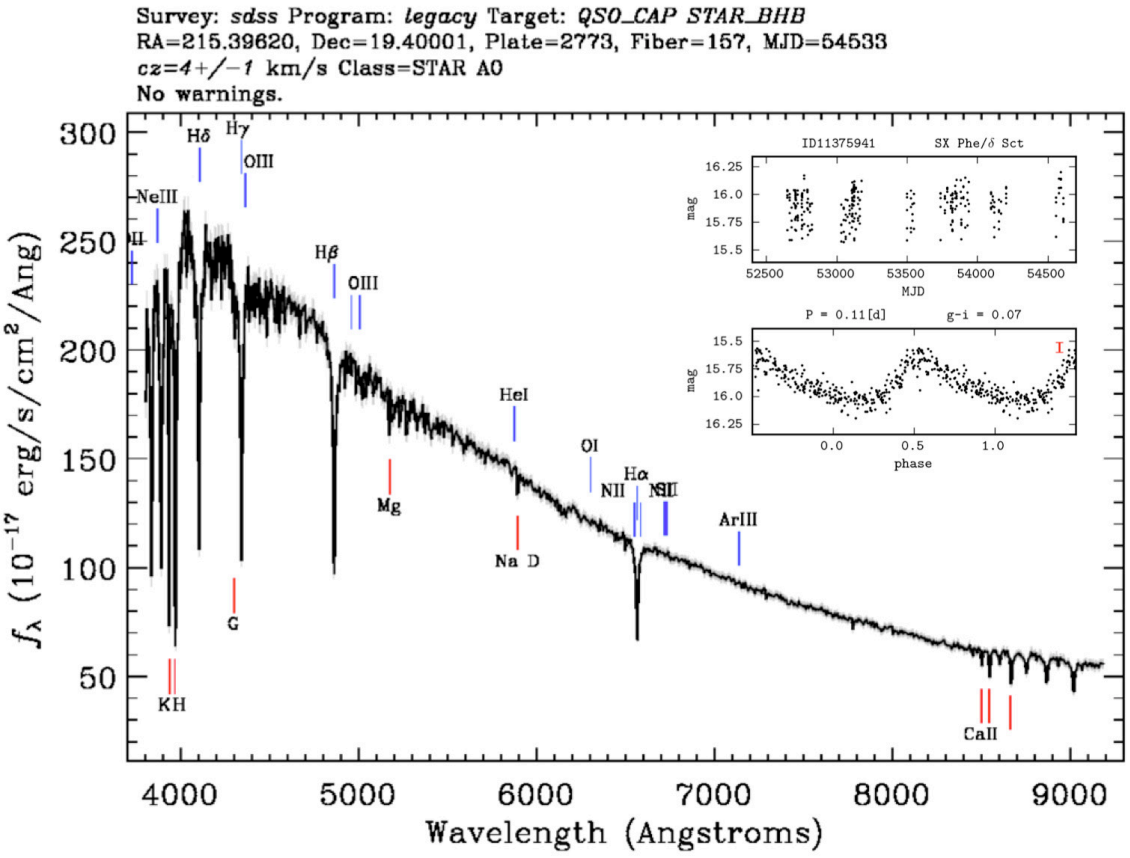

Fig. 11.— Default SDSS visualization of the SDSS spectrum for an SX Phe candidate (LINEAR ID=11375941). The inset shows the observed and phased LINEAR light curves.

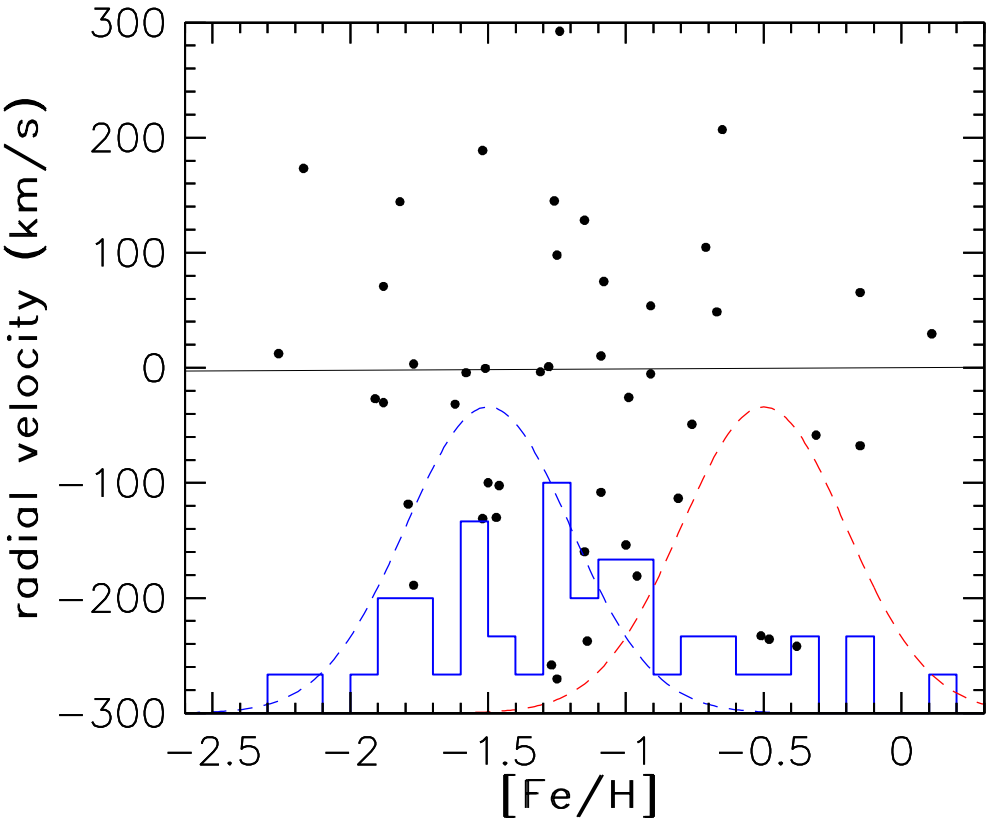

Fig. 12.— The symbols show metallicity vs. radial velocity distribution of 34 candidate SX Phe stars with SDSS spectra (repeated measurements for 7 stars are also shown). The histogram shows the marginal distribution of metallicity. The two Gaussians illustrate expected metallicity distributions for halo stars (left) and disk stars (right), taken from Ivezić et al. (2008b). The metallicity is below the traditional boundary for separating halo and disk stars at $[Fe/H] = 1.0$ dex for 57% of measurements. For these stars, the radial velocity dispersion is 135 km/s, and fully consistent with the halo hypothesis (Bond et al. 2010). Only the four stars with $[Fe/H] > 0.5$ dex and small radial velocity are consistent with the disk hypothesis.

