

## MIT Open Access Articles

*Exploring the variable sky with linear.  
III. classification of periodic light curves*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Palaversa, Lovro, Zeljko Ivezić, Laurent Eyer, Domagoj Ruzdjak, Davor Sudar, Mario Galin, Andrea Kroflin, et al. "Exploring the Variable Sky with Linear. III. Classification of Periodic Light Curves." *The Astronomical Journal* 146, no. 4 (September 16, 2013): 101. © 2013 The American Astronomical Society

**As Published:** <http://dx.doi.org/10.1088/0004-6256/146/4/101>

**Publisher:** IOP Publishing

**Persistent URL:** <http://hdl.handle.net/1721.1/92739>

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of Use:** Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



## EXPLORING THE VARIABLE SKY WITH LINEAR. III. CLASSIFICATION OF PERIODIC LIGHT CURVES

LOVRO PALAVERSA<sup>1</sup>, ŽELJKO IVEZIĆ<sup>2,3,4</sup>, LAURENT EYER<sup>1</sup>, DOMAGOJ RUŽDIJAK<sup>4</sup>, DAVOR SUDAR<sup>4</sup>, MARIO GALIN<sup>5</sup>,  
ANDREA KROFLIN<sup>3</sup>, MARTINA MESARIĆ<sup>3</sup>, PETRA MUNK<sup>3</sup>, DIJANA VRBANEC<sup>3</sup>, HRVOJE BOŽIĆ<sup>4</sup>, SARAH LOEBMAN<sup>2</sup>,  
BRANIMIR SESAR<sup>6</sup>, LORENZO RIMOLDINI<sup>1,7</sup>, NICHOLAS HUNT-WALKER<sup>2</sup>, JACOB VANDERPLAS<sup>2</sup>, DAVID WESTMAN<sup>2</sup>,  
J. SCOTT STUART<sup>8</sup>, ANDREW C. BECKER<sup>2</sup>, GREGOR SRDOČ<sup>9</sup>, PRZEMYSŁAW WOZNIAK<sup>10</sup>, AND HAKEEM OLUSEYI<sup>11</sup>

<sup>1</sup> Observatoire Astronomique de l'Université de Genève, 51 chemin des Maillettes, CH-1290 Sauverny, Switzerland; [lovro.palaversa@unige.ch](mailto:lovro.palaversa@unige.ch)

<sup>2</sup> Department of Astronomy, University of Washington, P.O. Box 351580, Seattle, WA 98195-1580, USA

<sup>3</sup> Department of Physics, Faculty of Science, University of Zagreb, Bijenička cesta 32, 10000 Zagreb, Croatia

<sup>4</sup> Hvar Observatory, Faculty of Geodesy, Kačićeva 26, 10000 Zagreb, Croatia

<sup>5</sup> Faculty of Geodesy, Kačićeva 26, 10000 Zagreb, Croatia

<sup>6</sup> Division of Physics, Mathematics, and Astronomy, Caltech, Pasadena, CA 91125, USA

<sup>7</sup> ISDC Data Centre for Astrophysics, Université de Genève, chemin d'Ecogia 16, CH-1290 Versoix, Switzerland

<sup>8</sup> Lincoln Laboratory, Massachusetts Institute of Technology, 244 Wood Street, Lexington, MA 02420-9108, USA

<sup>9</sup> Saršoni 90, 51216 Viškovo, Croatia

<sup>10</sup> Los Alamos National Laboratory, 30 Bikini Atoll Road, Los Alamos, NM 87545-0001, USA

<sup>11</sup> Florida Institute of Technology, Melbourne, FL 32901, USA

Received 2013 June 9; accepted 2013 July 30; published 2013 September 16

### ABSTRACT

We describe the construction of a highly reliable sample of  $\sim 7000$  optically faint periodic variable stars with light curves obtained by the asteroid survey LINEAR across  $10,000 \text{ deg}^2$  of the northern sky. The majority of these variables have not been cataloged yet. The sample flux limit is several magnitudes fainter than most other wide-angle surveys; the photometric errors range from  $\sim 0.03 \text{ mag}$  at  $r = 15$  to  $\sim 0.20 \text{ mag}$  at  $r = 18$ . Light curves include on average 250 data points, collected over about a decade. Using Sloan Digital Sky Survey (SDSS) based photometric recalibration of the LINEAR data for about 25 million objects, we selected  $\sim 200,000$  most probable candidate variables with  $r < 17$  and visually confirmed and classified  $\sim 7000$  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 an unsupervised machine learning approach. The resulting sample of periodic LINEAR variables is dominated by 3900 RR Lyrae stars and 2700 eclipsing binary stars of all subtypes and includes small fractions of relatively rare populations such as asymptotic giant branch stars and SX Phoenicis stars. We discuss the distribution of these mostly uncataloged variables in various diagrams constructed with optical-to-infrared SDSS, Two Micron All Sky Survey, and *Wide-field Infrared Survey Explorer* photometry, and with LINEAR light-curve features. We find that the 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 ( $\beta$  Lyrae and W UMa): as the color-based spectral type varies from K4 to F5, the median period increases from 5.9 hr to 8.8 hr. These large samples of robustly classified variable stars will enable detailed statistical studies of the Galactic structure and physics of binary and other stars and we make these samples publicly available.

**Key words:** binaries: eclipsing – blue stragglers – catalogs – Galaxy: halo – stars: statistics – stars: variables: general

**Online-only material:** color figures, machine-readable and VO tables

### 1. INTRODUCTION

Variability is an important phenomenon in astrophysical studies of structure and evolution, in stellar, Galactic, and extragalactic realms. 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. 2008b), 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; Dubath et al. 2011; Richards et al. 2011). Most such methods require reliable training samples; in addition to the astrophysical motivation for improved understanding of

the optical variability of faint sources, a goal of the analysis presented here is to construct a large training sample of periodic variable stars that probes both a large sky area and a faint magnitude range.

This paper is the third in a series based on light-curve data collected by the LINEAR (Lincoln Near-Earth Asteroid Research) asteroid survey over a period roughly from 1998 to 2009. In the first paper (hereafter Paper I; Sesar et al. 2011), we described the LINEAR survey and photometric recalibration based on SDSS stars acting as a dense grid of standard stars. In the overlapping  $\sim 10,000 \text{ deg}^2$  of sky between LINEAR and SDSS, photometric errors range from  $\sim 0.03 \text{ mag}$  for sources not limited by photon statistics to  $\sim 0.20 \text{ mag}$  at  $r = 18$  (here,  $r$  is the SDSS  $r$ -band magnitude). LINEAR data provide time domain information for the brightest 4 mag of the SDSS survey, with

250 unfiltered photometric observations per object on average (rising to  $\sim 500$  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;  $\sim 5$  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 Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) catalog entries are also available for the entire sample. In this work, we also provide positional matches to the *Wide-field Infrared Survey Explorer* (WISE) catalog entries (Wright et al. 2012) for confirmed periodic variables.

In Paper I, we compared the LINEAR dataset with other prominent contemporary wide-area variability surveys comparable 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 the distance limit for Galactic structure studies based on RR Lyrae stars by a factor of four (to about  $\sim 30$  kpc; see the second paper in this series for details, hereafter Paper II; Sesar et al. 2013). Thanks to the improved faint limit, the sample includes over a thousand quasars (for  $r < 17$ ; for a detailed analysis see Ruan et al. 2012). The large sky area, with a resulting increase in sample size, enables robust statistical studies of samples such as eclipsing binary stars and searches for rare objects (e.g., field SX Phoenixis (SX Phe) stars and asymptotic giant branch (AGB) 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 of samples of selected variable objects need to be understood and quantified. There are a number of automated methods proposed in the literature for selecting variable objects and classifying their light curves (e.g., Eyer & Blake 2005; Debosscher et al. 2007; Dubath et al. 2011; Richards et al. 2011, and references therein). Measuring the performance of these methods in the LINEAR dataset requires a reliable training sample and a full understanding of the photometric error distribution. It would be difficult to quantify the performance of these methods in the LINEAR dataset because there are no reliable training samples and the photometric error distribution is not yet fully understood. 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 in Paper I is close to a Gaussian, various tests show that of the order of 1% of measurements can have anomalous errors (defined here as errors at least three times larger than the reported errors) that are hard to recognize using the available metadata (such as photometric zeropoint information and the photometric scatter for calibration stars). This problem could be explained by the acquisition of data in non-photometric conditions (e.g., thin clouds or haze). A part of the problem may also be the fact that a large fraction of observations were obtained along the ecliptic where contamination by blended main belt asteroids is non-negligible.

Despite the fraction of measurements with anomalous errors being as small as 1%, the resulting sample contamination can

still 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 a practical cutoff in rms is about 0.1 mag for the LINEAR dataset and excluding quasars that are not numerous at the magnitudes probed by LINEAR (fewer than 0.1% of objects in the LINEAR sample with  $r < 18$  are quasars), robustly detectable variability is 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.

The LINEAR observing strategy produces repeat photometric data for stars on several timescales ranging from 15 to 20 minute intervals between images within a frameset to a few days between repeat visits during one lunation to the month-long timescale between lunar months to yearly timescales. More details on the sampling patterns can be found in Appendix A of Paper I.

In order to better understand the behavior of photometric errors in the LINEAR sample and to ultimately enable the 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 the visual classification and the construction of the resulting sample of about 7000 robust periodic variables are described in Section 2. The distribution of periodic variables, dominated by roughly equal fractions of RR Lyrae stars and eclipsing binary stars, in various color-color diagrams and other diagrams, is discussed in Section 3. We compare our results with existing variable star catalogs in Section 4 and with supervised and unsupervised machine learning classification methods in Section 5. Our main results are discussed and summarized in Section 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 the phased light curves, eight human classifiers extracted about 7000 likely periodic variables from a starting set of about 200,000 candidate variables and also obtained initial light-curve classifications. In the following two steps, a single expert refined the selection and classification of the smaller sample of 7000 likely periodic variables, first by repeating the visual classification and then by further refining the candidate sample by adding into the classification procedure the parameters measured from the light curves and other information such as multicolor photometry. In this section, we first describe the initial sample selection and period estimation and then we discuss the visual classification procedures in detail. 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<sup>12</sup> using the following criteria:

<sup>12</sup> Available at <https://astroweb.lanl.gov/lineardb/>

1. *Brightness limit.*  $14.5 < \langle m_{\text{LINEAR}} \rangle < 17$ , where  $\langle m_{\text{LINEAR}} \rangle$  is the median value of the white-light LINEAR magnitude.<sup>13</sup>
2. *Likely variability.*  $\chi^2_{\text{dof}} > 3$ , where the  $\chi^2$  per degree of freedom is computed using the unweighted mean magnitude and the photometric errors reported in the database.
3. *Variability amplitude.*  $\sigma > 0.1$  mag, where  $\sigma$  is the rms scatter (standard deviation) of the recalibrated LINEAR magnitudes.

The majority of about 200,000 selected objects are found in the region bounded by  $125^\circ < \text{R.A.} < 268^\circ$  and  $-13^\circ < \text{decl.} < 69^\circ$  (corresponding to the North Galactic Cap scanned by SDSS). An additional  $\sim 8000$  objects are found in the SDSS Stripe 82 region ( $-50^\circ < \text{R.A.} < 60^\circ$  and  $|\text{decl.}| < 1^\circ 266$ ). The selected objects contain both true variable objects and spurious candidates. We limit our classification to objects exhibiting mono-periodic variability (light curves  $m(t)$  that satisfy  $m(t + P) = m(t)$ , where  $P$  is the period and  $t$  is positive; assuming no noise) and use phased light curves for visual inspection. Phased light curves are constructed by plotting  $m(t)$  as function of phase:

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

where the function  $\text{int}(x)$  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 an implementation of the Supersmoother algorithm (Friedman 1984; Reimann 1994). This non-parametric method smooths the light curve using a variable smoothing length and uses a 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 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 Section 2.3.7). For this reason, we also included two additional algorithms for estimating periods: the Lomb–Scargle (LS) and the 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 the 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 therefore 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 (e.g., Sesar et al. 2010) or to use multiple harmonics in the Fourier expansion, which we have not considered here (e.g., Figures 4 and 5).

## 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 2500 light curves between the samples of the “adjacent” classifiers were used to verify classification consistency (which was assessed as described in Sections 2.3.2 and 2.3.4).

### 2.3.1. Initial Visual Classification

The initial visual classification was performed using the 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 the predicted classes of variable objects. Classifiers answered three questions with fixed possible answers.

The first question was whether the displayed phased light curves have a “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), and “definitely yes” (3). Unless the answer to the first question is “definitely no,” classifier proceeds to the second question related to 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), and “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 s on average to answer all three questions, for a throughput of  $\sim 700$  objects  $\text{hr}^{-1}$  (about a week’s worth of full-time work per classifier or about two full-time-equivalent person months for the whole effort, assuming 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 8044 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 could also be performed (described further below).

<sup>13</sup> The faint magnitude limit adopted in Paper II is 0.5 mag fainter than that adopted here because the ab type RR Lyrae discussed in Paper II are easier to recognize than other types of variable objects discussed here.



```
% READCOL: 317 valid lines read
```

```
=====
Question 1:
Does the light curve looks "good" for ANY period?
That is, does the scatter of data points
around the medians, shown by the yellow line,
looks "coherent"?
=====
```

```
The answers can be:
0: definitely no
1: probably no, but not sure
2: probably yes, but not sure
3: definitely yes
=====
```

```
Your response: 3
% READCOL: Format keyword not supplied - All columns
% READCOL: Skipping Line 1
% READCOL: 5 valid lines read
% READCOL: 509 valid lines read
% READCOL: 317 valid lines read
% READCOL: 511 valid lines read
% READCOL: 309 valid lines read
% READCOL: 474 valid lines read
=====
```

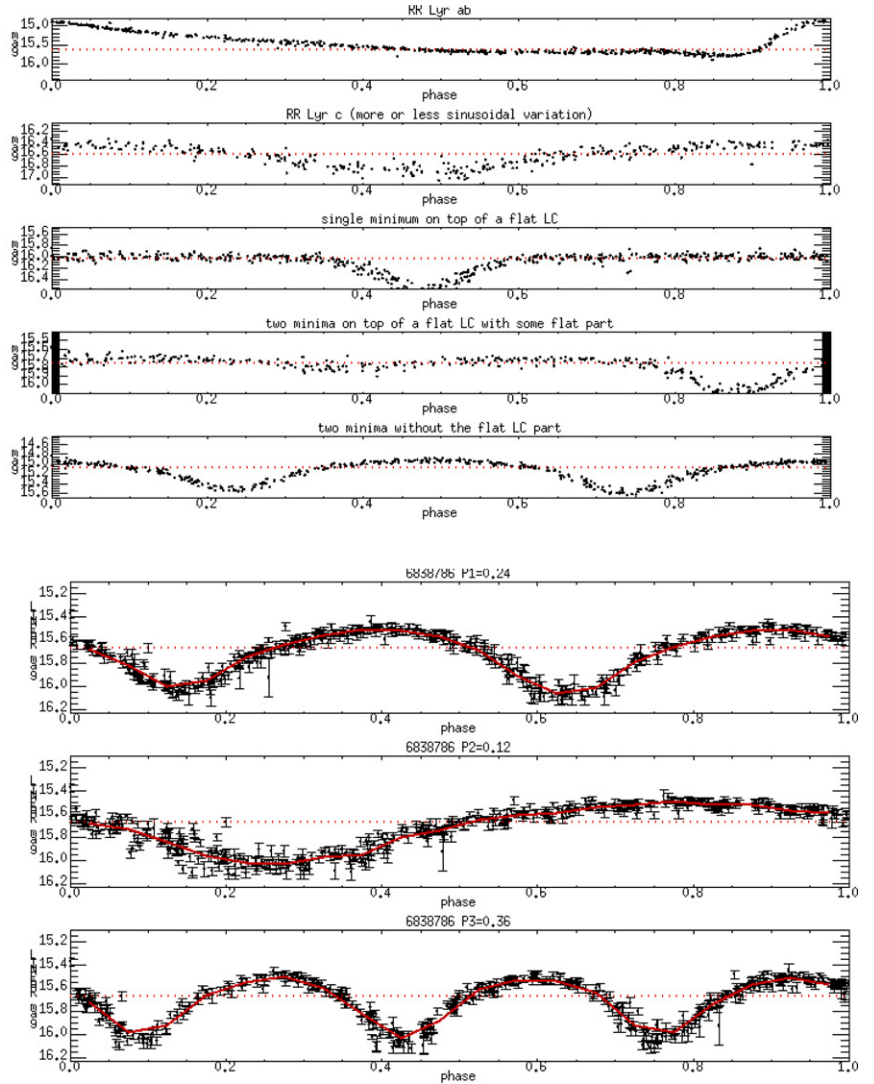
```
Question 2:
What type of light curve does the object
appear to have?
=====
```

```
The answers can be:
0: does not look like any template
1: RR Lyr ab (fast rise, slow fading,
  amplitudes about 0.2-0.8 mag)
2: RR Lyr c (more or less sinusoidal variation)
3: single minimum on top of a flat LC
4: two minima on top of a flat LC with some
  flat part
5: two minima without the flat LC part
=====
```

```
Your response: 5
=====
```

```
Question 3:
Which period is the best period to use?
Options are 1,2,3 (default 1).
=====
```

```
Your response: 1
=====
```



**Figure 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 the classification. (A color version of this figure is available in the online journal.)

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,  $\sigma_{A1}$ , to quantify the dispersion in the 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% of 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 re-inspected all 438 light curves with  $A1 > 1.1$  (classes 1–3), as well as 1000 light curves from class 0 with the 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 unconvincing 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

**Table 1**  
The Classification Statistics for the SDSS Stripe 82 Subsample

Class	$A1_{\min}$	$A1_{\max}$	$\sigma_{A1}$	$N$	$\langle \chi^2_{\text{dof}} \rangle$	$\langle R \chi^2_{\text{dof}} \rangle$
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

**Notes.** 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  $\chi^2$  per degree of freedom and the last column lists the median robust  $\chi^2$  per degree of freedom (5% of the most outlying points are excluded from the computation).

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 $\chi^2$ Selection

The LINEAR light-curve database contains two values of  $\chi^2$ : the standard value and the so-called robust  $\chi^2$ ,  $R\chi^2$ , determined by excluding both the brightest and faintest 10% of the points from the computation (note that despite its name, the measured  $\chi^2$  does not follow the statistical  $\chi^2$  distribution expected for Gaussian photometric errors). The robust  $\chi^2$  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 completeness (that is, whether  $R\chi^2$ -based selection could be used instead of visual pruning of the candidate sample). If an  $R\chi^2 > 3$  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 the 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 re-inspected these puzzling cases and found that they all were 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 their 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  $\chi^2 > 3$ .

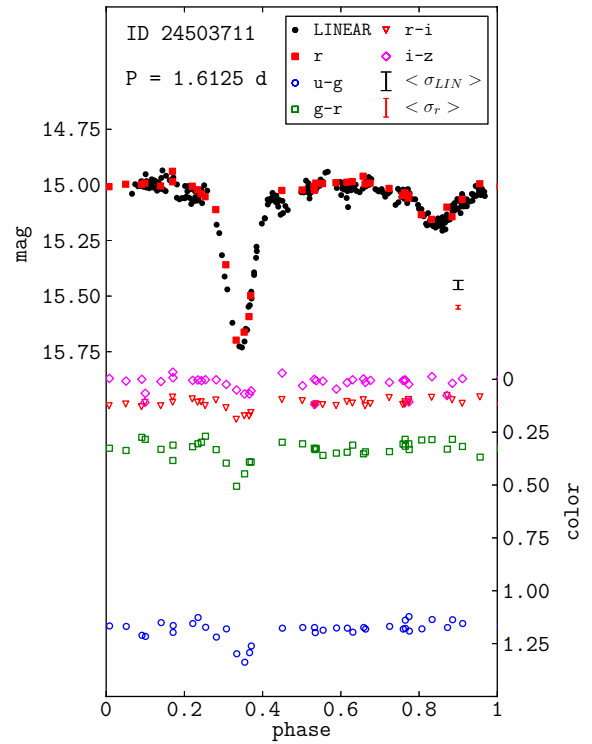
We have also re-inspected a random sample of light curves with  $A1 < 1.2$  and  $R\chi^2 > 3$ , that is, light curves that show significant variability according to  $R\chi^2$  but were not visually classified as periodic variables. About a half of these light curves show significant variability that 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 AGB stars. Therefore, their rejection from the periodic light-curve sample during visual classification is justified.

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

### 2.3.4. Comparison with the Variable Star Sample from the SDSS Stripe 82

SDSS has obtained multiple observations (about 50 on average) in the 300 deg<sup>2</sup> large, so-called Stripe 82 region. These data were used to select 67,507 candidate variable point sources<sup>14</sup> (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 the LINEAR sample because the former is much deeper ( $g < 20.5$  versus  $r < 17.5$ ) and has a more inclusive cutoff for variability rms (0.05 versus 0.1 mag). We have used this SDSS catalog to assess the reliability and completeness of candidate variables visually selected from the LINEAR database.

Out of the 8044 LINEAR objects found in the Stripe 82 region, 543 have positional matches within 2 arcsec with candidate



**Figure 2.** Example of LINEAR/SDSS synergy. The scale on the left corresponds to unfiltered LINEAR magnitudes and the scale on the right corresponds to SDSS colors. The top and bottom bars (black and red in the online version) show the 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.

(A color version of this figure is available in the online journal.)

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 the 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 the 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 five-band photometry.

About 45% of SDSS variables that are sufficiently bright to be in the LINEAR sample are not selected from the LINEAR database using the criteria listed in Section 2.1 and  $A1 > 1.2$  based on visual classification. About one third of those variables 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 the visual classification. Therefore, relative to the SDSS subsample limited to a similar depth, the completeness of the LINEAR sample is in the range of 55%–70%, depending on the adopted  $A1$  cut (most of the LINEAR incompleteness is due to the larger adopted minimum rms variability, 0.1 mag versus 0.05 mag).

<sup>14</sup> Light curves are publicly available from <http://www.astro.washington.edu/users/ivezic/sdss/catalogs/S82variables.html>

**Table 2**  
The Rectangular Boundaries in the Period–Amplitude–Skewness–Color Space Used for Classification

Type	$\log(P)$ (days)	$\log(A)$ (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	$> -0.6$	$\langle -0.67, 0.14 \rangle$	$\langle 1, 3.7 \rangle$	$\langle -1.2, 3.8 \rangle$
$\beta$ Lyr and W UMa	$\langle -0.67, -0.4 \rangle$	$\langle -0.56, -0.09 \rangle$	$\langle -0.1, 1.6 \rangle$	$\langle 0.1, 1.8 \rangle$
SX Phe/ $\delta$ Sct	$\langle -1.38, -1.05 \rangle$	$\langle -0.63, -0.12 \rangle$	$\langle -1.0, 0.7 \rangle$	$\langle -0.5, 0.2 \rangle$

**Note.** The boundaries were iteratively tuned to maximize the ratio of correctly selected and classified objects (with respect to the visual classification).

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%. An additional 57 objects have periods aliased by a factor of two 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 primary and secondary minima depths.

### 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 that 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 Sections 2.3.2 and 2.3.4. To improve the sample purity and the 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 eight subsamples were noticed, in agreement with the conclusions from the previous sections.

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 stars cluster in different regions of the multi-dimensional attribute space. Using selection boundaries based on the color, period, amplitude, and light-curve skewness listed in Table 2 and discussed in more detail in the next subsection (Section 2.3.6), 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 7194 visually selected and classified periodic variables. Among those, 6876 light curves (96%) have been assigned a definite type, while the remainder are classified as “Other.” The latter group contains objects that are variable, but not periodically, and objects for which 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 are listed in Table 3 and discussed in more detail in the next section. Hereafter, we refer to this entire sample as the “visually confirmed sample of periodic LINEAR

**Table 3**  
The Main Light-curve Classification Results

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	Heartbeat <sup>a</sup>	<1	1
9	BL Her	<1	6
11	ACEP	<1	5
0	Other	4	318
	Total	100	7194

**Notes.** The first column lists a numerical class name used in the public catalog and the second column lists its more descriptive name. The fraction of all cataloged objects in a given class is listed in the third column. The number of visually confirmed variable stars exceeding 98% is listed in the fourth column. Class “SM” corresponds to flat light curves with a single minimum and class “Other” contains periodic variables that could not be reliably classified and non-periodic variables. During the submission process we received information that 59 likely d type RR Lyrae can be found in the PLV (Poleski 2013).

<sup>a</sup> Candidate for a class of stars with tidally induced distortions and oscillations.

variables,” or simply the “PLV” sample. The resulting catalog is made publicly available.<sup>15</sup>

Table 3 quantitatively summarizes the results of the 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., it was either too shallow to be detected or too similar in depth to the primary minimum; recall Section 2.3). For that class of objects, correct periods could be twice as long as those listed in the catalog. The light curves classified as “5” include two types of eclipsing binaries: EB (or  $\beta$  Lyrae) and EW (W Ursae Majoris), which are grouped together because they are hard to distinguish using only LINEAR light curves. Motivated by the distribution in period–color and period–amplitude diagrams, we introduced two additional classes: class “6” (containing SX Phe

<sup>15</sup> Available from [http://www.astro.washington.edu/users/ivezic/r\\_data depot.html](http://www.astro.washington.edu/users/ivezic/r_data depot.html).



and  $\delta$  Scuti candidates) and class “7” (long-period variables defined here as variables with periods longer than 50 days and as semi-regular variables). Further explanations regarding the introduction of these two additional classes can be found in Sections 3.4 and 3.5.

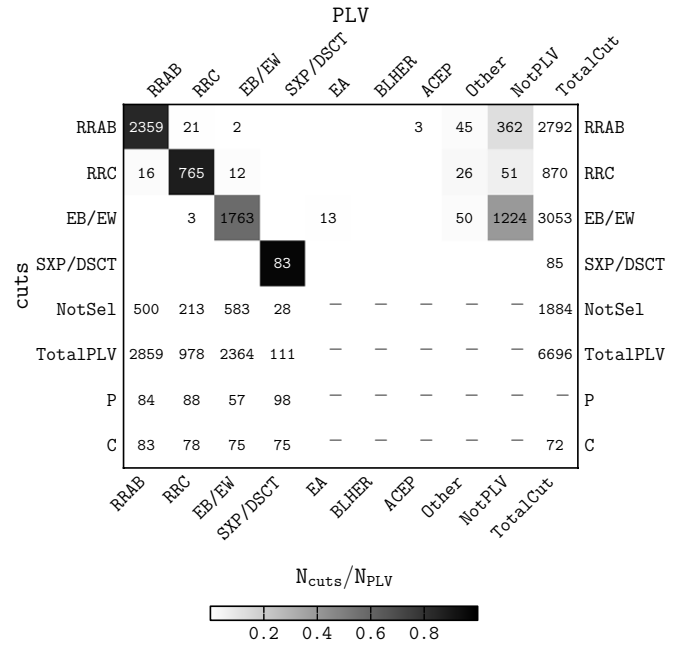
### 2.3.6. Simple Automated Classification with the Aid of Other Attributes

The clustering of objects in different regions of the multi-dimensional attribute space offers an opportunity to develop automated classification methods. Here, we define selection boundaries using simple, rectangular cuts in the four-dimensional attribute space (period, amplitude, skewness, and  $g - i$  color). Alternative approaches based on machine learning algorithms are discussed in Section 4. 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 Phe/ $\delta$  Scuti ( $\delta$  Sct) candidates. We do not include classes whose sizes fail to exceed 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 could easily have an ensemble of paired objects with presumably few common physical characteristics). An analogous diversity is expected among long-period variables that include both Miras and semi-regular variables and possibly other classes of variable stars. 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 Miras are actually fundamental mode pulsators—a physical characteristic that differentiates them from other long-period variables (e.g., Wood & Sebo 1996; Soszyński et al. 2009; Spano et al. 2011).

In order to maintain analysis uniformity, we use best-fit periods found by the classic LS method. Objects with unreliably measured SDSS colors and LS periods close to one day and half a day ( $\pm 0.05$  tolerance in  $\log(P)$ ) were excluded from the analysis. The performance of this supervised classification is statistically compared with our visual classification results in Figure 3. We have visually re-examined all 3270 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, were rejected by the period cut, or had at least one of the attributes outside the allowed interval. This selection fraction varies little among the four types (see the bottom row in Figure 3).

The automated selection method selected an additional 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 those 835 objects, 246 correspond to ab type RR Lyrae; the majority are located very close to the red cutoff for  $g - i$  color. Approximately 15% of these 246 objects have light curves hinting at ab type RR Lyrae, but are not of sufficient quality to enable a 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 (a 1.4% effect). In case of c type RR Lyrae, 44 objects not in the PLV catalog are uniformly distributed throughout the selection volume. About 30% of these objects have light curves that might be classified



**Figure 3.** Statistical performance comparison between the visually confirmed and classified variable sample listed in the final PLV catalog (7194 objects) and a simple supervised classification algorithm applied to the full sample of all 200,000 candidate 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 “Other” corresponds to variable PLV objects that do not belong to any of other chosen variability types and the “NotPLV” column corresponds to objects that satisfy the selection cuts applied to the full sample but that were not visually tagged as variable and included in the PLV catalog. The first four rows correspond to the four analyzed subsamples of variables defined by the applied selection cuts. The last column lists the total number of objects selected by each automated cut. The fifth row, labeled “NotSel,” corresponds to PLV objects not selected by automated selection cuts and the sixth row, labeled “TotalPLV,” gives the sum of the fifth row and the number of PLV objects of a given type correctly classified by the automated method. The intersection regions are color-coded by the fraction of objects in each row falling into a given region, that is, the fraction of selected objects with a type confirmed by the PLV catalog, that are also listed in the penultimate row. The last row, labeled “C,” lists the completeness of the automated selection method compared to the PLV catalog for the four analyzed variability classes.

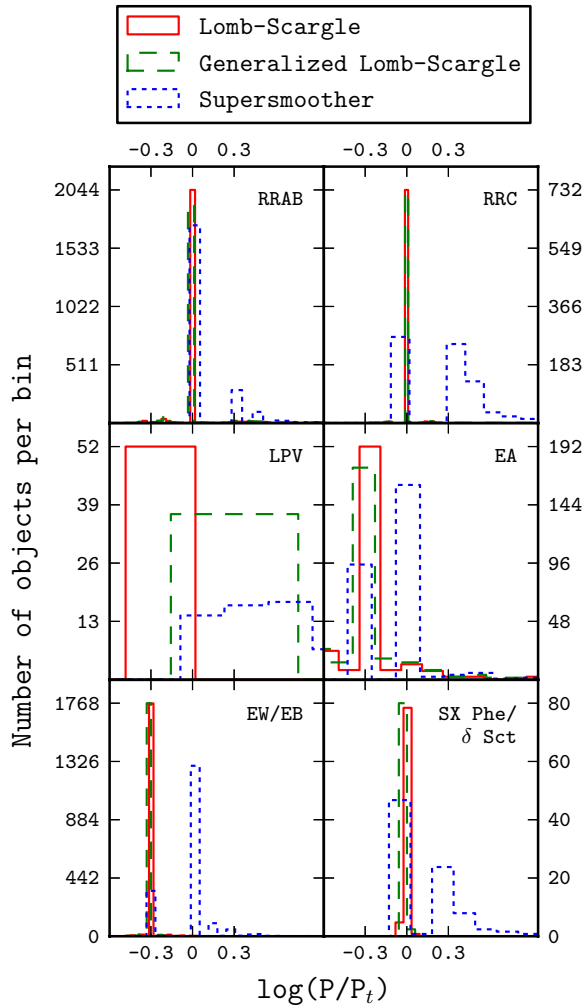
as c type RR Lyrae, although not reliably. A similar behavior is displayed in the EB/EW case, with only about 10% of 545 objects not in the PLV catalog being potentially classifiable as reliably periodic. Therefore, the PLV catalog is only slightly incomplete relative to the initial sample of 200,000 candidates (by about 1%–2% at most).

The automated classification is correct for a high fraction of PLV objects: 97% for ab type RR Lyrae, 78% for c type RR Lyrae, 87% for EB/EW, and 100% for SX Phe/ $\delta$  Sct. In summary, this analysis provides further support that the PLV catalog is highly complete relative to the initial sample of 200,000 candidate variables, has exceedingly low contamination, and has a high rate of correct light-curve classification.

### 2.3.7. 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: the primary and secondary minima are often of similar depths and are therefore often misidentified as the same feature in the phased light curve. This error, however, is not seen consistently: not all of the objects





**Figure 4.** Comparison of the three period finding methods, shown 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/ $\delta$  Sct variables, EW/EB eclipsing binaries ( $\beta$  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 two bias corresponds to 0.30 on a logarithmic scale). Note that the Lomb-Scargle methods consistently underestimate the period of EA and EW/EB light curves by a factor of two (this systematic effect has been corrected in the public catalog).

(A color version of this figure is available in the online journal.)

with similar minima depths have periods that are too short by a factor of two.

Given the final sample of 6876 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 4. We left 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 minima depths, but with periods that are too short by a factor of two. Another explanation could be that the secondary minima for these objects are too shallow to be detected in the LINEAR data.

Our results show that the LS and GLS methods typically outperform the Supersmoother algorithm for all variability types. For c type RR Lyrae, long-period variables, and SX Phe/ $\delta$  Sct 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 the histograms shown in Figure 4 is larger).

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

The reason for this consistent bias in period estimation by the LS and GLS 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 this method is more computationally expensive). As illustrated in Figure 5, a Fourier series model with six terms correctly recognizes the two minima in the light curve of an eclipsing binary star. For an additional discussion, please see Hoffman et al. (2009) and Wyrzykowski et al. (2003).

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 implementation might be more difficult since the values of period, amplitude, and skewness are in large part similar for c type RR Lyrae and EB and EW binaries. The addition of appropriate color information (e.g.,  $g - i$ ) easily breaks this degeneracy (see Sections 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.

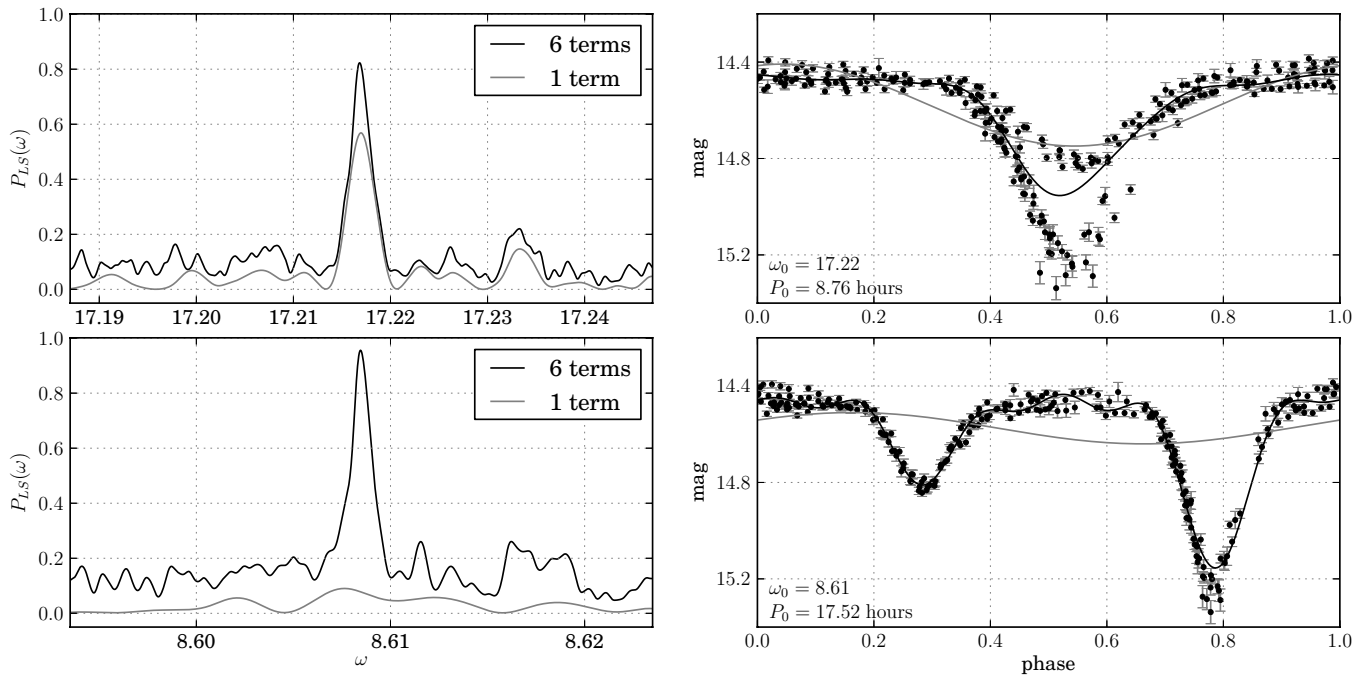
### 3. ANALYSIS OF PERIODIC LINEAR VARIABLES

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 a robust and efficient classification of objects into various classes of the variable population. These features are not seen in the full sample of 200,000 candidate variable objects and thus strongly suggest that visual classification successfully extracted the true variables.