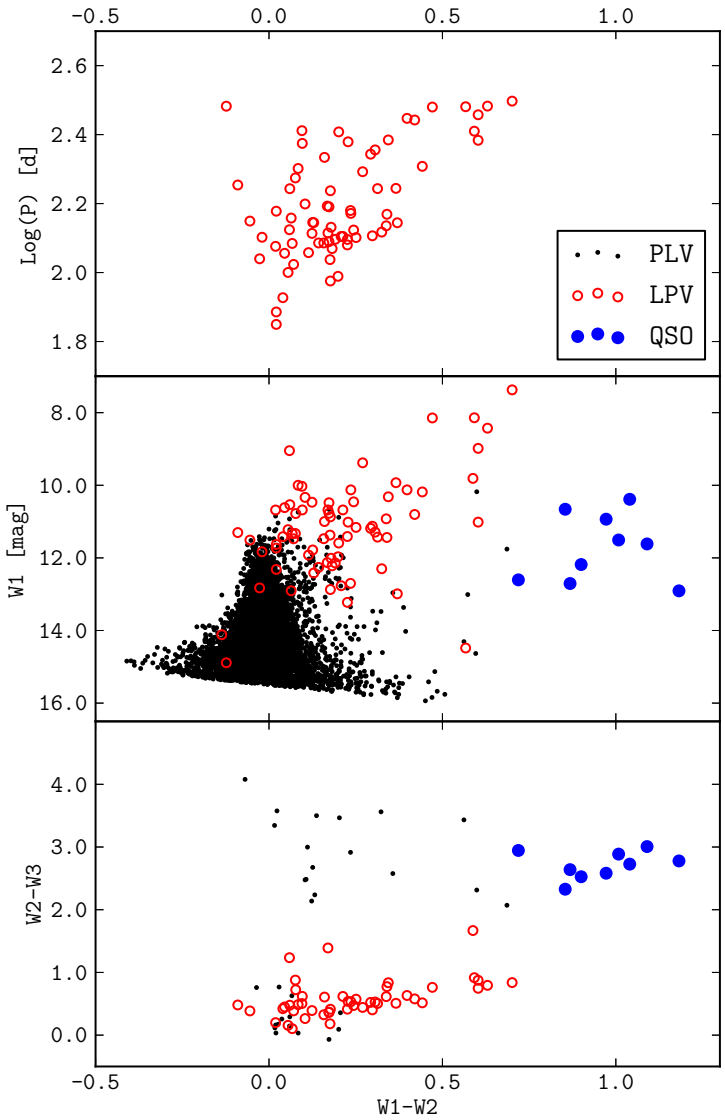

Fig. 13.— The symbols in the middle panel show the distribution of a subsample of periodic variables in the PLV catalog that are detected by the WISE survey and have WISE magnitudes $W1 < 16.5$ and $W2 < 15.5$ (5-σ detection limits). Objects classified as "long-period variables" (defined as variables with periods longer than 50 days, and also include semi-regular variables) are shown as open circles; the majority display infrared excess compared to colors of dust-free stars ($W1 - W2 \sim 0$). Objects with light curves classified as "Other" and with quasar-like infrared colors, $W1 - W2 > 0.7$, are shown as large dots. The bottom panel shows a WISE color-color diagram for the subset of objects that have $W3 < 11.2$ (note that the majority of objects without significant infrared emission do not satisfy this condition). The top panel shows the period-color diagram for long-period variables.

Fig. 14.— Default SDSS *gri* composite images of 18 resolved objects that display visually confirmed variability in LINEAR data. The top number in each panel is the object's LINEAR ID. The first eight objects are clearly galaxies. The light curves for the remaining ten objects are shown in Figure 15. The extremely red source (the second panel in the bottom row) is the brightest carbon-rich AGB star, CW Leo (IRC+10216).