We first discuss the distribution of classified variables in diagrams constructed with the three light-curve parameters and then we 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 of likely Population II field SX Phe stars (compared with the number of currently known objects), 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.



**Figure 5.** 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 a star with ID = 14752041 (the data and the *python* code to produce this figure, including the 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. This figure is adapted from Ivezić et al. (2013) and can be reproduced using code available at <http://www.astroML.org> (VanderPlas et al. 2012).

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, depending on the sampling, might not be 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 it quickly becomes very noisy), Fourier coefficients (which help greatly in classifying eclipsing binary subtypes; Pojmański 2002, or RR Lyrae subtypes; Soszyński et al. 2011), or even non-parametrically using 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 an informative description of the light-curve behavior. Nevertheless, exploring these other options would be a worthwhile analysis to undertake.

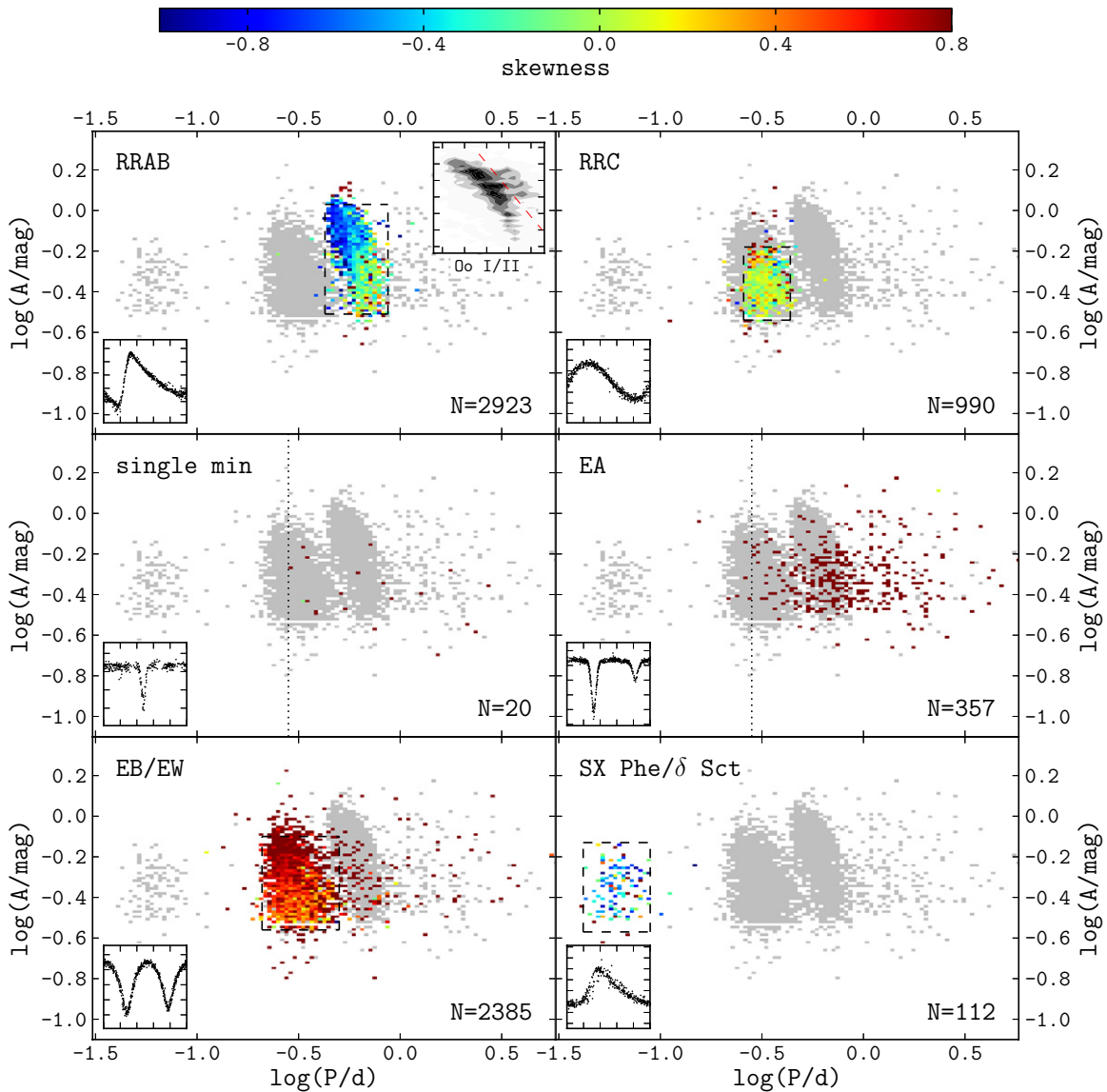
The distribution of variables in period–amplitude–skewness space is illustrated separately for each of the six main variability classes in Figure 6. The period distribution of the PLV sample is multi-modal, as further quantified in Figure 7. Even the period alone enables a remarkable, although not perfect, classification of periodic variables: SX Phe/ $\delta$  Sct candidates clearly stand out ( $P < 0.1$  days) 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 (i.e., the EB/EW class has a much larger skewness than c type RR Lyrae; compare the symbol color in the top-right and bottom-

left panels in Figure 6). In other words, the visual classification of light curves in essence reflects the distribution of these three parameters (and also the light-curve smoothness). We analyze the performance of automated classification methods based on this behavior in Section 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 6 (note also the strong correlation between amplitude, skewness, and period for ab type RR Lyrae). The average periods of the 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 the periods of RR Lyrae ab stars in Oosterhoff type I clusters are 0.1 days shorter than 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 a better performance from the 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 five (the three light-curve attributes and at least two color attributes).



**Figure 6.** Period–amplitude diagram for visually confirmed periodic LINEAR variables. Each panel represents a given class of variable stars confirmed by visual classification. The width of the bins is 0.03 in the color dimension and 0.02 in the  $\log(P/\text{day})$  dimension. Bins are color coded by the median value of skewness (per bin). The gray background corresponds to all PLV sample variables. The insets in the panels show a typical light curve for the variability type in the given panel.  $N$  is the total number of objects of a given type. The top-right inset in the “RRAB” panel shows the separation between Oosterhoff type I and type II RR Lyrae ab, where the former are to the left of and below the dashed line (red in the online version) and the latter are to the right of and above the dashed line. The gray map shows the density of ab type RR Lyrae per bin. The axes in the folded light-curve diagrams correspond to phase and magnitude. In the inset in the Oosterhoff diagram for RR Lyrae, the axes correspond to  $\log(P/\text{day})$  and  $\log(A/\text{mag})$ .

(A color version of this figure is available in the online journal.)

We emphasize that both SDSS and 2MASS photometry are single-epoch measurements obtained at random light-curve phases. Therefore, while the observed color range tracks the intrinsic color range of a given population, the distribution of objects *within* that range is affected by the color light-curve shape (i.e., ab type RR Lyrae stars spend more time close to minimum than to maximum light; since RR Lyrae are redder when fainter, their instantaneous color distribution is skewed redward compared to their mean color distribution).

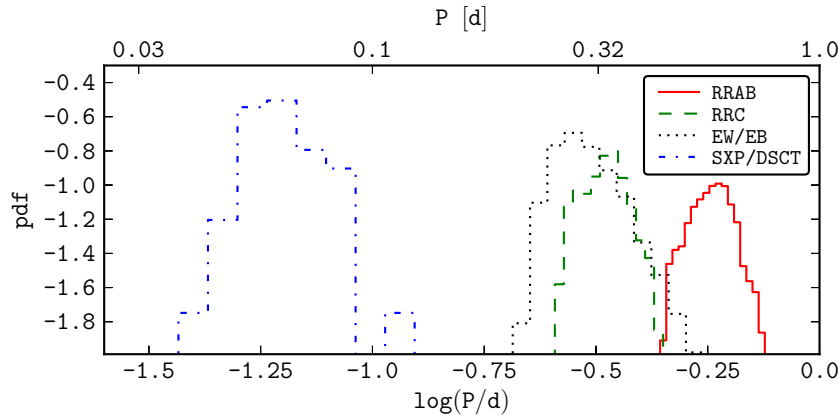
Figure 8 demonstrates that the addition of just one color to the period, here the SDSS  $g - i$  color, which is a good measure of the effective temperature (Ivezić et al. 2008a), helps to clearly separate c type RR Lyrae from EB/EW binaries. A more detailed illustration of the correlations between the  $g - i$  color and light-curve properties is shown in Figure 9. Note in particular how EA and EB/EW objects are well separated in this diagram. The

EB/EW subsample displays a good correlation between period and color, discussed in more detail in Section 3.3.

### 3.2.1. The $g - r$ versus $u - g$ Diagram

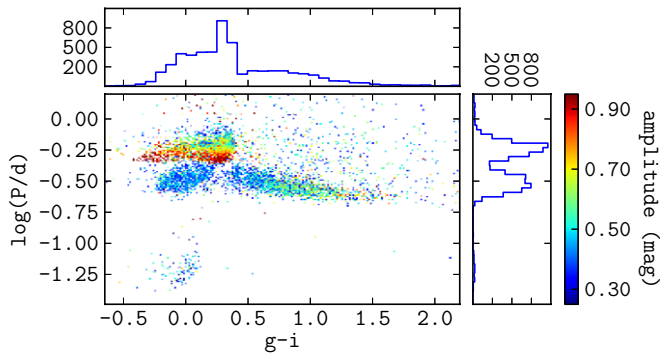
In addition to the three-dimensional  $g - i$  color–period–amplitude projection of the full multi-dimensional attribute space discussed above, the three-dimensional projection spanned by the SDSS  $u - g$  and  $g - r$  colors and light-curve skewness is also rich in content. The  $g - r$  versus  $u - g$  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, 2008a).

The distribution of variables in the  $g - r$  versus  $u - g$  versus skewness space is shown separately for each of the six main



**Figure 7.** Period distribution for the five most populous variability classes. While SX Phe/ $\delta$  Sct candidates clearly stand out ( $P < 0.1$  days), and ab type and c type RR Lyrae are fairly well separated by  $P = 0.4$  days, the eclipsing binaries overlap with the period range of RR Lyrae stars (especially the EW/EB type eclipsing binaries and the c type RR Lyrae).

(A color version of this figure is available in the online journal.)



**Figure 8.** Distribution of periodic variables in the color-period diagram. Bins are color coded by the median value of the light-curve amplitude according to the legend on the right. The two histograms show the marginal distributions of the period and the  $g - i$  color.

(A color version of this figure is available in the online journal.)

variability classes in Figure 10. As known from previous work based on SDSS data, the 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 10.

Based on the  $g - r$  versus  $u - g$  color-color diagram and the skewness distributions, we identified approximately 25% suspected misclassifications between c type RR Lyrae and EB/EW eclipsing binaries (from the first classification round) and visually re-inspected their light curves. We found that approximately 80% of those classifications were indeed likely incorrect and their type was subsequently revised. The cross-contamination of these two subsamples is easy to understand; a light curve of an eclipsing binary with similar minima depths can easily be misidentified as a nearly symmetric (sinusoidal) c type RR Lyrae light curve. This ambiguity is particularly problematic in the 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 versus period versus amplitude diagram and the  $g - r$  versus  $u - g$  versus 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 AGB 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

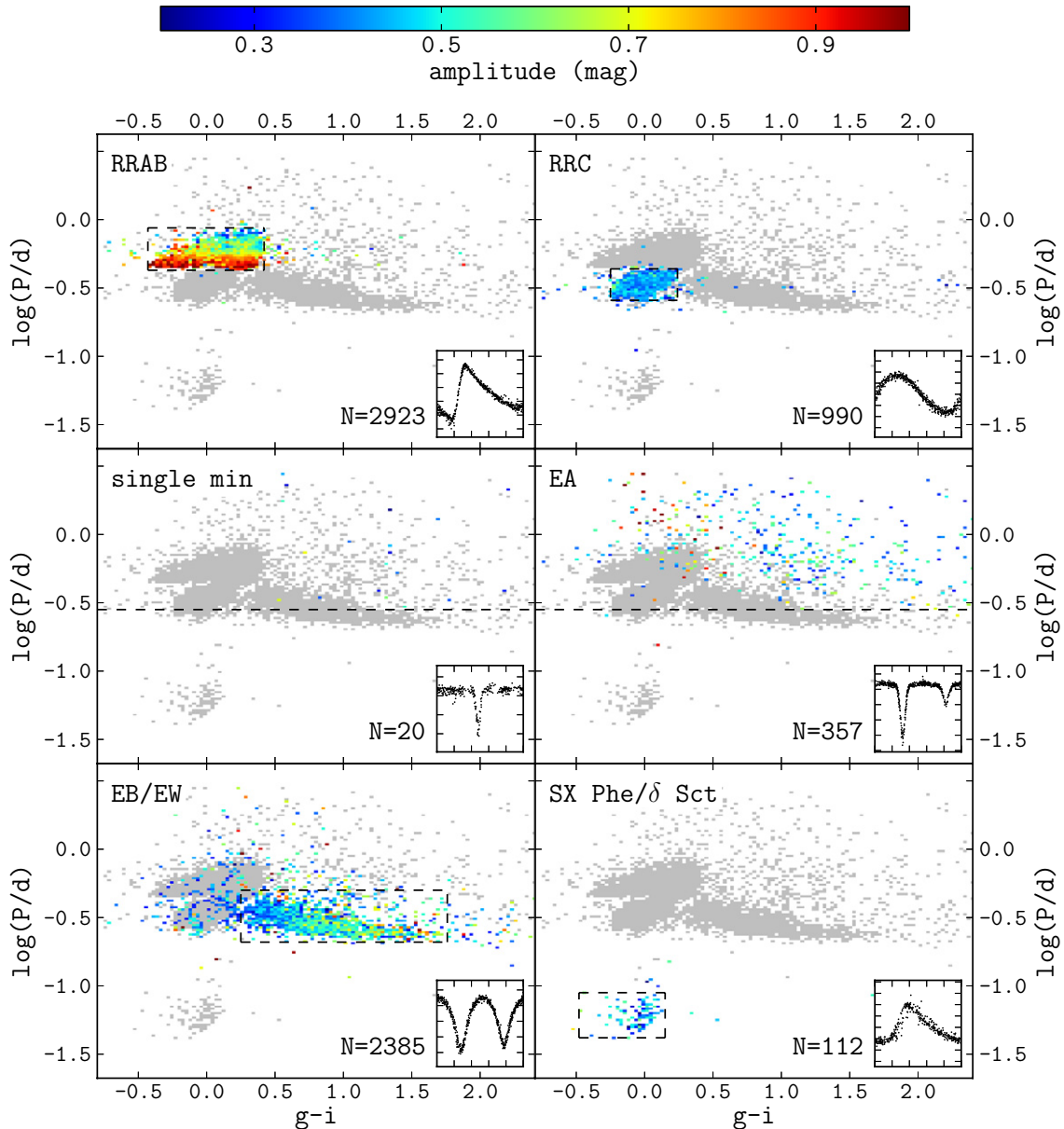
The distribution of EB ( $\beta$  Lyrae) and EW (W Ursae Majoris) eclipsing binary stars is remarkably well outlined in the period versus  $g - i$  color diagram (see the bottom left panel in Figure 9 and a zoomed-in version in Figure 11). Since the sample selection is primarily driven by the light-curve shapes and substantial selection effects in the  $g - i$  color and period in the relevant ranges are not expected, this 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 corresponds to spectral types from 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 Ursae Majoris (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  $g - i$  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:

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

Due to the large sample size, the random errors on 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 hr to 8.8 hr as the color-based spectral type varies from K4 to F5. An alternative form based on the Johnson  $B - V$  color, derived using transformations between the SDSS and Johnson systems from Ivezić et al.





**Figure 9.** Distribution of visually confirmed periodic LINEAR variables (PLV) in the color–period diagram. Each bin has been color coded by the median amplitude of objects inside it, according to the color bar above. The width of the bins is 0.03 in the color dimension and 0.02 in the  $\log(P/\text{day})$  dimension. The gray background represents all PLV sample variables. The insets in the panels represent a typical light curve for the variability type in that given panel.  $N$  is the number of objects of a given type. The 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 Section 2.3.6.

(A color version of this figure is available in the online journal.)

(2007b), is

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

and valid over the range  $0.3 < B - V < 1.1$ . This relation agrees well with a similar relation obtained by Rucinski (1997) for  $\sim 400$  W UMa stars observed by the OGLE project in Baade’s Window (note that we fit the median relation while 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).