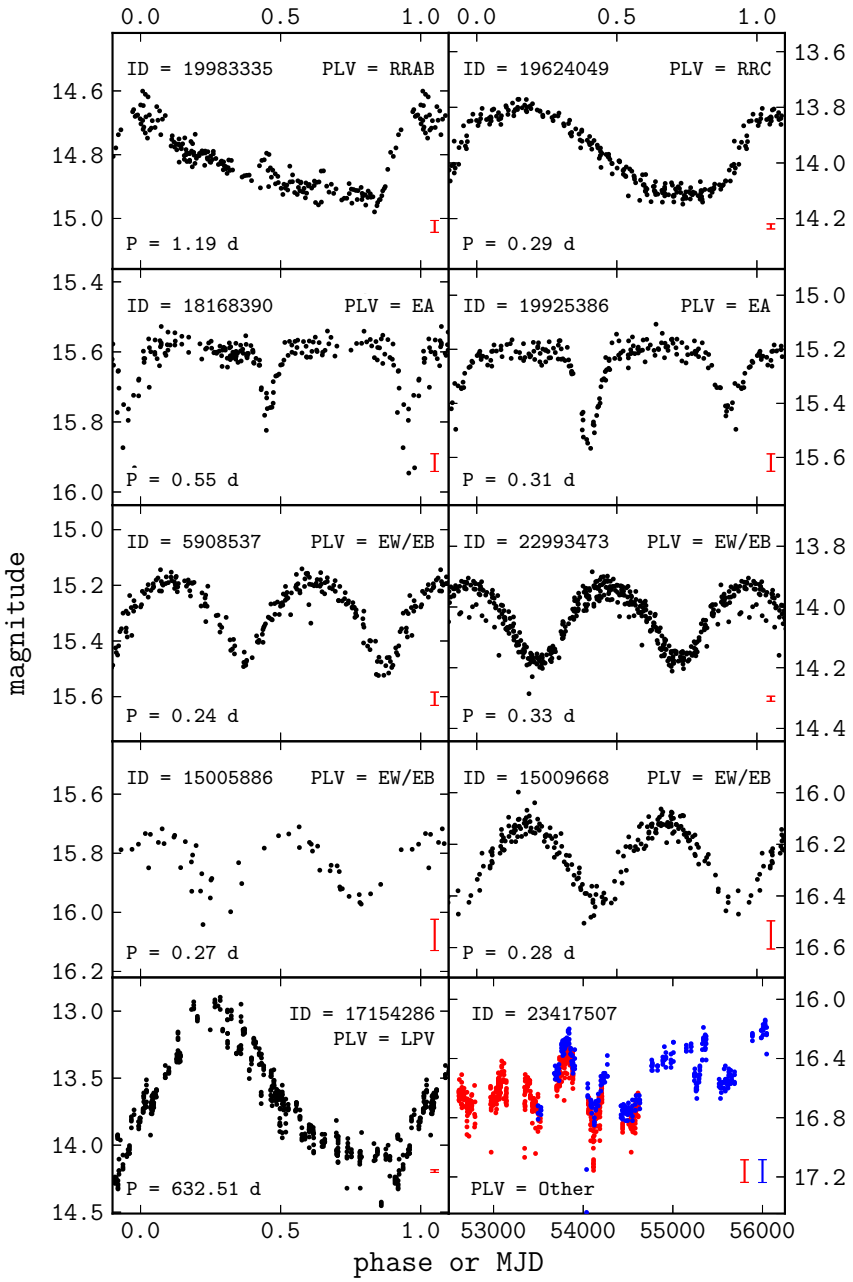

Fig. 15.— The LINEAR light curves for 10 objects that are optically resolved in the SDSS imaging data but do not appear as well-resolved galaxies. Each panel lists LINEAR ID, its visual light curve classification from the PLV catalog, and the best-fit period (in days). The vertical error bars show typical photometric errors for each light curve. All panels except the bottom right panel display phased light curves. The light curve of a quasar in the bottom right panel combines LINEAR (red) and CSDR2 (blue) data and confirms its quasi-periodic behavior (full light curve is shown in this panel).

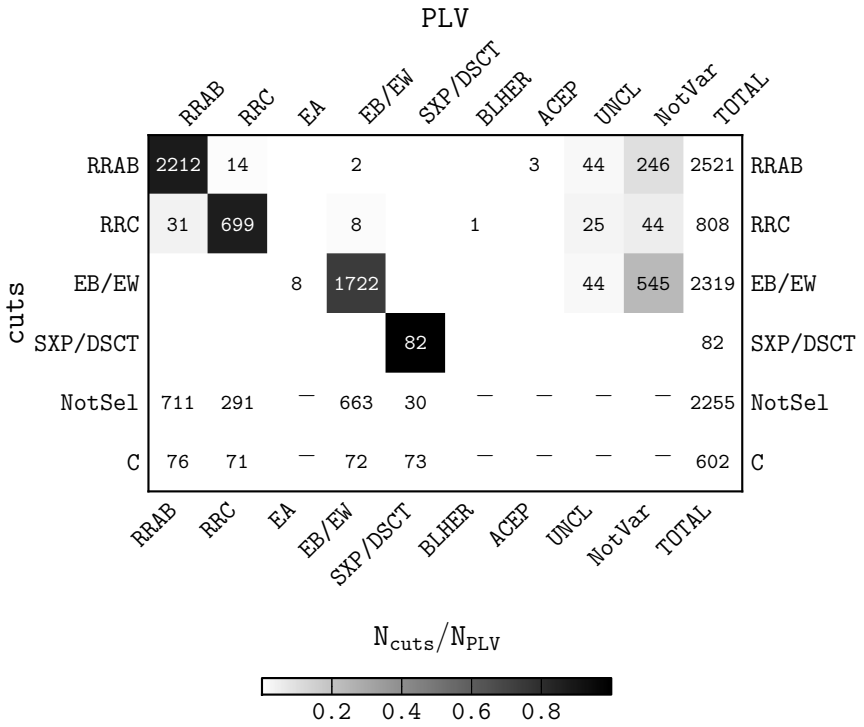

Fig. 16.— A statistical performance comparison between the visually confirmed and classified variable sample listed in the PLV catalog (7,194 objects), and a simple supervised classification algorithm aplied to full sample of all 200,000 candidates LINEAR variables. The selection boundaries for the latter are listed in Table 2. The columns correspond to light curve types used in the PLV catalog; in addition, the column labeled "UNCL" corresponds to variable PLV objects that do not have reliable variability type, and the "NotVar" column corresponds to objects that satisify the selection cuts applied to the full sample but that were not visually tagged as variable and included in PLV. The first four rows correspond to the four analyzed subsamples of variables defined by applied selection cuts. The fifth row, labeled "NotSel", corresponds to objects not selected by automated selection cuts, but nevertheless visually identified as periodic and listed in PLV. The intersection regions are color-coded by the fraction of objects in each row falling into a given region. The last row, labeled "C", lists the completenes of the automated selection method when compared to PLV for the four analyzed variability classes. **Lovro, please remove "602" from the bottom rigth corner; also, I am getting C values of 75 and 70 for the first two, not 76 and 71 as in the figure...**