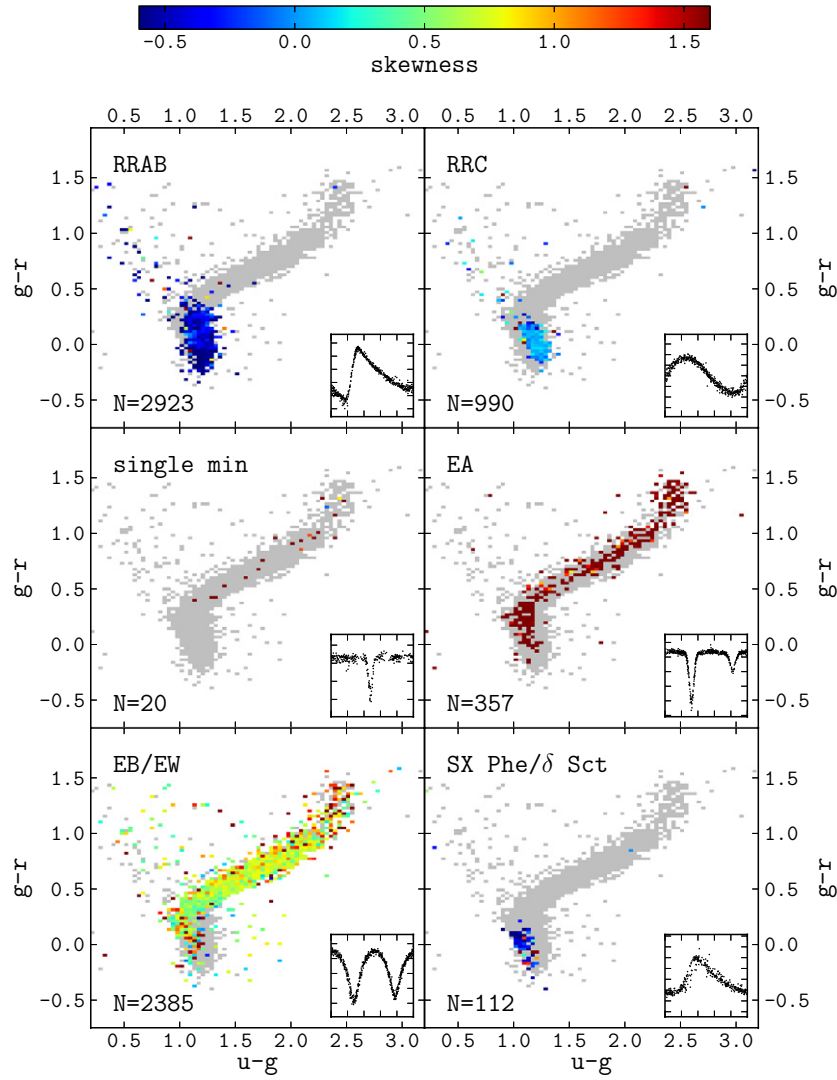
These findings are related to the fact that the period distribution for contact binary star systems appears to have a well-defined lower limit at 0.22 days (Rucinski 1992). More recent data show that this limit may be a bit smaller ( $\sim 0.20$  days; see Dimitrov & Kjurkchieva 2010; Davenport et al. 2013), but the existence 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 spectral type M0, we obtain a period of 0.22 days, in good agreement with other studies.

In Figure 12, we show several examples of these short-period binaries. Several objects have periods below 0.2 days and test the value of the aforementioned period boundary.

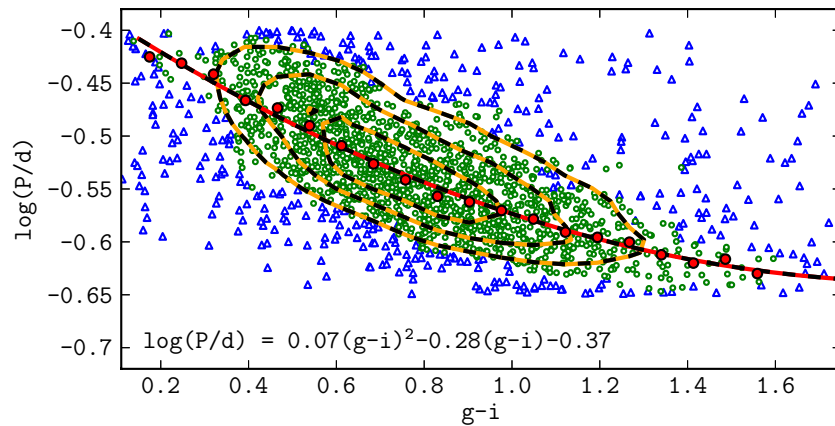
### 3.4. Candidate SX Phe Stars

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



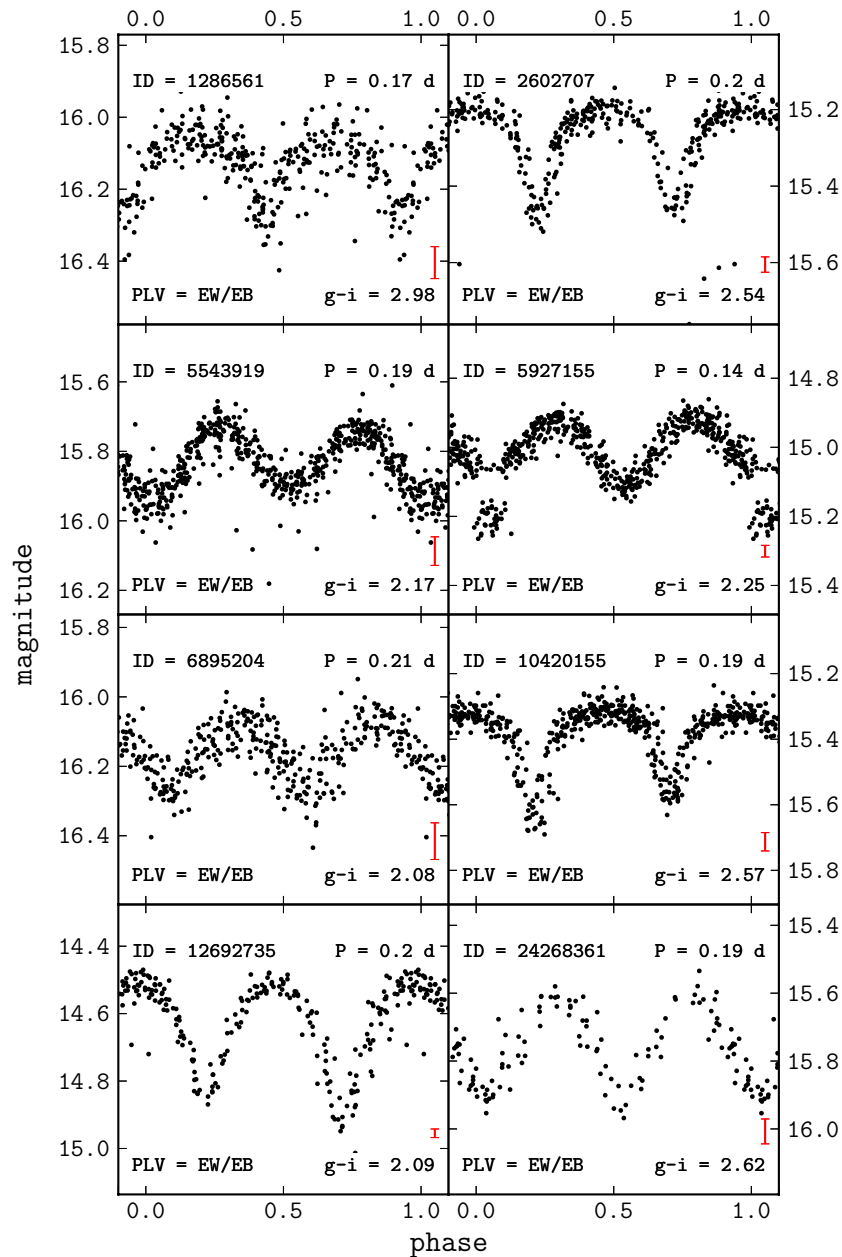
**Figure 10.** Analogous to Figure 9, except that the PLV sample distribution is shown in the SDSS  $u - g$  vs.  $g - r$  color-color diagram and the color coding is based on the light-curve skewness.

(A color version of this figure is available in the online journal.)



**Figure 11.** Quadratic fit to the correlation between the period and the color of EB/EW binaries. Selected objects were visually classified as EB/EW binaries and satisfy the following criteria:  $0.1 < g - i < 1.8$  and  $-0.67 < \log(P) < -0.4$ . Data were binned in 0.1 wide bins in  $(g - i)$  and 0.05 wide bins in  $\log(P)$ . Objects outside the 5% and 95% points of the distribution in color and period per bin were removed from the subsequent fitting procedure (triangles; blue in the online version). The remaining objects (circles; green in the online version) were again binned in 0.1 wide bins in  $(g - i)$  and the median of the logarithm of the period in days per bin was calculated (filled points; red in the online version). A quadratic function was fit to those points. Dashed contours designate areas of 5, 10, and 15 points  $\text{bin}^{-1}$ , i.e., the density of objects that passed the 5% and 95% cuts.

(A color version of this figure is available in the online journal.)



**Figure 12.** Examples of short-period contact binaries. Some periods are shorter than 0.2 days and test the value of the boundary mentioned in Section 3.3. Our most likely candidate for the eclipsing binary with the shortest period is in the top left corner. Vertical error bars show the typical photometric errors for each light curve. Note the unusual light curve of the object with LINEAR ID 5927155 (follow-up is in progress.).

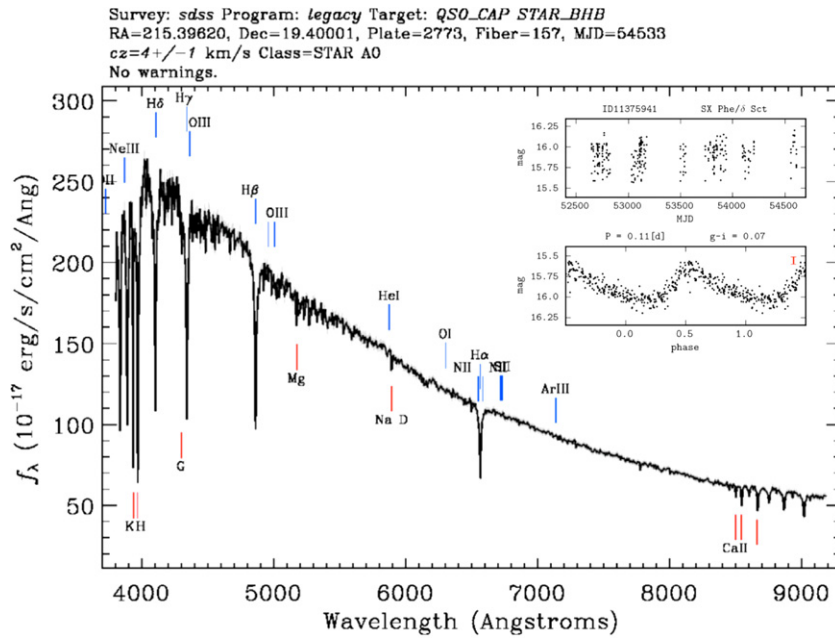
(A color version of this figure is available in the online journal.)

from Ivezić et al. 2007b) with very short periods (1–2.5 hr) and asymmetric light curves (see bottom-right panel in Figures 6 and 9). These stars can be identified as a mixture of  $\delta$  Sct and SX Phe stars (e.g., see Figure 8 in Eyer & Mowlavi 2008). Both types of stars are usually considered 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 the  $\delta$  Sct subsample belonging to Population I disk stars and the SX Phe subsample belonging 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 that all of these

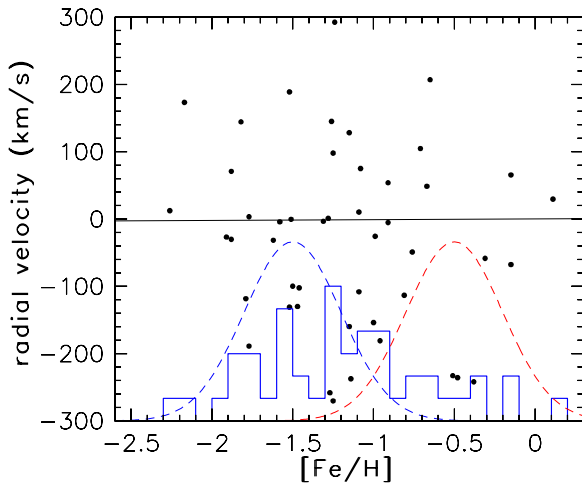
objects are likely radial mode pulsators. Given the apparent magnitude limits of our sample, the implied distances span the range 2–10 kpc, that is, many disk scale heights away, and thus SX Phe stars probably dominate because they are Population II (halo) objects. We note that the  $B - V$  color distribution 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  $g - i < -0.1$ ).

SDSS spectra for 34 stars from the SX Phe candidate sample are available. A much higher fraction of those spectra is consistent with the SX Phe hypothesis. All the spectra appear very uniform and characteristic for A stars; an example is shown in Figure 13. Using the default SDSS metallicity and radial velocity estimates (see Figure 14), we find that the sample is dominated



**Figure 13.** 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.

(A color version of this figure is available in the online journal.)



**Figure 14.** Radial velocity vs. metallicity for 34 candidate SX Phe stars with SDSS spectra (the repeated measurements for seven stars are also shown). The histogram shows the marginal distribution of the metallicity. The two Gaussians illustrate the expected metallicity distributions for halo stars (left) and disk stars (right), taken from Ivezić et al. (2008a). The metallicity is below the traditional boundary for separating halo and disk stars at  $[Fe/H] = -1.0$  dex for 57% of the measurements. For these stars, the radial velocity dispersion is  $135 \text{ km s}^{-1}$ , fully consistent with the halo hypothesis (Bond et al. 2010). Only the four stars with  $[Fe/H] > -0.5$  dex and small radial velocities are consistent with the disk hypothesis.

(A color version of this figure is available in the online journal.)

by stars with  $[Fe/H] < -1$ , low metallicities characteristic of halo stars, with a large velocity dispersion ( $134 \text{ km s}^{-1}$ ) that is also consistent with a presumed halo population (for a review of recent observational constraints on the differences between the metallicity and kinematic distributions of disk and halo stars; see, e.g., Ivezić et al. 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 & Breger 2001, there are only 17 known field SX Phe stars). 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 to explore sufficiently faint flux levels, over a large sky area, with an appropriate cadence. We are currently undertaking photometric and spectroscopic follow-up efforts to better characterize this sample.

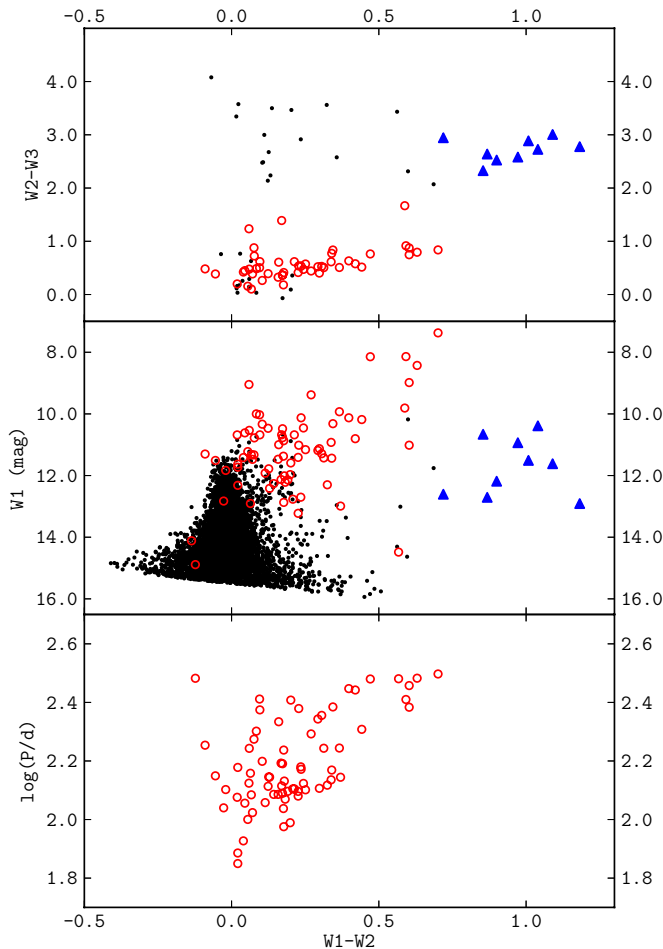
### 3.5. Candidate AGB Stars and WISE Color Distributions

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

The best available infrared sky survey was conducted by the recent *WISE* mission (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  $22 \mu\text{m}$  with  $5\sigma$  point source sensitivities better than 0.08, 0.11, 1, and 6 mJy (corresponding to Vega-based magnitudes of 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  $0''.15$ . *WISE* is photometrically calibrated to the Vega system and thus objects with infrared excesses should have colors greater than zero (not accounting for the measurement noise).

We have positionally matched the PLV catalog and the *WISE* catalog with a matching radius of 3 arcsec and obtained 7123 *WISE* matches for objects listed in the PLV catalog. Our analysis of this sample is shown in Figure 15. The distribution of *WISE* colors for objects classified as “long-period variables”





**Figure 15.** Symbols in the middle panel show the distribution of a subsample of 7123 variables (out of 7194) in the PLV catalog that are detected by the *WISE* survey and have *WISE* magnitudes  $W1 < 16.5$  and  $W2 < 15.5$  ( $5\sigma$  detection limits). Objects classified as “long-period variables” (defined as variables with periods longer than 50 days, and semi-regular variables) are shown as open circles (74 objects); the majority display infrared excesses compared with the colors of dust-free stars ( $W1 - W2 \sim 0$ ). Nine objects with light curves classified as “Other” and quasar-like infrared colors,  $W1 - W2 > 0.7$ , are shown as large triangles. The top panel shows a *WISE* color-color diagram for the subset of 99 objects that have  $W3 < 11.2$  (note that the majority of objects without significant infrared emission do not satisfy this condition; out of 74 long-period variables, 49 satisfy the  $W3$  brightness limit, as well as nine objects with quasar-like colors and 41 other objects). The bottom panel shows the period-color diagram for the 74 long-period variables. Examples of their light curves are shown in Figure 16. (A color version of this figure is available in the online journal.)

is consistent with the majority of them being genuine AGB stars (Tu & Wang 2013; 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 Section 3.6.1). The paucity of long-period variables with  $W1 > 13$  is a Galactic structure effect—at the high latitudes probed by the LINEAR sample (due to the requirement of overlap with the SDSS footprint) this magnitude cutoff corresponds to several tens of kpc and thus reaches many disk scale heights away from the plane (N. Hunt-Walker et al., in preparation).