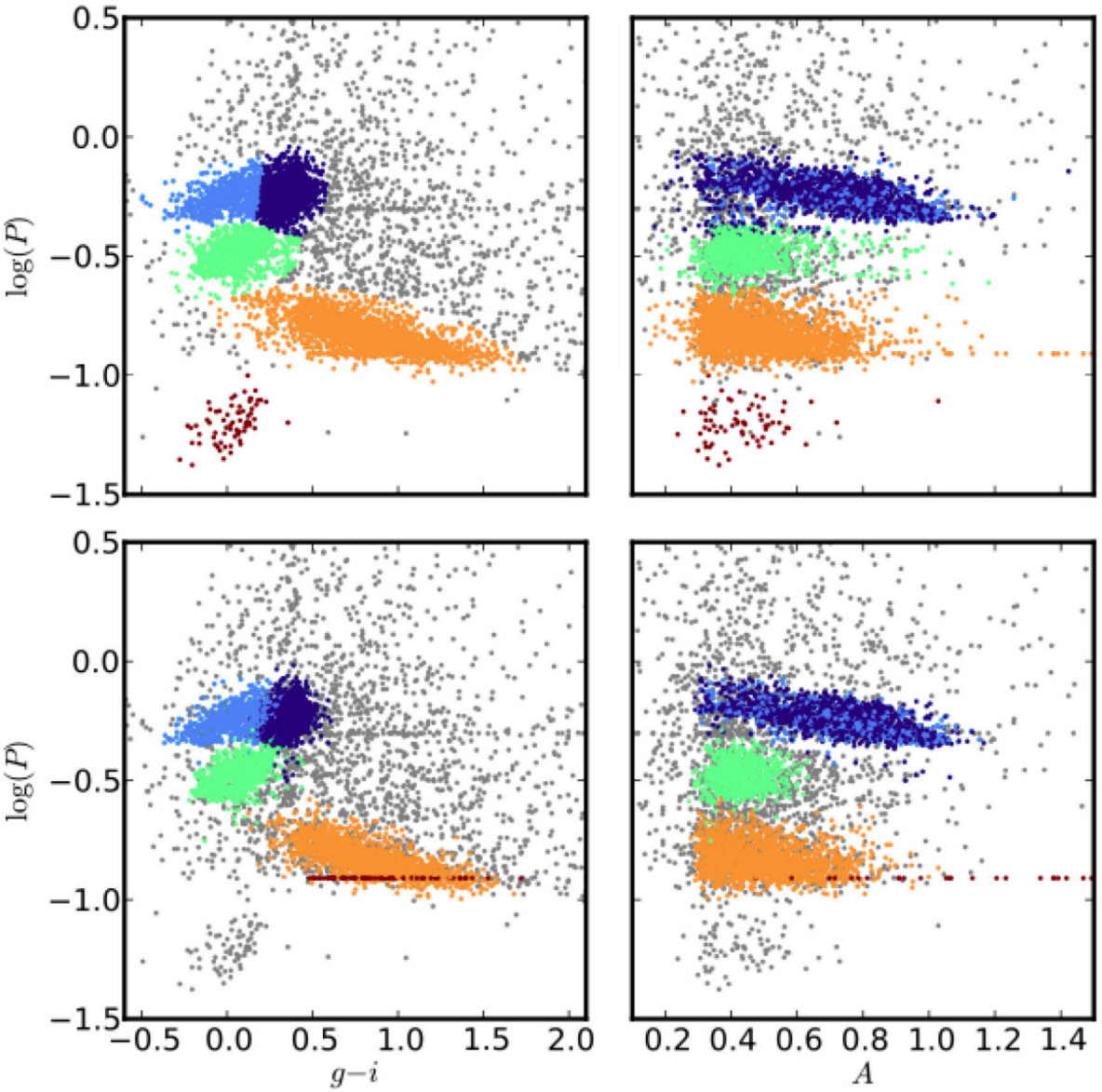

Fig. 17.— Clustering analysis of periodic variable stars from the LINEAR dataset. The top row shows clusters derived using two attributes ($g-i$ and $\log(P)$) and a mixture of 12 Gaussians. The colorized symbols mark the five most significant clusters. The bottom row shows analogous diagrams for clustering based on seven attributes (colors $u-g$, $g-i$, $i-K$, and $J-K$, $\log(P)$, light curve amplitude, and light curve skewness), and a mixture of 20 Gaussians. See Figure 18 for data projections in the space of other attributes for the latter case.

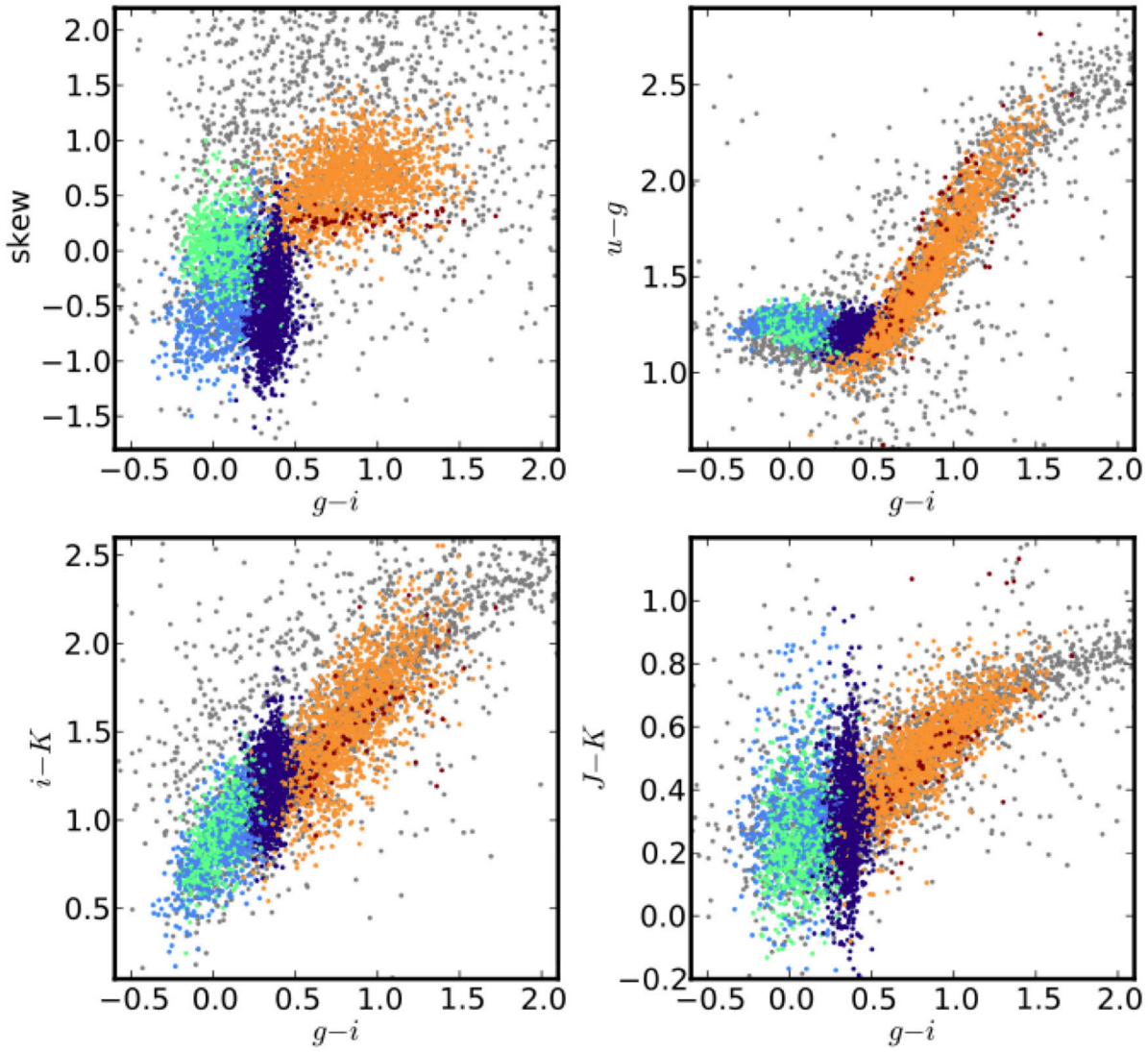

Fig. 18.— Clustering analysis of periodic variable stars from the LINEAR dataset. Clusters are derived using seven attributes (colors $u-g$, $g-i$, $i-K$, and $J-K$, $\log(P)$, light curve amplitude, and light curve skewness), and a mixture of 20 Gaussians. The $\log(P)$ vs. $g-i$ diagram and $\log(P)$ vs. light curve amplitude diagram for the same clusters are shown in the lower panels of Figure 17.