The top panel in Figure 15 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  $\log(P)$  at a fixed color of about 0.2 dex is in good agreement with earlier work (e.g., see Whitelock et al. 2006 and

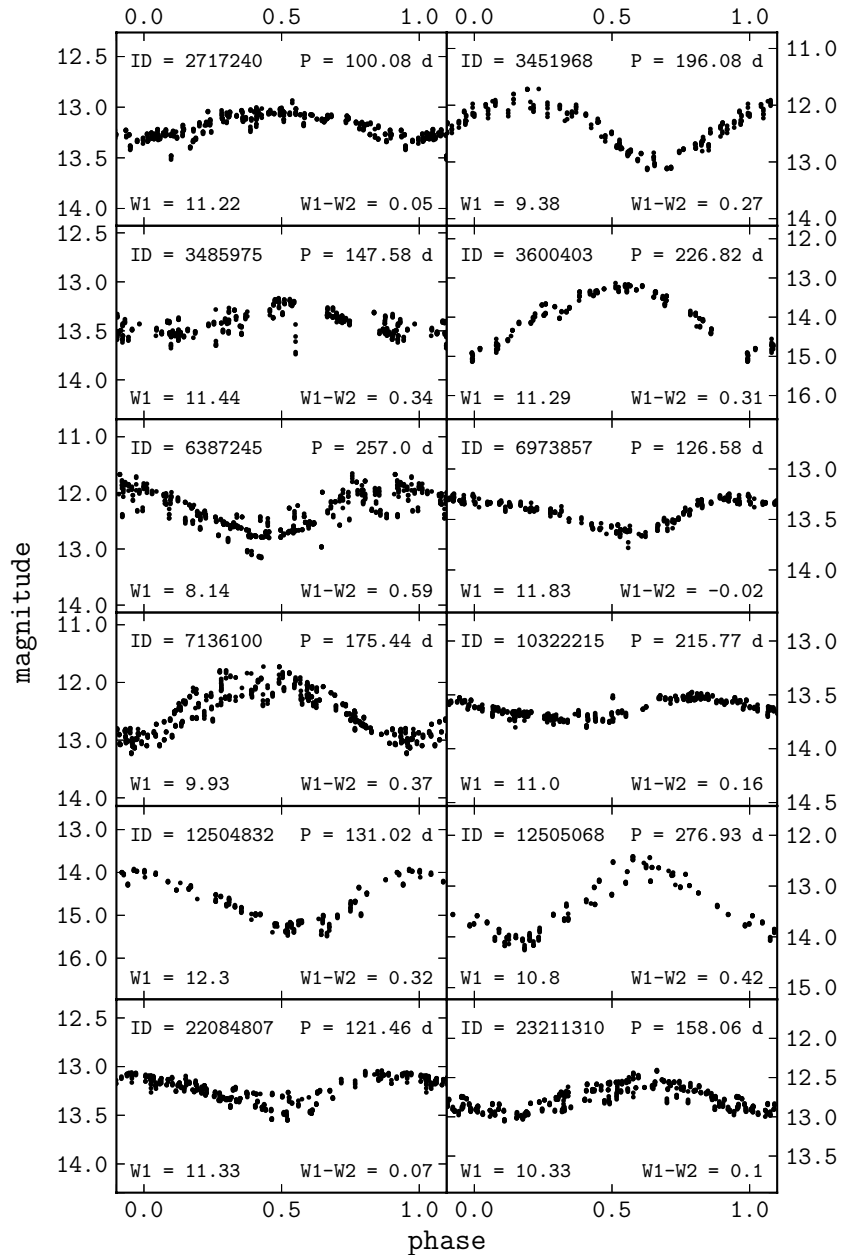
references therein). Examples of LINEAR light curves for long-period variables are shown in Figure 16. We note that the scatter in the phased light curves is much larger than the photometric errors and reflects the fact that light curves for these stars are not exactly reproducible between different cycles.

There are nine objects with light curves classified as “Other” that show infrared colors consistent with quasars ( $W1 - W2 > 0.7$ ; see, e.g., Yan et al. 2013). In addition, there are 14 objects with  $W2 - W3 > 2.0$ , implying strong infrared excesses that are likely inconsistent with AGB stars, but also with blue  $W1 - W2 < 0.5$  colors inconsistent with quasars (R. Nikutta et al., in preparation). A few but not all of these results could be chance positional coincidences with background quasars that would mostly affect the  $W3$  and  $W4$  measurements (based on a quasar surface density of several hundred per square degree and a matching radius of 3 arcsec).

### 3.6. Noteworthy Objects

There are six interesting sources that deserve direct mention by name. There is one case of a likely Type Ia supernova (LINEAR ID = 7682813; see the bottom-left panel in Figure 17) that increased in brightness by 0.8 mag over about 10 days and then gradually returned to its 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. For the standard cosmology, the implied absolute magnitude at maximum light is  $M = -19.4$ , which is consistent with a supernova classification. The absolute magnitude of its blue host galaxy is  $M = -18.6$ , in agreement with expectations. The object with LINEAR ID = 17655724 (see the bottom-right panel in Figure 17) steadily increased in brightness by 0.5 mag over about 5 yr. If this trend continues, in 400 yr it would outshine the Sun; nevertheless, this fact 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). In addition, the Catalina Sky Survey (Drake et al. 2009) data demonstrate that the brightness increase is slowing down.

Given its light curve that shows large variations (i.e., a decrease in brightness of  $\sim 1$  mag over  $\sim 200$  days; see the top left panel in Figure 17) and its *WISE* colors, the object with LINEAR ID = 2752114 is a good candidate for an R Coronae Borealis star, a supergiant carbon-rich star with episodic mass loss (Tisserand 2012; Tisserand et al. 2013). On the other hand, an object with a similar light curve and *WISE* colors, LINEAR ID = 3766947, is a confirmed BL Lac object at a redshift of 0.1325. The object LINEAR ID = 7455728 (see the top right panel in Figure 17) 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 star (6 arcsec away), 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,  $P \sim 604$  days, amplitude  $\sim 0.4$  mag; see the bottom-right panel in Figure 18). A periodogram of this object shows a strong peak, however the shape of the light curve is not fully repeatable. 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



**Figure 16.** Examples of light curves for objects classified as long-period variables. Each panel lists LINEAR ID, best-fit period in days, and the *WISE* W1 magnitude and W1 – W2 color (see Figure 15). The scatter in the light curves is much larger than the photometric errors and reflects the fact that the light curves for these stars are not exactly reproducible between different cycles.

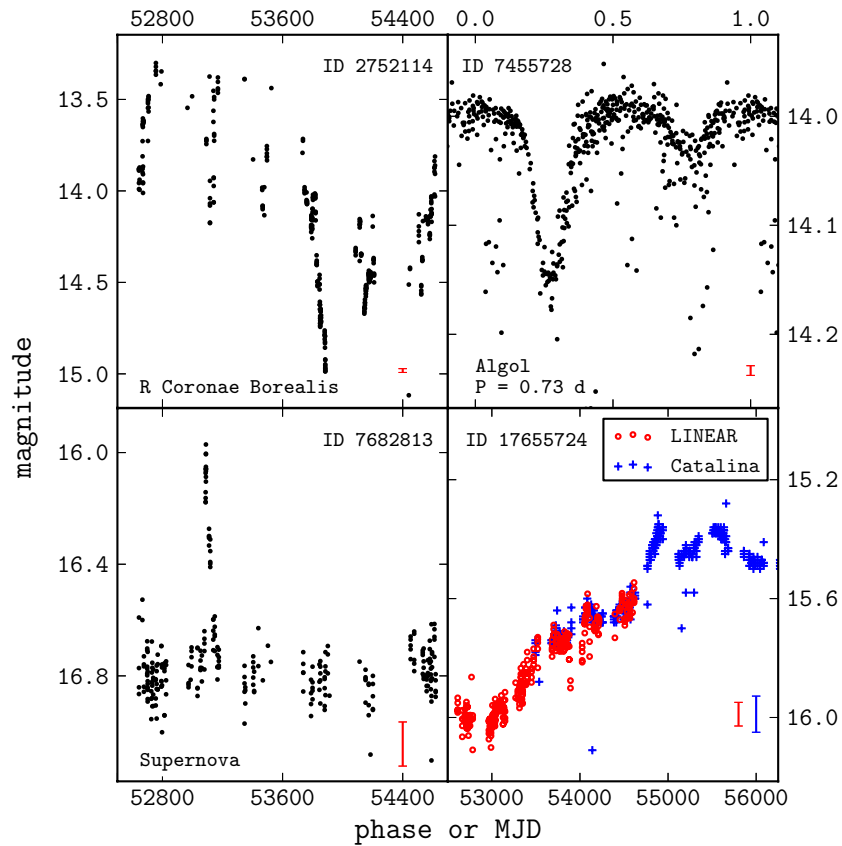
sample of  $\sim 9000$  quasars from the SDSS Stripe 82 region with spectroscopic confirmation and SDSS light curves. MacLeod et al. (2010) declared all the objects to be unconvincing cases of periodicity because their best-fit periods were roughly the same as the span of the observations—that is, only 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 follow-up study.

### 3.6.1. Optically Resolved, Periodically Variable Objects

Among the 7194 objects listed in the PLV catalog, 18 are optically resolved (sufficiently large difference between the point spread function (PSF) and model magnitudes) in the SDSS imaging data and an additional 116 objects have unreliable size measurements. Their SDSS image stamps are shown in

Figure 19. As is evident, eight objects are clearly galaxies and their variability may be at least to some extent due to photometric measurement difficulties when using the LINEAR images. Nevertheless, three objects (LINEAR IDs = 7682813, 8440571, and 9183803) show spectroscopic evidence for active galactic nucleus (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 10 objects that do not appear to be well-resolved galaxies are shown in Figure 18. The object in the middle-right panel (LINEAR ID = 22993473; the fourth object in the third row in Figure 19) 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 PSFs (including the known RR Lyrae star V368 Her, shown in the top-left panel); such gradients can be a sign of a binary nature or possibly fast



**Figure 17.** LINEAR light curves for four objects with unusual light curves (top left: an R Coronae Borealis candidate; top right: an Algol-like variable; bottom left: a supernova candidate; bottom right: a quasar with steady brightness increase; for more details see Section 3.6). Each panel lists the object’s LINEAR ID and its visual light-curve classification from the PLV catalog. The vertical error bars show typical photometric errors for each light curve. The top right panel shows a phased light curve. The bottom right panel also shows the Catalina Sky Survey data (small crosses).

(A color version of this figure is available in the online journal.)

changes in the PSF 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 18 have already been discussed: the carbon-rich AGB star CW Leo and a quasar with a nearly periodic light curve. For the latter, we have added data from the Catalina Sky Survey; in the overlap with the LINEAR data, the two light curves are consistent. These additional data provide further support for the quasi-periodic light variations displayed by this quasar.

#### 4. CLASSIFICATION BASED ON MACHINE LEARNING ALGORITHMS

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 this behavior can enable robust and efficient automated classification of objects into various classes of the variable population. We consider two classification methods based on machine learning algorithms.

First, we analyze the performance of an unsupervised classification algorithm that attempts to recognize existing variability classes in the PLV catalog using only their clustering in the multi-dimensional attribute space, but not the results of the visual light-curve classification. The motivation here is that these clusters correspond to different physical classes of objects (different types of variable stars) and an automated method might

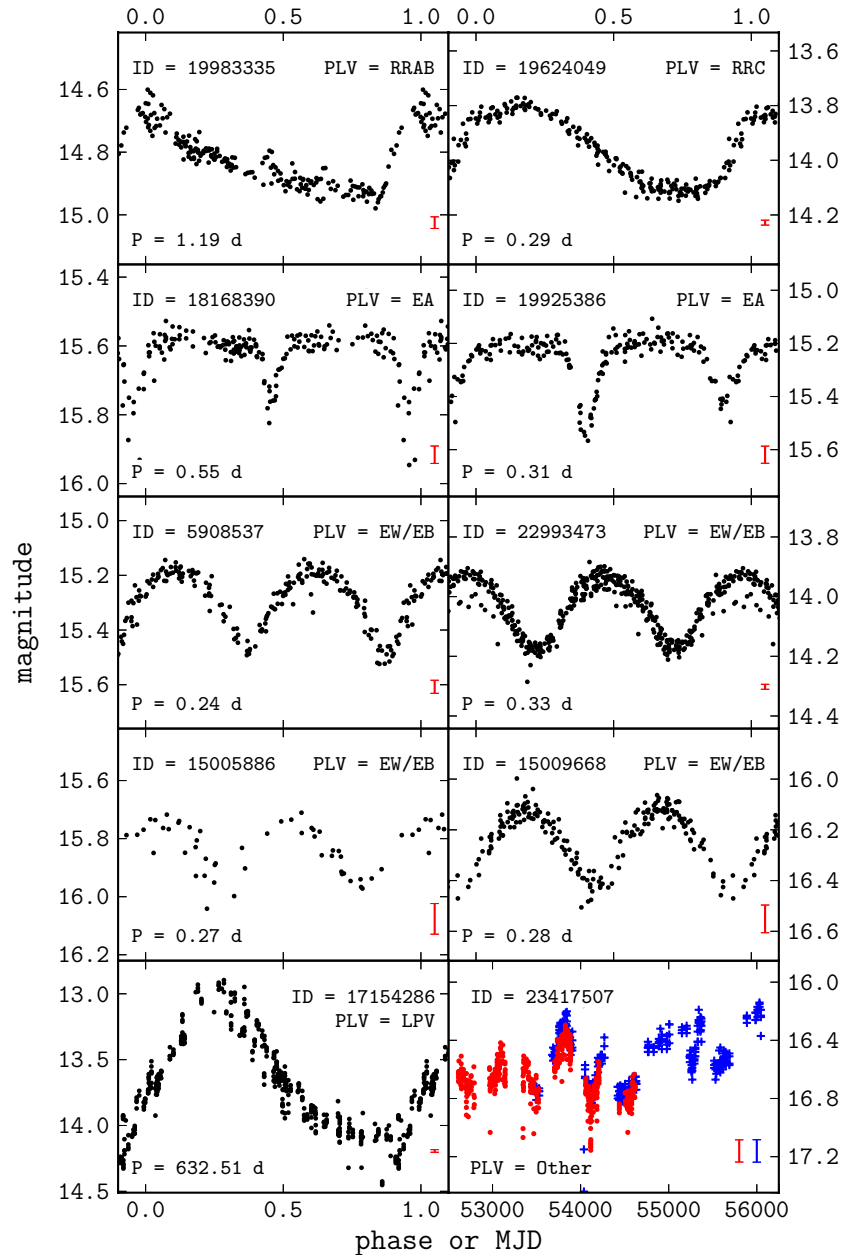
pick additional clusters. We also perform the so-called “supervised classification,” where a training sample is used to define selection boundaries. The main goal is to quantify whether visual classification could be improved, or perhaps entirely bypassed.

In order to avoid the impact of objects with unreliable measurements, the starting sample of 7194 variables is cleaned of sources with unreliable periods, bad SDSS photometry, and sources without 2MASS detections. We consider only the five most populous classes (ab type and c type RR Lyrae, EA and EW/EB eclipsing binaries, and SX Phe/ $\delta$  Sct candidates). The resulting cleaned sample of 6146 variables is publicly available from the same site as the main catalog.<sup>16</sup>

##### 4.1. Unsupervised Classification Based on a Gaussian Mixture Model

The strong clustering of objects in the multi-dimensional attribute space, visually classified in six different types using their light curves, 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 (GMM) 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

<sup>16</sup> Available from [http://www.astro.washington.edu/users/ivezic/r\\_datadepot.html](http://www.astro.washington.edu/users/ivezic/r_datadepot.html)



**Figure 18.** 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 the object’s 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 (circles; red in the online version) and CSDR2 (crosses; blue in the online version) data and confirms its quasi-periodic behavior (note that its full light curve, and not its phased light curve, is shown in this panel).

(A color version of this figure is available in the online journal.)

a light curve, are also possible (e.g., Debosscher et al. 2007; Richards et al. 2011, and references therein).

The 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 are assigned. 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<sup>17</sup> (VanderPlas et al. 2012)

data mining and machine learning tools implemented in *python*. Figures 20 and 21 show the GMM results for two cases.

The top panel in Figure 20 shows a 12-component GMM using only the two most discriminative data attributes, the  $g - i$  color and  $\log(P)$ . The number of components is determined automatically using the Bayesian Information Criterion (see *astroML* documentation for details). Out of the 12 clusters, 6 are very compact, while the rest seem to describe the background. Three clusters correspond to ab and c type RR Lyrae stars. Interestingly, the former are separated into two clusters. The reason is that the  $g - i$  color is a single-epoch color from SDSS that corresponds to a random phase. Since ab type RR Lyrae stars spend more time close to minimum than maximum

<sup>17</sup> See <http://www.astroML.org>





**Figure 19.** Default SDSS *gri* composite images of 18 resolved objects that display visually confirmed variability in the 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 10 objects are shown in Figure 18. The extremely red source (the second panel in the bottom row) is the brightest carbon-rich AGB star, CW Leo (IRC + 10216).

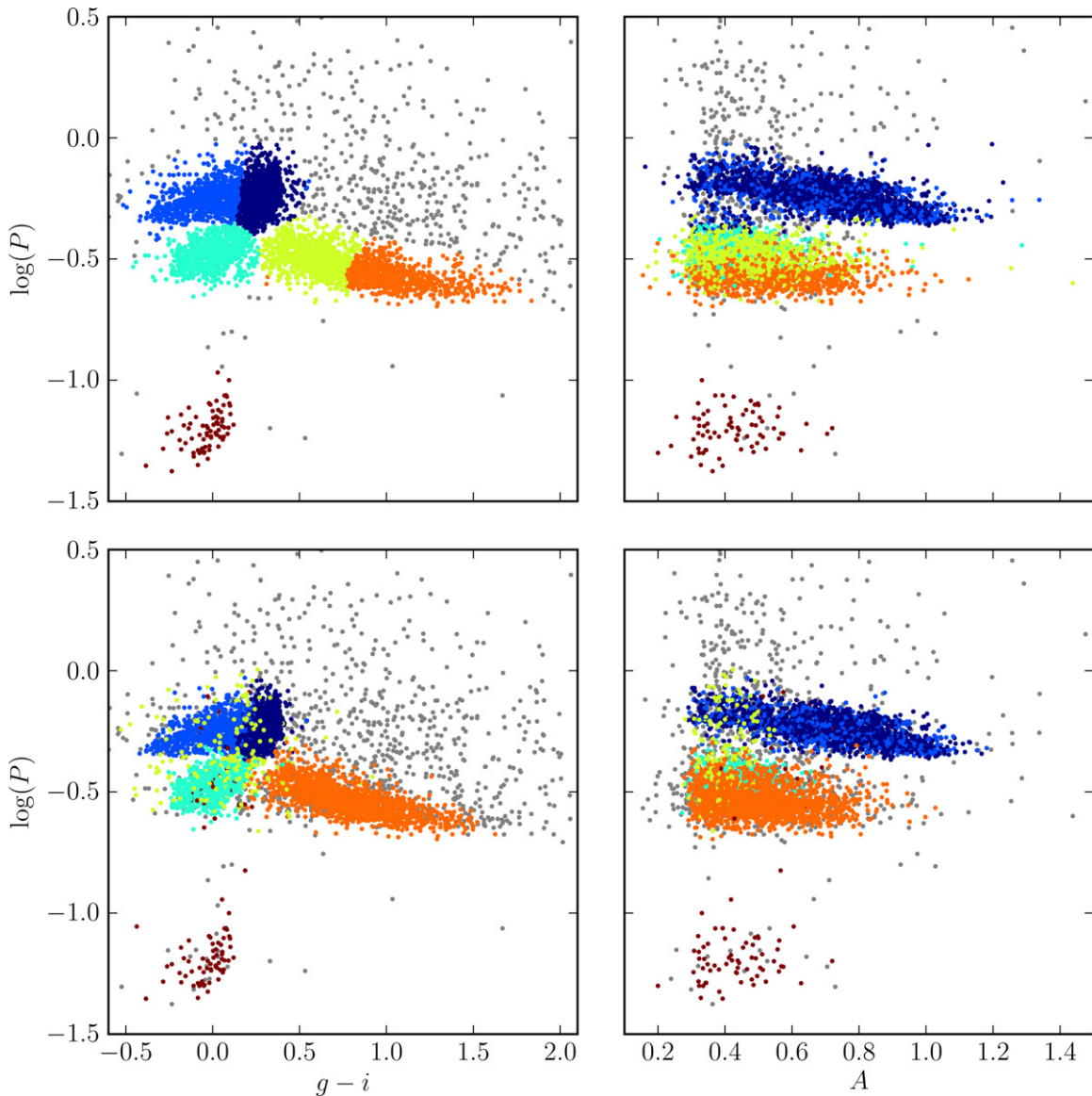
(A color version of this figure is available in the online journal.)

light, when their colors are red compared with their colors at maximum light, their color distribution deviates strongly from a Gaussian. The elongated sequence populated by various types of eclipsing binary stars is also split into two clusters because its shape cannot be described by a single Gaussian either. The upper-right panel shows the clusters in a different projection,  $\log(P)$  versus light-curve amplitude. The top four clusters are still fairly well localized in this projection due to  $\log(P)$  having significant discriminative power.

In another instance of GMM analysis, the clustering attributes included four photometric colors based on SDSS and 2MASS measurements ( $u - g$ ,  $g - i$ ,  $i - K$ , and  $J - K$ ) and three parameters determined from the LINEAR light-curve data ( $\log(P)$ , amplitude, and light-curve skewness). A

15-component GMM to this seven-dimensional dataset yields the clusters shown in the bottom panels of Figure 20. The clusters derived from all seven features are remarkably similar to the clusters derived from just two features; this result shows that the additional data add very little new information (equivalently, this result shows that the seven attributes are strongly correlated). The main difference compared with the two-attribute case is that the EB/EW sequence is now described by a single component. Figure 21 shows the locations of the six most compact clusters in the space of the other attributes.

As is evident from a visual inspection of Figures 20 and 21, the most discriminative attribute is the period. A few clusters that have very similar period distributions are separated in  $g - i$  and  $i - K$  colors, which are a measure of the star's effective



**Figure 20.** Unsupervised clustering analysis of periodic variable stars from the LINEAR dataset using the *astroML* code for the GMM algorithm. The top row shows clusters derived using two attributes ( $g - i$  and  $\log(P)$ ) and a mixture of 12 Gaussians. The colored symbols mark the six most compact 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 15 Gaussians. See Figure 21 for data projections in the space of other attributes for the latter case. This figure is adapted from Ivezić et al. (2013) and can be reproduced using code available at <http://www.astroML.org> (VanderPlas et al. 2012).

(A color version of this figure is available in the online journal.)

temperature; see Covey et al. (2007). In summary, although there are many Gaussian components in the chosen mixture models, no new compact classes were revealed by this automated analysis.

#### 4.2. Supervised Classification with Support Vector Machine

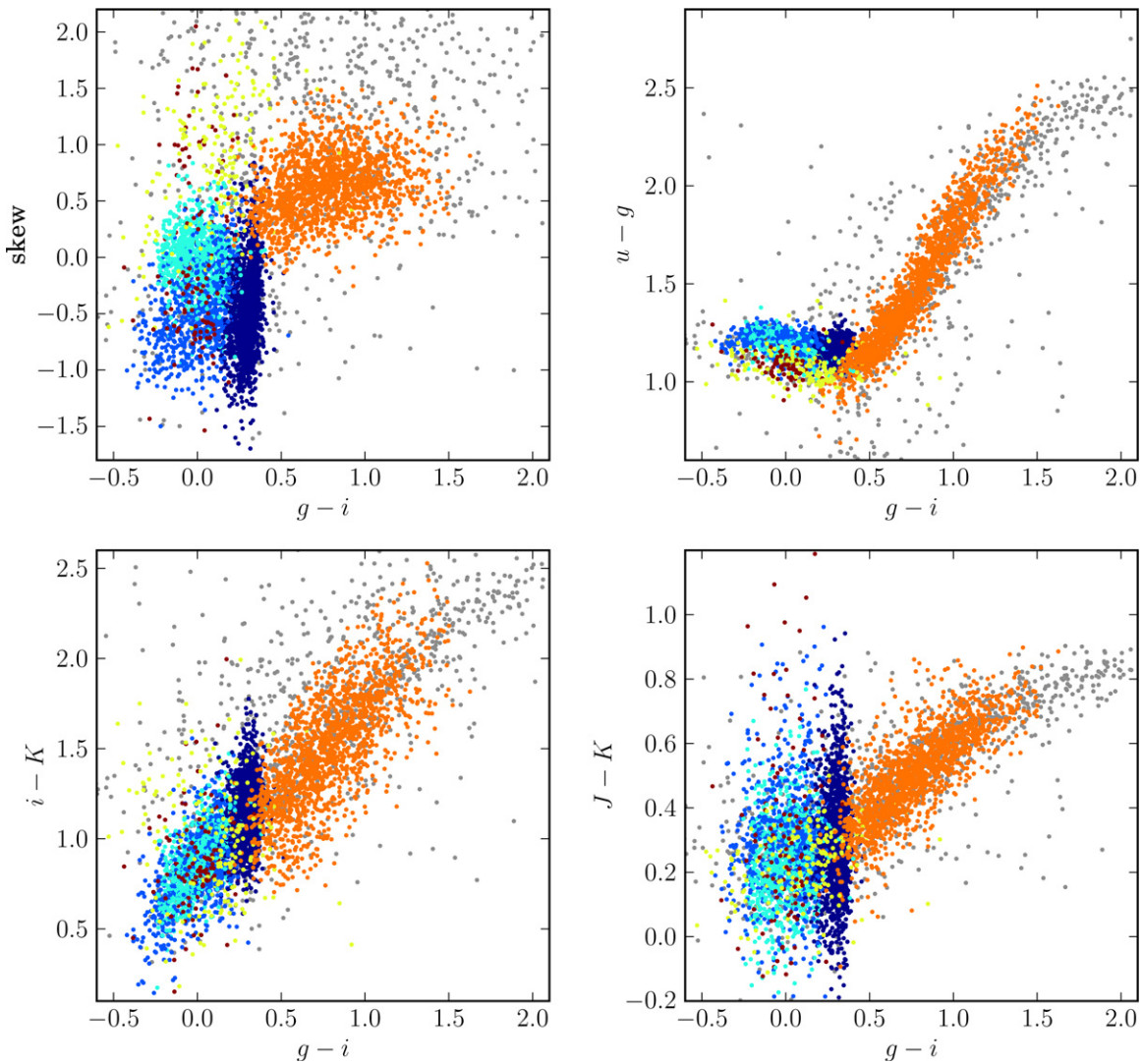
Given the results of visual classification, we attempt to reproduce it in an automated fashion using supervised classification and a machine learning method called support vector machine (SVM; Cortes & Vapnik 1995). SVM uses linear classification boundaries, but unlike our simple method described in Section 2.3.6, it does not need to be aligned with the coordinate axes. The optimal classification boundaries are those that maximize the class separation or margin (the training points that are found on the margin are called support vectors).

We used a multi-label SVM from the *scikit-learn* package (Pedregosa et al. 2011), via *astroML*. A randomly selected third

of the sample is used for training SVM and the remaining two thirds of the sample is used for measuring the classification performance. Figures 22 and 23 illustrate the SVM results for two cases,<sup>18</sup> and Table 4 provides a quantitative summary.

As with unsupervised GMM clustering, both the two-attribute and seven-attribute cases are considered. SVM assigns a large fraction of the EA class (Algol-type eclipsing binaries) to the EB/EW class (contact binaries). This result is not necessarily a problem with the SVM method because these two classes are hard to distinguish given LINEAR light curves. Compared with the simple method discussed in Section 2.3.6, the precision of the SVM classification relative to visual classification is a bit better (especially for c type RR Lyrae stars). Furthermore, the SVM code from *astroML* was much

<sup>18</sup> This part of the analysis can be easily reproduced using the public and open-sourced *astroML* code and the datasets available at <http://www.astroML.org>



**Figure 21.** Unsupervised clustering of periodic variable stars from the LINEAR dataset using the GMM algorithm. 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 15 Gaussians. The colored symbols mark the six most compact clusters. 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 20. This figure is adapted from Ivezić et al. (2013) and can be reproduced using code available at <http://www.astroML.org> (VanderPlas et al. 2012).

(A color version of this figure is available in the online journal.)

**Table 4**

The Performance of Supervised Classification Using the Support Vector Machines Method in the Seven-Attribute Case

Class	$N$	RRAB	RRC	EA	EB/EW	SX Phe
RRAB	1772	95.9	0.3	1.4	2.4	0.0
RRC	583	1.5	91.3	0.2	7.0	0.0
EA	228	5.3	1.3	67.5	25.9	0.0
EB/EW	1507	2.1	4.0	3.1	90.7	0.1
SX Phe	56	0.0	1.8	0.0	1.8	96.4
Purity	...	97	88.4	68.4	90.5	98.2

**Notes.** Each row corresponds to an input class listed in the first column (RRAB: ab type RR Lyrae; RRC: c type RR Lyrae; EA: Algol-type eclipsing binaries; EB/EW: contact eclipsing binaries; SX Phe: SX Phe and  $\delta$  Sct candidates). The second column lists the number of objects in each input class and the remaining columns list the percentage of sources classified into the classes listed in the top row. The bottom row lists the classification contamination in percent for each class listed in the top row.

easier to deploy than developing the manual method from Section 2.3.6.

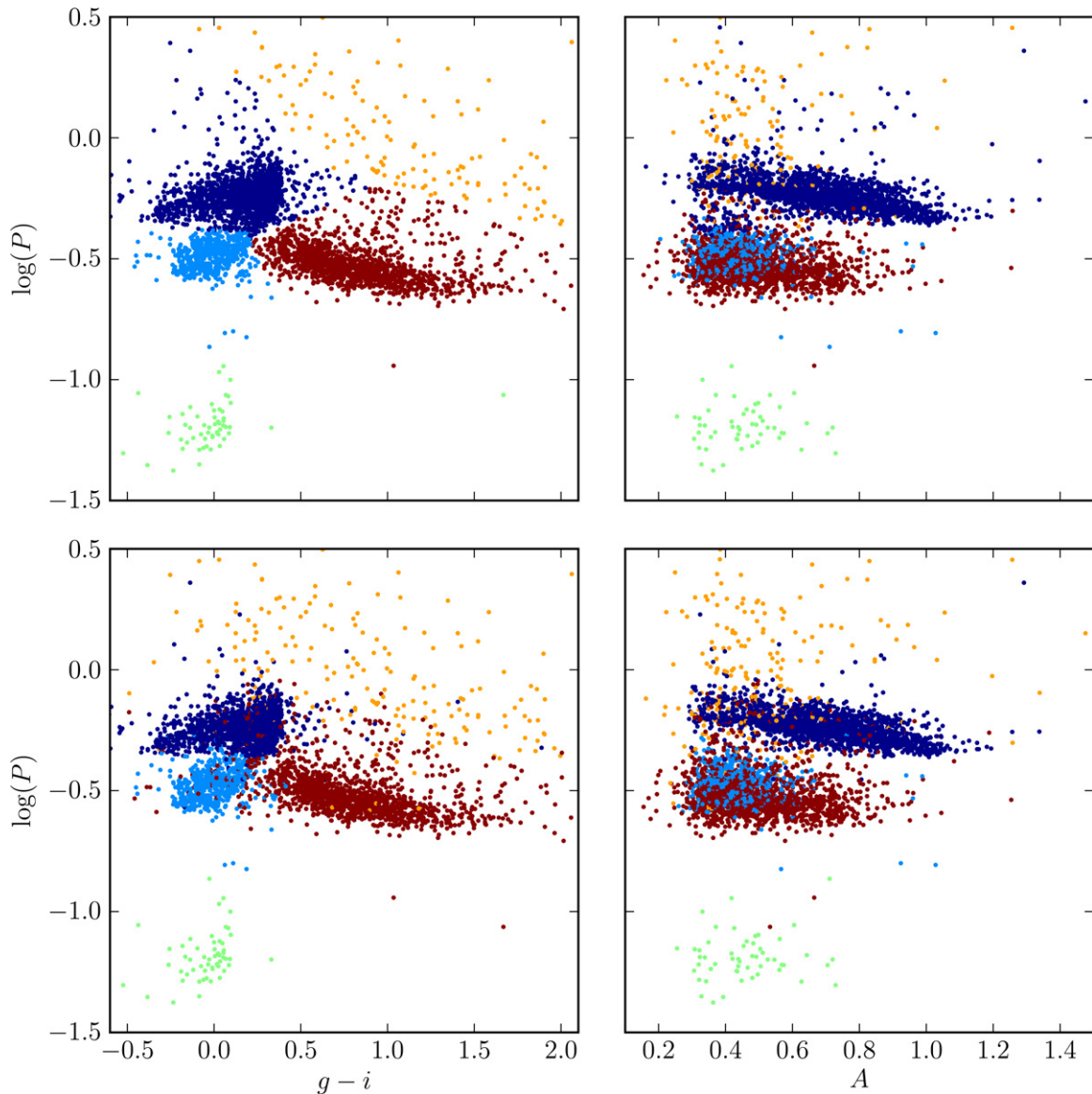
## 5. DISCUSSION AND CONCLUSIONS

We described the creation of a catalog of visually confirmed periodic variable stars selected from data acquired by the LINEAR asteroid survey, the “PLV” catalog. This publicly available catalog consists of 7194 variable objects, where over 96% are likely periodic variable stars.<sup>19</sup> Combined with a large sky coverage ( $\approx 10,000 \text{ deg}^2$ ) and a flux limit several magnitudes fainter than most other wide-angle surveys ( $14 < r < 17$ ), this catalog is useful for a wide variety of research topics such as studies of Galactic halo structure and the physics of pulsating stars and eclipsing binaries.

The completeness of the PLV catalog, relative to the initial sample of 200,000 candidate variables, is very high ( $>98\%$ ); nevertheless, it is subject to the selection criteria listed in Section 2.1 that were used to select the initial sample based

<sup>19</sup> We provide examples of the online data in Tables 5–7.





**Figure 22.** Supervised classification analysis of periodic variable stars from the LINEAR dataset using the *astroML* code for the SVM algorithm. The top row shows classes derived using visual classification results for five classes and two attributes ( $g - i$  and  $\log(P)$ ). One third of the sample was used as a training sample. The colored symbols mark objects from the five classes adopted by SVM. The bottom row shows the analogous diagrams for classification based on seven attributes (colors  $u - g$ ,  $g - i$ ,  $i - K$ , and  $J - K$ ,  $\log(P)$ , light-curve amplitude, and light-curve skewness). See Figure 23 for data projections in the space of other attributes for the latter case. This figure can be reproduced using code available at <http://www.astroML.org> (VanderPlas et al. 2012).

(A color version of this figure is available in the online journal.)