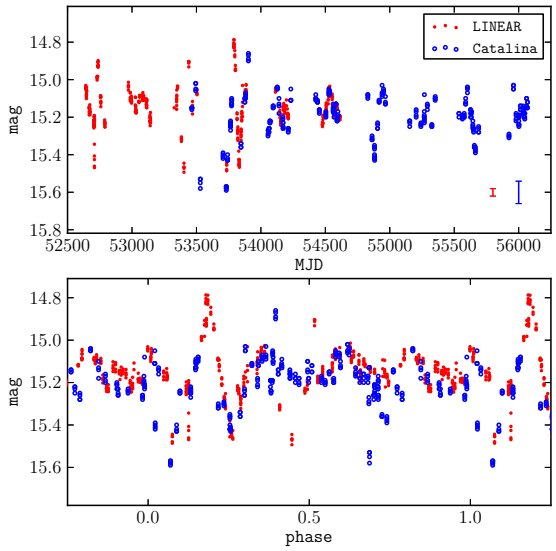

Fig. 19.— Candidate hearbeat star with period of 526.4 days. LINEAR (red points) and CSDR2 (blue circles) have been combined.

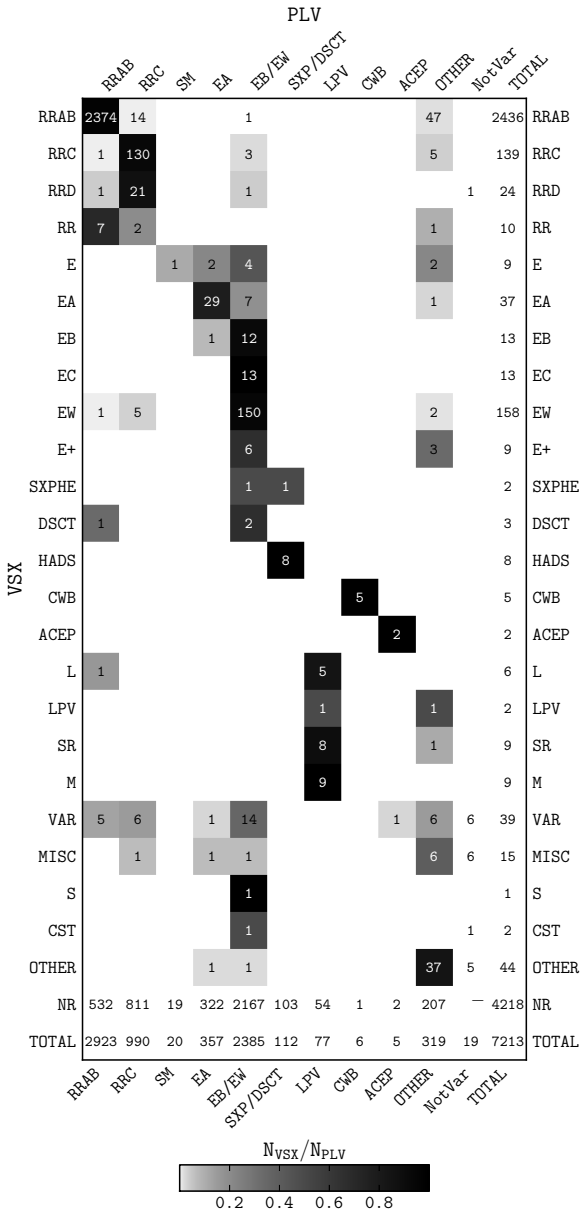

Fig. 20.— The PLV vs. VSX confusion matrix. The column labeled "UNCL" corresponds to variable PLV objects that do not have reliable variability type. The "NotVar" column corresponds to VSX objects that are not included in PLV, and the row "NR" to PLV objects not listed in the VSX catalog. The row "Other" corresponds to VSX variables with classes others than listed in this confusion matrix. The intersection regions are color-coded by the fraction of objects in each row falling into a given region. Acronyms are according to Watson et al. (2012).

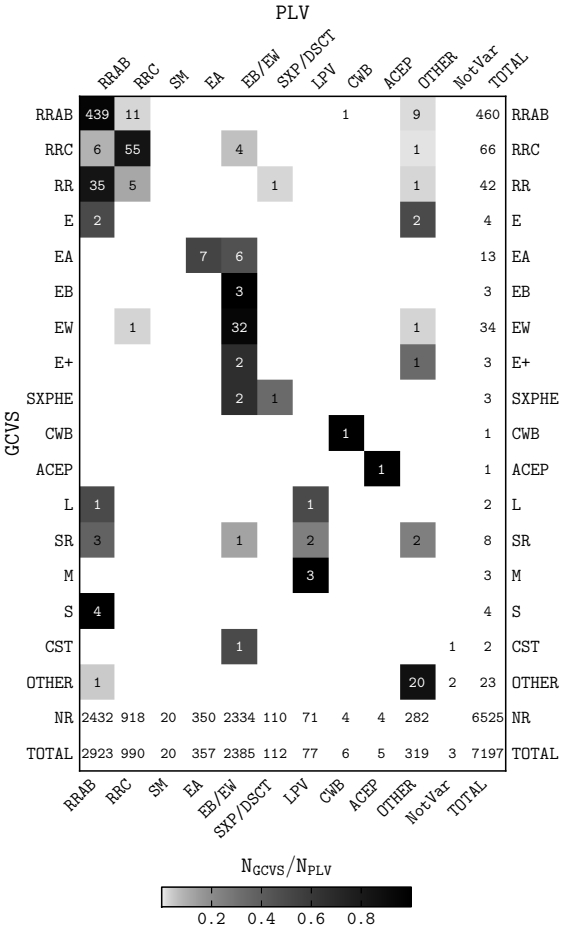

Fig. 21.— The PLV vs. GCVS confusion matrix. The column labeled "UNCL" corresponds to variable PLV objects that do not have reliable variability type. The "NotVar" column corresponds to GCVS objects that are not included in PLV, and the row "NR" to PLV objects not listed in the GCVS catalog. The row "Other" corresponds to GCVS variables with classes others than listed in this confusion matrix. The intersection regions are color-coded by the fraction of objects in each row falling into a given region. Acronyms are according to Samus et al. (2009).

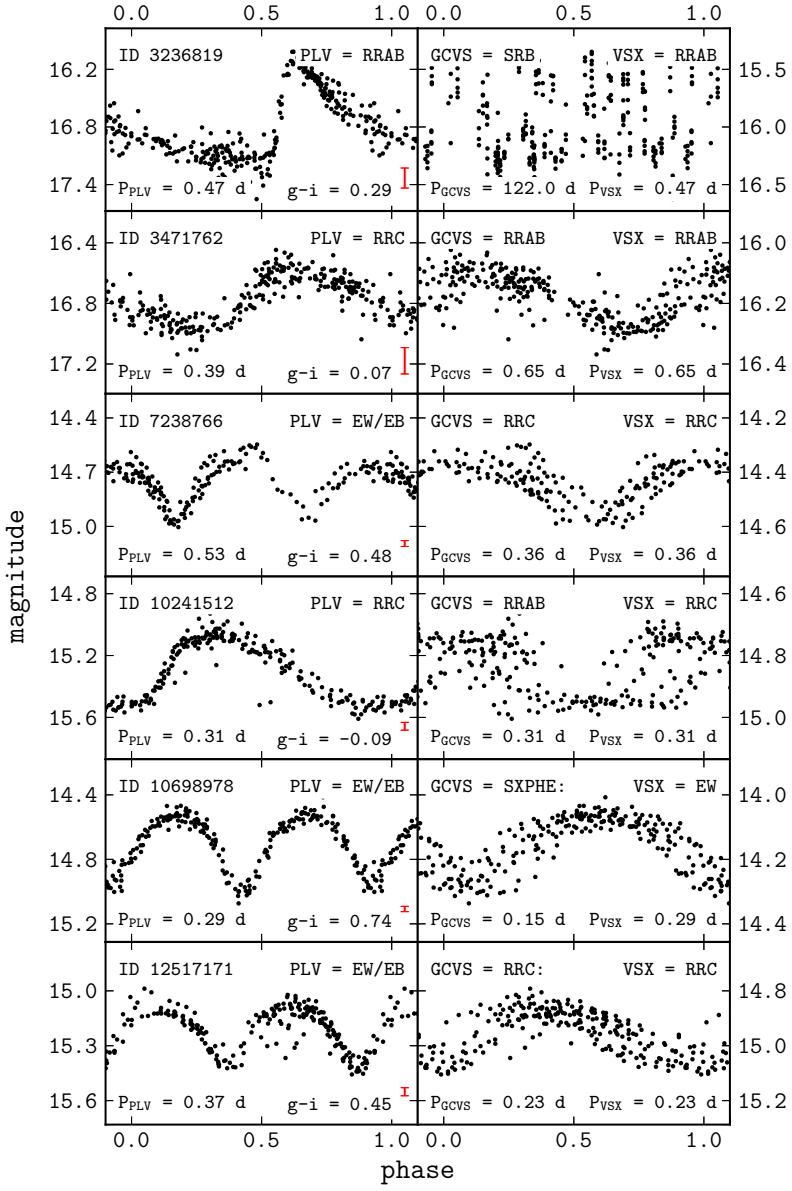

Fig. 22.— Examples of light curves of objects for which the GCVS or VSX period and classification does not agree with the PLV classification. The vertical red bars in each panel in the left column show the median error for LINEAR data. Plots on the left are folded with the PLV periods, and plots on the right are folded with the GCVS periods. It is evident that the PLV periods produce smoother folded light curves and thus are more likely to be correct. The objects are (top to bottom): BC CVn, GZ Com, V0533 Hya, BE Boo, UW CVn, V0593 Vir.

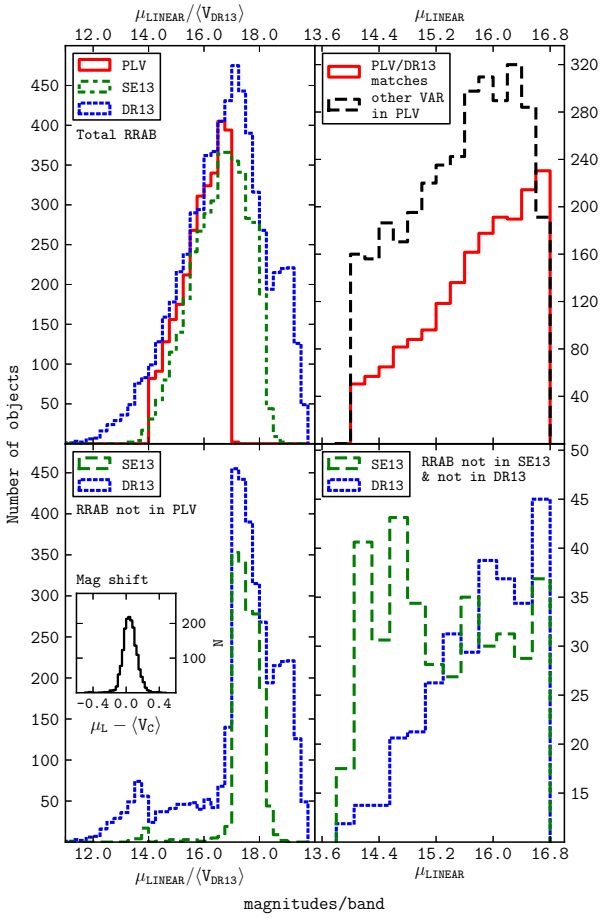

Fig. 23.— Comparison between PLV, Sesar et al. (2013) and DR13. The median LINEAR magnitude is designated as μ_{LINEAR} and mean DR13 V magnitude as $\langle V_{DR13} \rangle$. All four plots are made for the area in which PLV and DR13 have overlap (approximately $125° <$ R.A. $< 268°$ and $-13° <$ Dec $< 69°$). The histograms in the top left panel show the number of ab type RR Lyrae found in these three catalogs. The histograms in the bottom left panel show ab type RR Lyrae present in Sesar et al. (2013) and DR13, but not in the PLV catalog. The inset shows the difference in brightness for matched objects in the photometric systems used by LINEAR (unfiltered, μ_L) and DR13 (Johnson V band, $\langle V_C \rangle$). The histograms in the top right panel show the total number of matched objects between PLV and DR13, as well as the total number of other variable stars identified in PLV. The histograms in the bottom right panel show ab type RR Lyrae found by PLV and not listed in DR13 (line) and Sesar et al. (2013, dashed). **Better name for Mag Shift?**