**Table 5**  
PLV Catalog: Light Curve Data

ID	LCtype	$P$	$A$	mmed	stdev	rms	$L\chi^2_{\text{pdf}}$	nObs	Skew	Kurt	$LR\chi^2$	CUF	t2	t3
2522	5	0.238812	0.68	17.00	0.22	0.25	0.543	225	0.75	0.11	0.317	1	0	0

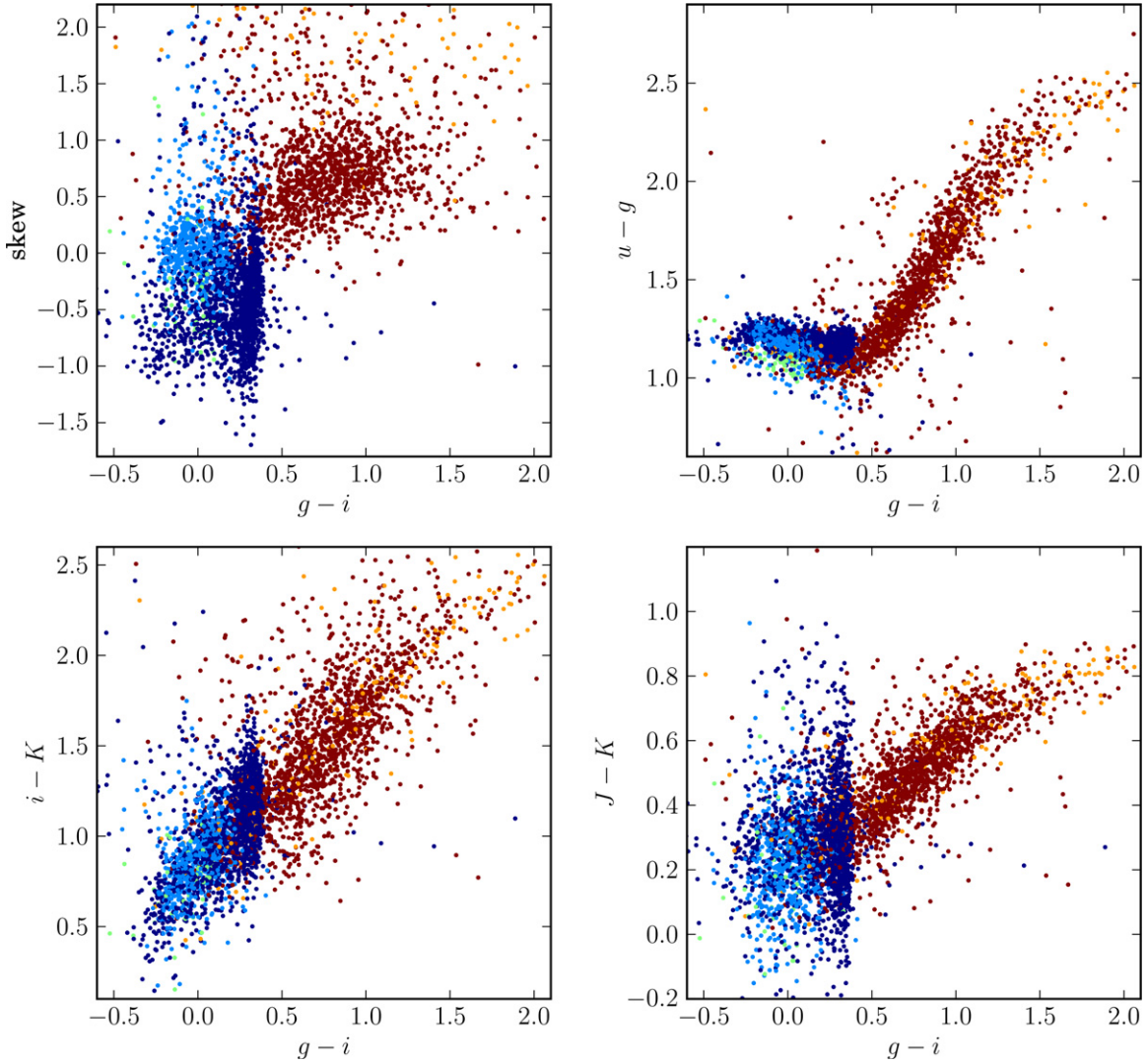
**Notes.** A detailed description of the entries is provided in the table header. Only the first entry is shown for illustration.

(This table is available in its entirety in machine-readable and Virtual Observatory (VO) forms in the online journal. A portion is shown here for guidance regarding its form and content.)

on visual classification. Based on a comparison with the SDSS Stripe 82 variable stars, we estimated that the completeness of the PLV catalog is 55%–70%; most of the LINEAR incompleteness is due to the larger adopted minimum rms variability, 0.1 mag versus 0.05 mag for the SDSS catalog.

The purity of the PLV catalog and the classification precision are both high (>96% of entries have an assigned light-curve type). The folded light curves of all the objects in the catalog were visually inspected several times. Additional attributes (SDSS, 2MASS, and *WISE* colors) were used to better characterize each of the objects and thus improve the classification





**Figure 23.** Supervised classification analysis of periodic variable stars from the LINEAR dataset using the SVM algorithm. Classes are derived using seven attributes (colors  $u - g$ ,  $g - i$ ,  $i - K$ , and  $J - K$ ,  $\log(P)$ , light-curve amplitude, and light-curve skewness). The colored symbols mark objects from the five classes adopted by SVM. The  $\log(P)$  vs.  $g - i$  diagram and  $\log(P)$  vs. light-curve amplitude diagram for the same classes are shown in the lower panels of Figure 22. This figure can be reproduced using code available at <http://www.astroML.org> (VanderPlas et al. 2012).

(A color version of this figure is available in the online journal.)

**Table 6**  
PLV Catalog: SDSS Data

ID	R.A.	Decl.	oType	nS	rExt	$u$	$g$	$r$	$i$	$z$	uErr	gErr	rErr	iErr	zErr
2522	117.99	48.67	6	1	0.139	20.24	18.24	17.28	16.89	16.67	0.06	0.01	0.01	0.01	0.01

**Note.** A detailed description of the entries is provided in the table header.

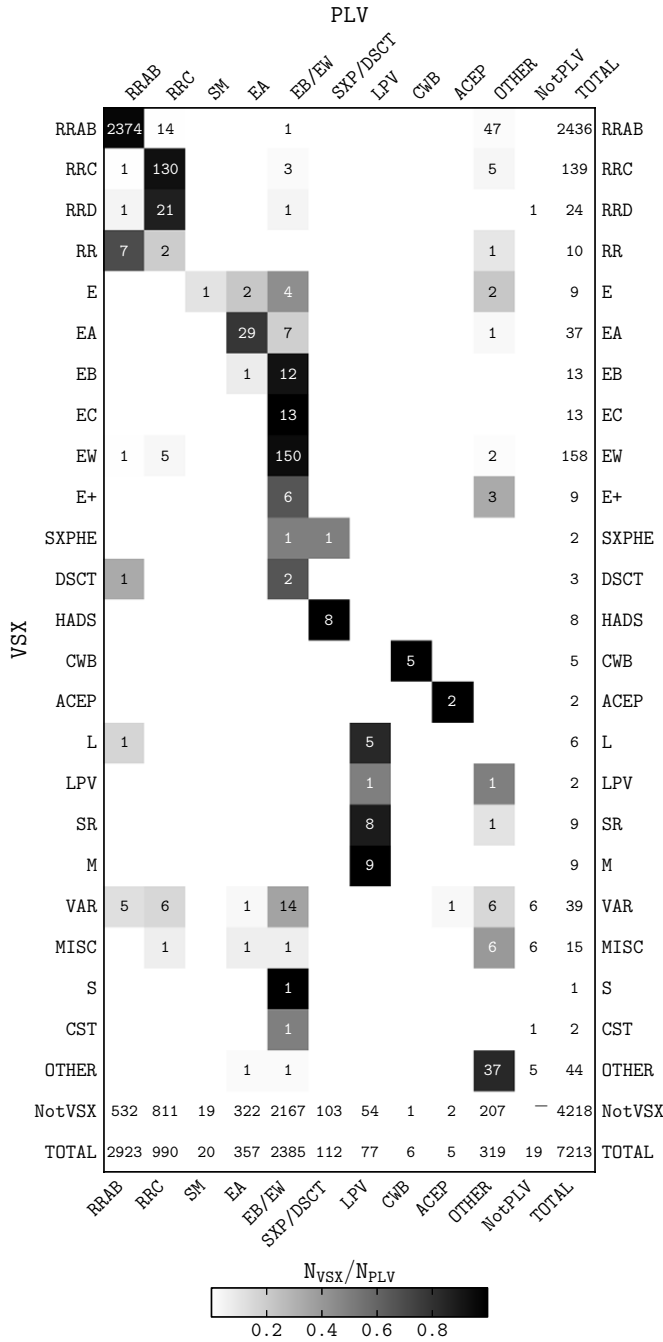
(This table is available in its entirety in machine-readable and Virtual Observatory (VO) forms in the online journal. A portion is shown here for guidance regarding its form and content.)

**Table 7**  
PLV Catalog: 2MASS and WISE Data

ID	$J$	$H$	$K$	Jerr	Herr	Kerr	$W_1$	$W_2$	$W_3$	$W_4$	W1err	W2err	W3err	W4err
2522	15.45	14.94	14.76	0.06	0.09	0.10	14.52	14.63	12.60	8.84	0.03	0.07	−9.90	−9.90

**Note.** A detailed description of the entries is provided in the table header.

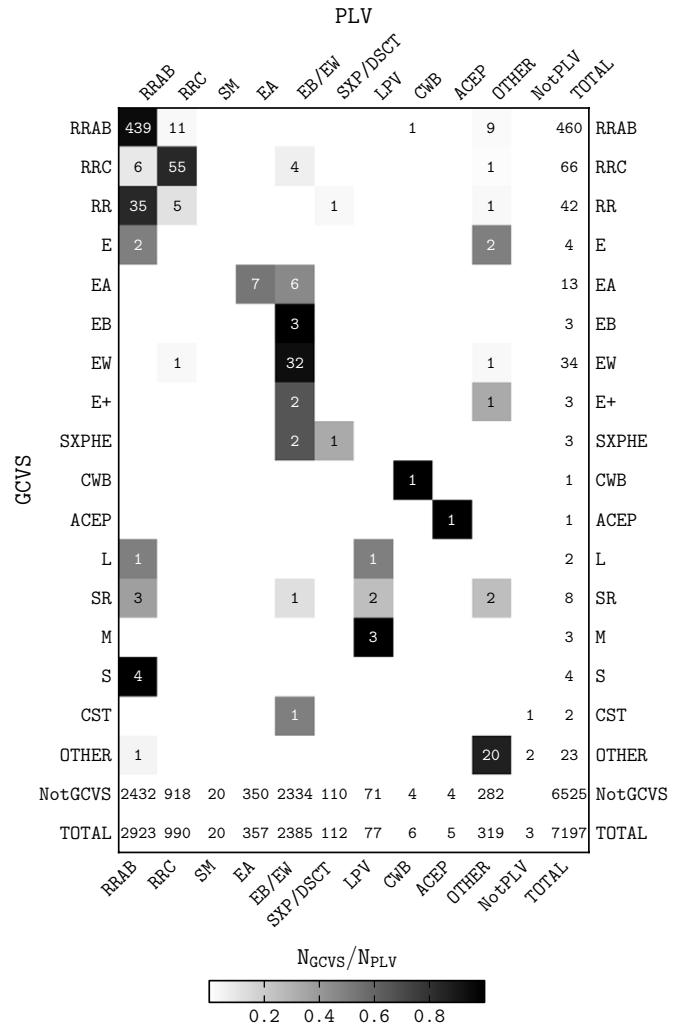
(This table is available in its entirety in machine-readable and Virtual Observatory (VO) forms in the online journal. A portion is shown here for guidance regarding its form and content.)



**Figure 24.** VSX vs. PLV confusion matrix. The column labeled “Other” corresponds to variable PLV objects that do not have a reliable variability type. The “NotPLV” column corresponds to VSX objects that are not included in the PLV catalog and the row “NotVSX” corresponds to PLV objects not listed in the VSX catalog. The row “Other” corresponds to VSX variables with classes others than those 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 (2006).

purity. Furthermore, we compared our results with the General Catalog of Variable Stars (GCVS), the Variable Star Index (VSX) catalog, and the RR Lyrae catalogs from the Catalina and Mount Lemmon Surveys (see the Appendix for details) in order to ascertain the effectiveness of our method. This analysis provides further support for our claim of a low contamination level by non-variable objects in the PLV catalog.

Our analysis was focused on the periodic variables, therefore many irregular and quasi-periodic variables did not make it into

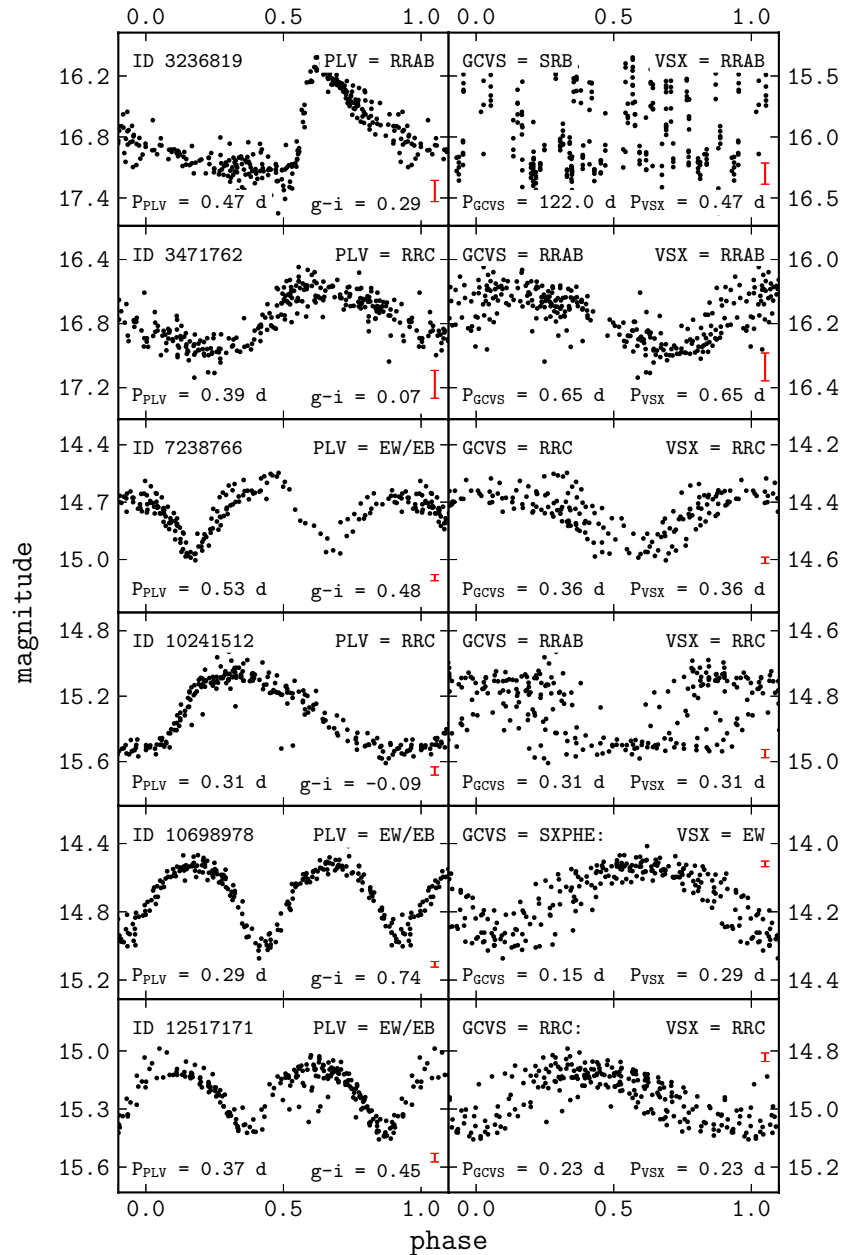


**Figure 25.** GCVS vs. PLV confusion matrix. The column labeled “Other” corresponds to variable PLV objects that do not have a reliable variability type. The “NotPLV” column corresponds to GCVS objects that are not included in the PLV catalog and the row “NotGCVS” corresponds to PLV objects not listed in the GCVS catalog. The row “Other” corresponds to GCVS variables with classes others than those 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).

the visual inspection stage. In the case that these objects passed the initial low level statistical cuts, they were most often ignored during the visual classification process. We did, however, stumble upon some of these non-periodic objects while examining the light curves. Some of those variables and transients (e.g., AGNs, AM Herculis, BL Lacertae, BY Draconis, cataclysmic variables, RS Canum Venaticorum) are grouped in the “Other” PLV class.

This result suggests that many other interesting object types could be extracted from the PLV catalog. Many of these objects are not periodic and therefore we made no true attempt to classify them.

The PLV catalog is dominated by RR Lyrae stars (3913 or 54%) and eclipsing binaries (2762 or 38%). We also found 112 (1%) candidate SX Phe/ $\delta$  Sct variables and 77 (1%) red variables with long regular or semi-regular (SR) periods (Mirae, long-period variable, and SR). As suspected in the Introduction, we confirm that variable sources fainter than  $V = 14$  consist of quite a different population mix than the brighter and better studied sources. Table 3 describes in detail the content of the PLV catalog.



**Figure 26.** 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 errors for the 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, and V0593 Vir.

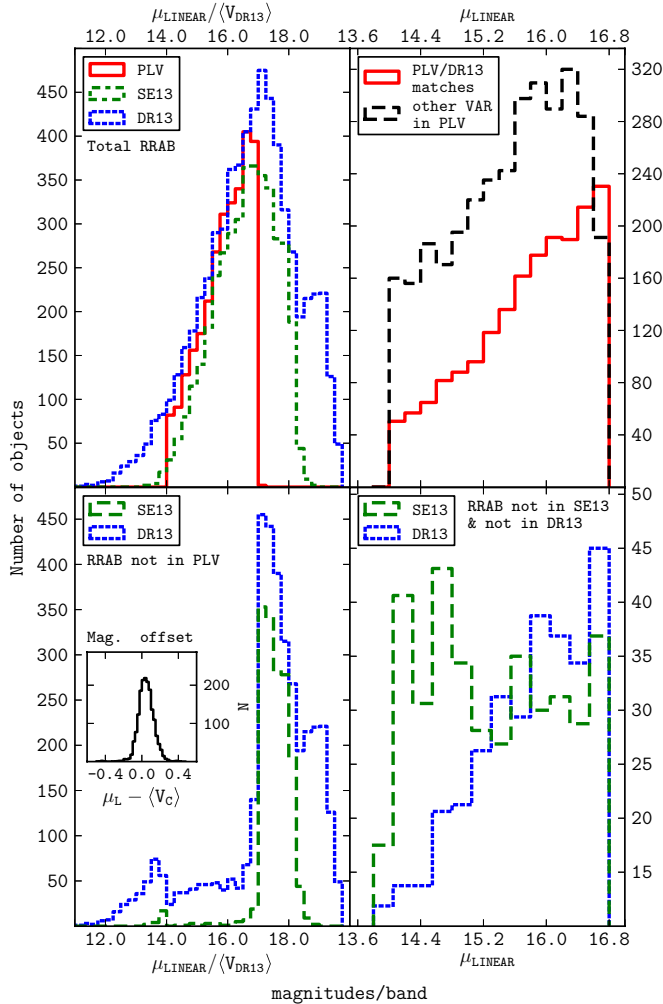
(A color version of this figure is available in the online journal.)