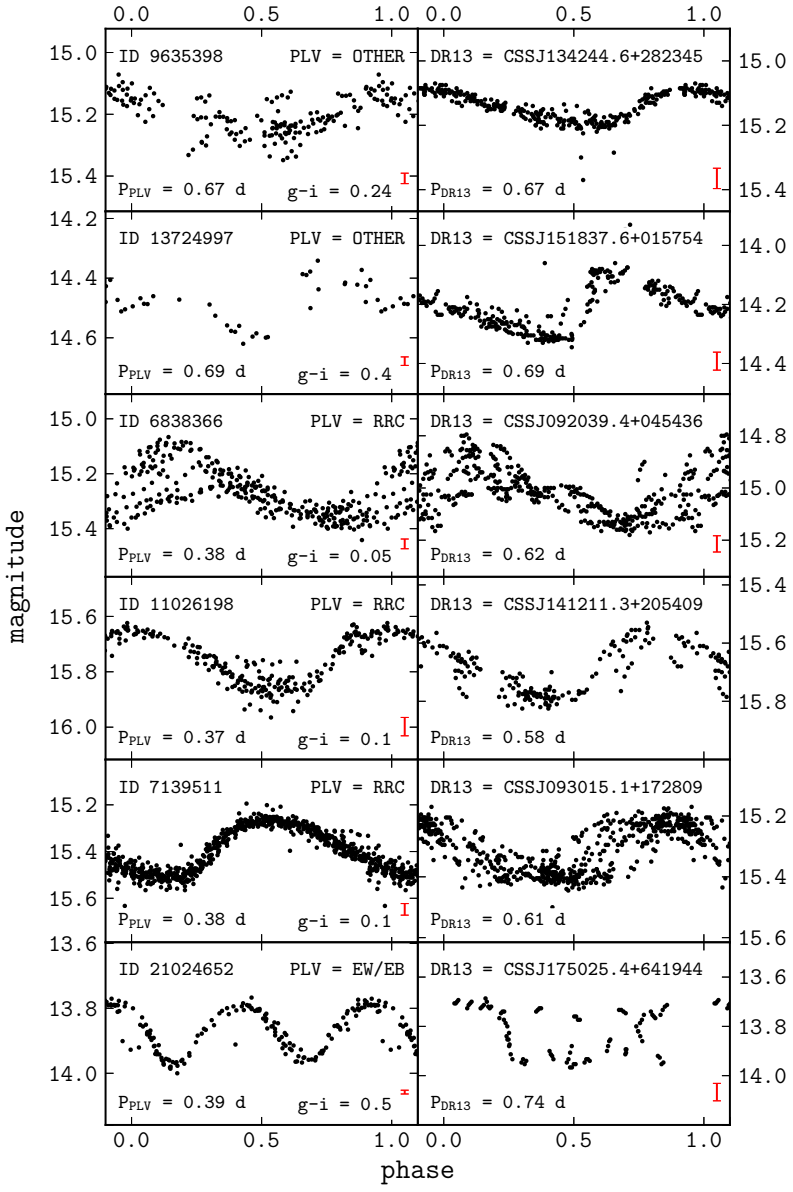

Fig. 24.— Examples of light curves of objects for which PLV and DR13 classification do not agree. Light curves on the left show LINEAR data folded with the PLV periods and figures on the right show DR13 data folded with the DR13 periods. The vertical red bars in each panel show the median errors (note that CRTS errors are larger than LINEAR errors). All of the objects were classified as ab type RR Lyrae in DR13.

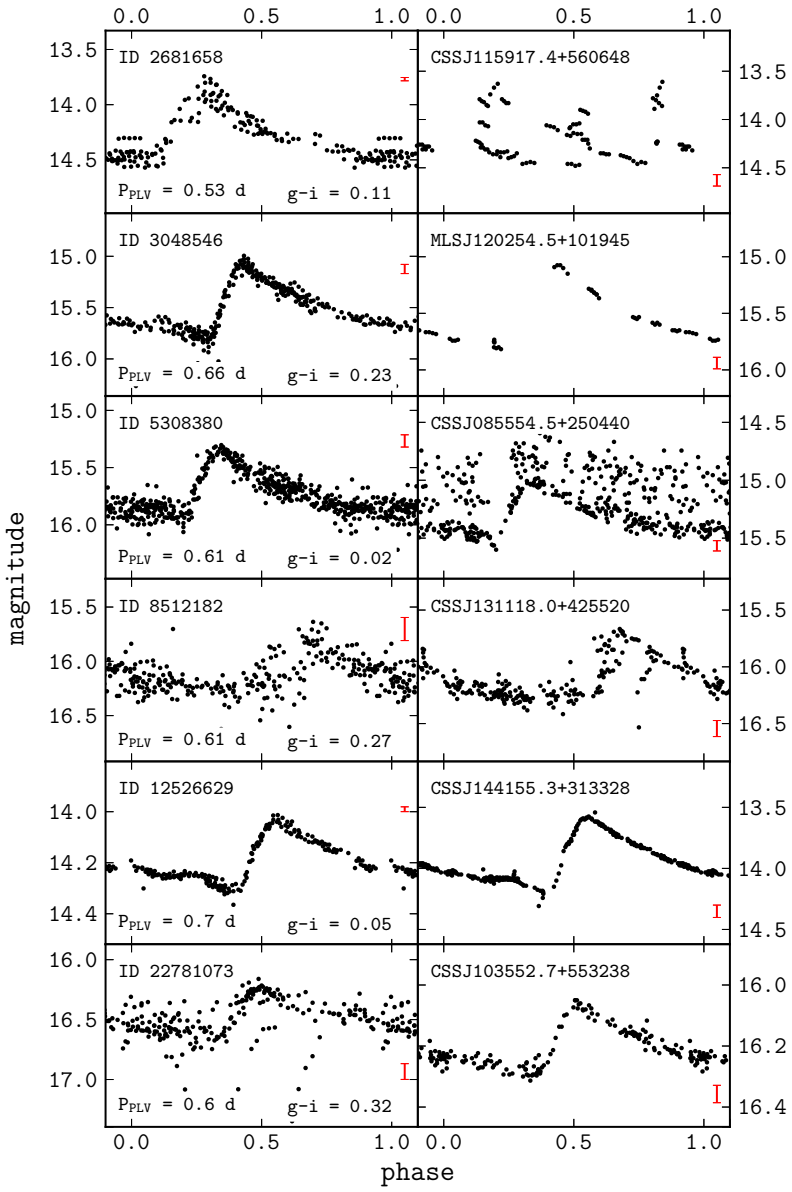

Fig. 25.— Examples of light curves of objects missing from DR13 but present in PLV. LINEAR data are shown in the left column and CSDR2 data in the right column. The PLV periods have been used to fold the light curves. The vertical red bars in each panel show the median errors.

Table 1. The classification statistics for the SDSS Stripe 82 subsample. Class is defined by the $A1$ range, the mean classification grade among eight classifiers, specified in the second and third columns. The standard deviation among the eight classifiers is listed in the fourth column, and the fifth column lists the number of light curves in each class. The sixth column lists the median χ^2 per degree of freedom, and the last column is the median robust χ^2 per degree of freedom (5% of the most outlying points are excluded from the computation).

Class	$A1_{min}$	$A1_{max}$	σ_{A1}	N	$< \chi^2_{dof} >$	$< R\chi^2_{dof} >$
0	0.0	1.1	0.38	7606	9.1	1.9
1	1.1	1.2	0.60	75	35.4	5.9
2	1.2	1.8	0.73	46	24.8	4.7
3	1.8	3.0	0.15	317	28.1	13.3

Table 2. The rectangular boundaries in the period-amplitude-skewness-color space used for classification. The boundaries were iteratively tuned to maximize the ratio of correctly selected and classified objects (with respect to the visual classification).

Type	log(P) [d]	log(A) [mag]	skewness	g-i
ab RR Lyr	$\langle -0.36, -0.05 \rangle$	$\langle -0.55, 0.05 \rangle$	$\langle -1.2, 0.2 \rangle$	$\langle -0.42, 0.5 \rangle$
c RR Lyr	$\langle -0.59, -0.36 \rangle$	$\langle -0.55, -0.15 \rangle$	$\langle -0.4, 0.35 \rangle$	$\langle -0.20, 0.35 \rangle$
single min	> -0.6	$\langle -0.7, 0 \rangle$	$\langle 0.32, 3.6 \rangle$	$\langle -0.2, 3 \rangle$
Algol	> -0.6	$\langle -0.67, 0.14 \rangle$	$\langle 1, 3.7 \rangle$	$\langle -1.2, 3.8 \rangle$
β Lyr & W UMa	$\langle -0.67, -0.4 \rangle$	$\langle -0.56, -0.09 \rangle$	$\langle -0.1, 1.6(3.2) \rangle$	$\langle 0.1(0.2), 1.8 \rangle$
SX Phe/δ Sct	$\langle -1.38, -1.05 \rangle$	$\langle -0.63, -0.12 \rangle$	$\langle -1.0, 0.7 \rangle$	$\langle -0.5, 0.2 \rangle$

Table 3. The main light-curve classification results. The first column lists a numerical class name used in the public catalog, and the second column lists its more descriptive names. The number of visually confirmed variable stars, with purity exceeding 99% (RR Lyrae, SX Phoenicis/δ Scuti), and 98% in case of eclipsing binaries, is listed in the fourth column, and the fraction of all cataloged objects in a given class is listed in the third column. Class "SM" corresponds to flat light curves with a single minimum, and class "Other" contains periodic variables which could not be reliably classified and non–periodic variables.

Class	Type	F [%]	N
1	RRAB	41	2923
2	RRC	14	990
3	SM	<1	20
4	EA	5	357
5	EB/EW	33	2385
6	SXP/DSCT	2	112
7	LPV	1	77
8	Hearbeat	<1	1
9	BL Her	<1	6
11	ACEP	<1	5
0	Other	4	318
	Total	100	7194

Table 4: The statistics for classes determined by Gaussian mixture modeling[a]

#	$u-g$	$g-i$	$J-K$	$\log(P)$	amplitude	skewness
1	1.23 ± 0.05	0.06 ± 0.12	0.24 ± 0.14	-0.48 ± 0.06	0.43 ± 0.07	0.01 ± 0.28
2	1.20 ± 0.05	0.35 ± 0.07	0.31 ± 0.15	-0.24 ± 0.06	0.70 ± 0.18	-0.47 ± 0.35
3	1.24 ± 0.05	0.06 ± 0.14	0.32 ± 0.17	-0.24 ± 0.05	0.70 ± 0.18	-0.42 ± 0.39
4	1.53 ± 0.33	0.83 ± 0.27	0.52 ± 0.14	-0.83 ± 0.07	0.50 ± 0.13	0.62 ± 0.29
5	1.65 ± 0.39	0.90 ± 0.29	0.59 ± 0.28	-0.91 ± 0.00	1.80 ± 0.68	0.28 ± 0.06

[a] The table lists statistics for the five most significant clusters out of 20 clusters identified in the seven-dimensional space of attributes (see Figures 17 and 18). The five clusters are identified as c type RR Lyrae (# 1), ab type RR Lyrae (# 2 and # 3), eclipsing binaries, and a set of objects with spurious periods.

Table 5. PLV object catalog.

objectID	Type	CUF	P	A	RA	DEC	objtype	mag_median
2522	5.0	1	0.238812	0.6764	117.988629	48.67131	6	16.999

Table 6. PLV object catalog.

objectID	uMod	gMod	rMod	iMod	zMod	uErr	gErr	rErr	iErr	
2522	1	20.492	18.435	17.418	16.991	16.751	0.062	0.007	0.006	0.006

Table 7. PLV object catalog.

objectID	rExt	J	H	K	JErr	HErr	KErr
2522	0.139	15.496	14.968	14.775	0.058	0.09	0.097

Table 8. PLV object catalog.

objectID	W1mag	W2mag	W3mag	W4mag	e_W1mag	e_W2mag	e_W3mag	e_W4mag
2522	14.516	14.63	12.599	8.843	0.034	0.066	blank	blank

Table 9. PLV object catalog.

objectID	noSATUR	stdev	rms	chi2pdf	nObs	skew	kurt	Rchi2pdf	Type2
2522	0.139	0.224	0.25	3.488	225	0.753	0.107	2.073	0