An exciting result of our effort is the discovery of 112 SX Phe/ $\delta$  Sct candidates. It is not possible to differentiate these two classes of objects on the basis of light-curve attributes and color. However, our preliminary analysis based on SDSS spectra and radial velocities (see Section 3.4 and Figure 14) shows that these candidates are consistent with Population II objects and therefore we assume that the sample is dominated by SX Phe stars. Until now, these stars have been found mostly in Galactic globular clusters ( $\approx 250$  objects in total) and only 17 field SX Phe stars are currently known. Therefore, if our assumption is correct, the PLV SX Phe sample would increase the number of currently known such stars by 30% and the number of known field SX Phe stars by as much as a factor of six. This increase in the sample size could play an important role in characterizing

not only this type of variable but blue stragglers as well. We are currently undertaking a follow-up program using several modest-size photometric telescopes (1.2 m and 0.25 m).

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

Compared to, e.g., 10,000 eclipsing binaries in the Galactic bulge fields discovered by OGLE II and analyzed by Devor (2004) or to  $\sim 2000$  eclipsing binaries discovered in the Kepler survey data (Prša & Zwitter 2005), our sample of  $\sim 2700$  stars is in the same realm of sample size. Its comparative advantage



**Figure 27.** Comparison of RR Lyrae catalogs between the PLV catalog, Sesar et al. (2013), and DR13. The median LINEAR magnitude is designated as  $\mu_{\text{LINEAR}}$  and the mean DR13 V magnitude is designated as  $\langle V_{\text{DR13}} \rangle$ . All four plots are made for the area in which the PLV catalog and DR13 overlap (approximately  $125^\circ < \text{R.A.} < 268^\circ$  and  $-13^\circ < \text{decl.} < 69^\circ$ ). 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,  $\mu_L$ ) and DR13 (Johnson V band,  $\langle V_C \rangle$ ). The histograms in the top right panel show the total number of matched objects between the PLV catalog and DR13, as well as the total number of other variable stars identified in the PLV catalog. The histograms in the bottom right panel show ab type RR Lyrae found by the PLV catalog and not listed in DR13 (dotted) and Sesar et al. (2013, dashed).

(A color version of this figure is available in the online journal.)

is its 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 possibly other parameters). We note that the period distribution of eclipsing binaries in the PLV catalog is generally in agreement with previous studies, e.g., (Giuricin et al. 1983; Devor 2004; Prša & Zwitter 2005).

We demonstrated that the availability of SDSS, 2MASS, and *WISE* data can enable analysis that is not possible with single-band light curves alone. For example, we derived a precise quantitative description of an interesting correlation between the colors of EB/EW type contact binaries and their period (Section 3.3): as the spectral type (determined from the  $g - i$

SDSS color) of these binaries changes from approximately K4 to F5, their median period increases from 5.9 to 8.8 hr. Since no consensus about the origin of the short-period boundary for contact binaries has yet been reached, the improvement in observational constraints enabled by the LINEAR data will be valuable for future studies of stellar evolution. We also showed how *WISE* colors can be used to better identify several populations, including AGB stars, R Coronae Borealis stars, and quasars.

We emphasize that the preliminary work described in Section 3 is by no means a complete analysis of the PLV catalog. To point out but a single example, detailed analyses 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 insights into the physics of such stellar systems. In addition, this variable star sample will be valuable to compare with the *Gaia* results, for example, to search for period evolution (e.g., Davenport et al. 2013).

We conclude by pointing out that processing the volume of light-curve data provided by the LINEAR survey is still (barely) manageable by human resources. However, with upcoming large surveys such as *Gaia* and LSST, automated schemes will have to be employed to classify the expected vast volumes of data. Examples of such methods, based on machine learning algorithms, are discussed in Section 4. In addition to the requirement for ever fainter training samples, we point out the need for the efficient automated recognition of outliers, a problem that we left for future work with the PLV catalog.

The authors thank *Gaia* Coordination Unit 7 (based at ISDC, Department of Astronomy, University of Geneva, Switzerland) for the help and infrastructure used in the calculation of Lomb–Scargle and generalized Lomb–Scargle periods.

L.P. acknowledges support from 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 No. 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.

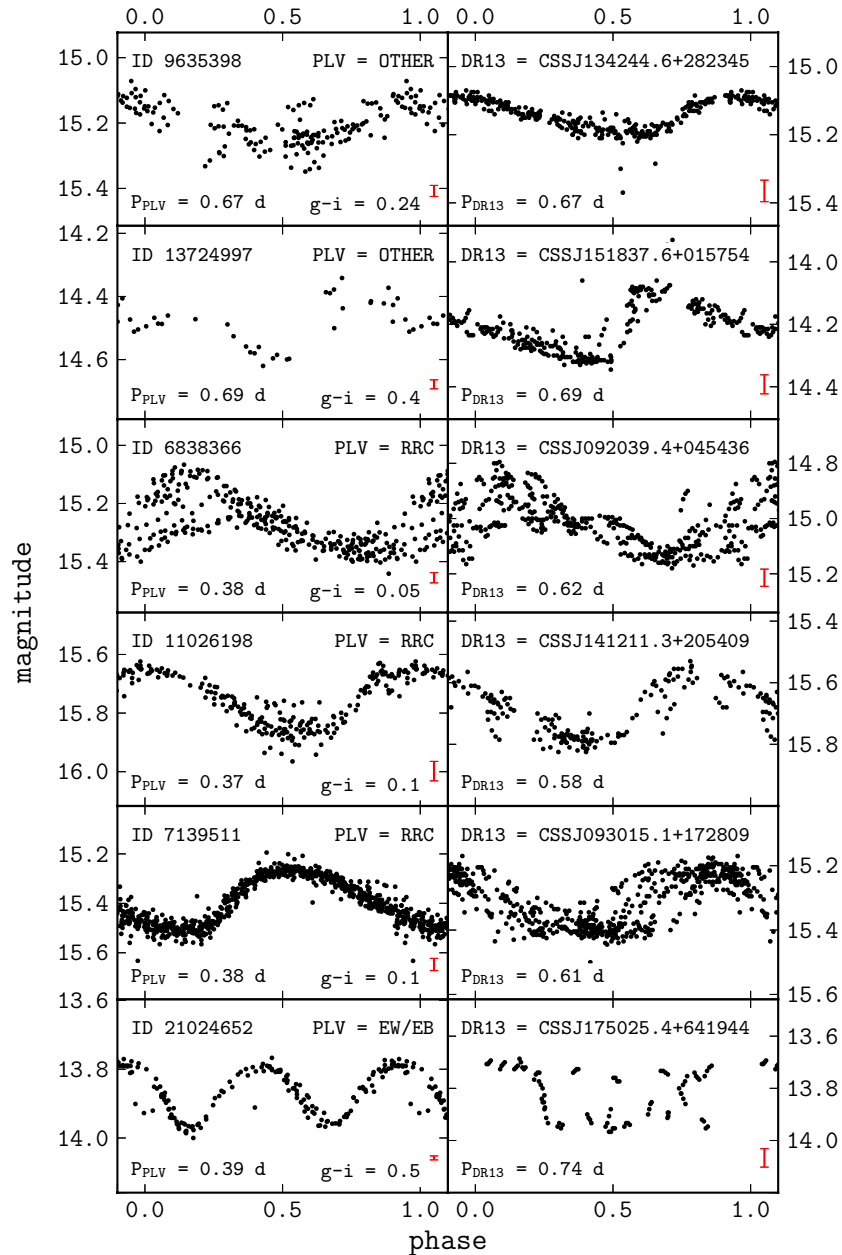
## APPENDIX

### COMPARISON WITH EXTANT CATALOGS OF VARIABLE STARS

#### A.1. Comparison with the General Catalog of Variable Stars and the AAVSO International Variable Star Index

In order to estimate the number of previously unknown variable stars in the PLV catalog, we compared this catalog with two online catalogs—the GCVS (Samus et al. 2009) and the American Association of Variable Star Observers International Variable Star Index (Watson 2006). The Topcat tool (Taylor 2005) was used to find positional matches within a 3 arcsec





**Figure 28.** Examples of light curves of objects for which the PLV catalog and DR13 classification do not agree. Light curves on the left show LINEAR data folded with the PLV periods and those on the right show DR13 data folded with the DR13 periods. The vertical error bars in each panel show the median errors (note that the CSDR2 errors are larger than the LINEAR errors). All of the objects were classified as ab type RR Lyrae by DR13.

(A color version of this figure is available in the online journal.)

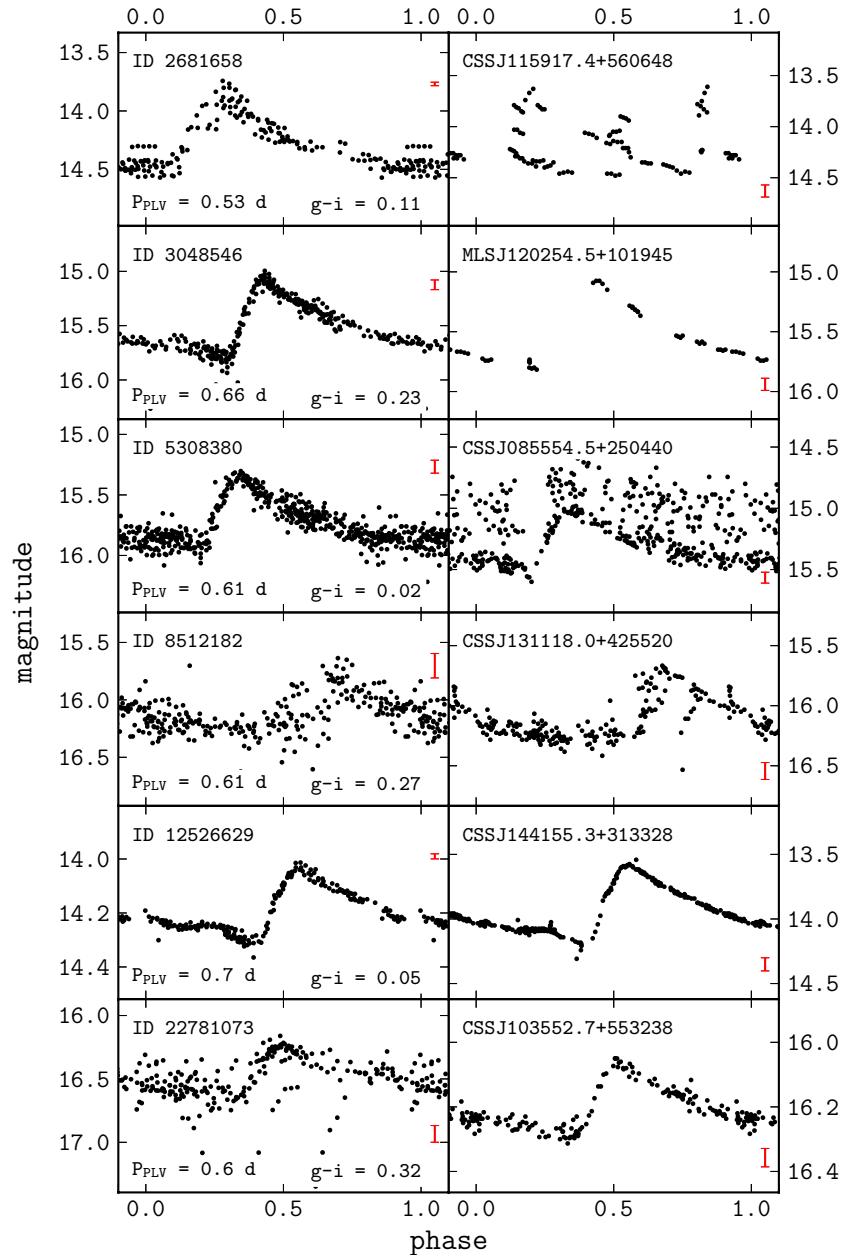
radius (in early 2013 February). Our results are summarized in Figures 24 and 25.

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. We note that the matching rate for the VSX catalog is higher than for matching to the SIMBAD database; only 1374 PLV entries, or 19%, have a SIMBAD object within 3 arcsec (with 41 different SIMBAD types, dominated by RR Lyrae stars and non-descriptive “Star” types, which account for  $\sim 70\%$  of matches). Therefore, the majority of PLV entries are previously uncataloged variable stars.

For both catalogs, the majority of unmatched objects are eclipsing binaries, followed by c type RR Lyrae, SX Phe/ $\delta$  Sct

candidates, and long period variables. Classification of the matched objects shows a good overall agreement between the 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 the VSX and the GCVS was performed and we stand by our classifications in all cases. In Figure 26, we show several examples where the classification from the GCVS and/or the VSX did not match the PLV classification.

This comparison with the VSX and the GCVS motivated us to introduce two more variable star classes: anomalous Cepheids and BL Herculis. Both can have light curves and colors that are very similar to those of ab type RR Lyrae. However, some



**Figure 29.** Examples of light curves of ab type RR Lyrae missing from DR13 but present in the PLV catalog (within the overlapping region). LINEAR data are shown in the left column and CSDR2 data are shown in the right column. The PLV periods have been used to fold the light curves. The vertical error bars in each panel show the median errors.

(A color version of this figure is available in the online journal.)

of them depart slightly from the locus populated by ab type RR Lyrae (in the color–period and other diagrams) and we have adopted the VSX and/or the GCVS classifications in these cases.

#### A.2. Comparison with the 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, 2013b). Their Catalina Surveys Data Release 2<sup>20</sup> catalog includes 15,000 ab type RR Lyrae selected from more than 200 million light curves obtained by the Catalina Schmidt Survey and the Mount Lemmon Survey over 20,000 deg<sup>2</sup> of sky, to a faint magnitude limit of  $V = 20$ . In the following text, we refer to this work as

DR13. Approximately 6460 DR13 objects are located inside an area covered by the PLV catalog (approximately  $125^\circ < \text{R.A.} < 268^\circ$  and  $-13^\circ < \text{decl.} < 65^\circ$ ). A cut in the magnitude range that corresponds to the brightness of objects potentially included in the PLV catalog ( $14 < V < 17$ ) selects approximately 3170 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 2612 objects (see Figure 27 for a statistical summary of the matched sources, which also includes a comparison with the deeper sample of RR Lyrae stars from Paper II). All but three matched sources are classified as variable and are included in the PLV catalog. Only 86 ( $\approx 3\%$ ) of the matched objects are not classified as ab type RR Lyrae in the PLV catalog. This agreement level is remarkable between two

<sup>20</sup> Available at <http://nessi.cacr.caltech.edu/DataRelease/>

catalogs that were derived from different datasets using different techniques. The latter group is dominated by objects that have poor LINEAR data (66 objects in total) and thus could not be reliably classified. Their median magnitude and coordinates are distributed roughly equally within the PLV brightness range and observed area. These objects were identified as variable and periodic in the PLV catalog, but the light-curve type could not be determined (they are classified as “Other” in the PLV catalog). Thirteen of the remaining objects with better data were classified as c type RR Lyrae, one was classified as an EB/EW eclipsing binary, one as a BL Herculis candidate, and two as anomalous Cepheids (in the VSX, these two objects were classified as ACEP and ACEP). Therefore, the only true disagreement in classification between LINEAR and DR13 is for those 13 c type RR Lyrae (0.5%). Several examples of light curves for objects where the PLV catalog and DR13 classifications did not match are shown in Figure 28.

Finally, we note that a total of 362 PLV ab type RR Lyrae (from the overlapping area and brightness range) do not show up in DR13. Some examples of these objects are shown in Figure 29.

## REFERENCES

- Becker, A. C., Bochanski, J. J., Hawley, S. L., et al. 2011, *ApJ*, **731**, 17
- Bond, B., Ivezić, Ž., Sesar, B., et al. 2010, *ApJ*, **716**, 1
- Cohen, R. E., & Sarajedini, A. 2012, *MNRAS*, **419**, 342
- Cortes, C., & Vapnik, V. 1995, *Mach. Learn.*, **20**, 273
- Covey, K., Ivezić, Ž., Schlegel, D., et al. 2007, *AJ*, **134**, 2398
- Davenport, J. R. A., Becker, A. C., West, A. A., et al. 2013, *ApJ*, **764**, 62
- Deb, S., & Singh, H. P. 2009, *A&A*, **507**, 1729
- Debosscher, J., Sarro, L. M., Aerts, C., et al. 2007, *A&A*, **475**, 1159
- Dempster, A. P., Laird, N. M., & Rubin, D. 1977, *J. R. Stat. Soc. Ser. B*, **39**, 1
- Devor, J. 2004, *ApJ*, **628**, 411
- Dimitrov, D. P., & Kjurkchieva, D. P. 2010, *MNRAS*, **406**, 2559
- 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
- Drake, A. J., Djorgovski, S. G., Mahabal, A., et al. 2009, *ApJ*, **696**, 870
- Dubath, P., Rimoldini, L., Süveges, M., et al. 2011, *MNRAS*, **414**, 2602
- Eggen, O. J. 1967, *MmRAS*, **70**, 111
- Eyer, L., & Blake, C. 2005, *MNRAS*, **358**, 30
- Eyer, L., Holl, B., Pourbaix, D., et al. 2013, *CEAB*, **37**, 115
- Eyer, L., & Mowlavi, N. 2008, *JPhCS*, **118**, 012010
- Eyer, L., Palaversa, L., Mowlavi, N., et al. 2012, *Ap&SS*, **341**, 207
- Friedman, J. H. 1984, *A Variable Span Smoother* (Stanford, CA: Stanford Univ. Press)
- Giuricin, G., Mardirossian, F., & Messetti, M. 1983, *A&A*, **119**, 218
- Hoffman, D. I., Harrison, T. E., & McNamara, B. J. 2009, *AJ*, **138**, 466
- Ivezić, Ž., Beers, T. C., & Jurić, M. 2012, *ARA&A*, **50**, 251
- Ivezić, Ž., Connolly, A. J., Vanderplas, J. T., & Gray, A. 2013, *Statistics, Data Mining and Machine Learning in Astronomy* (Princeton, NJ: Princeton Univ. Press)
- Ivezić, Ž., & Elitzur, M. 1995, *ApJ*, **445**, 415
- Ivezić, Ž., Goldston, J., Finlator, K., et al. 2000, *AJ*, **120**, 963
- Ivezić, Ž., Sesar, B., Jurić, M., et al. 2008a, *ApJ*, **684**, 287
- Ivezić, Ž., Smith, J. A., Miknaitis, G., et al. 2007a, *AJ*, **134**, 973
- Ivezić, Ž., Smith, J. A., Miknaitis, G., et al. 2007b, in *ASP Conf. Ser.* 364, *The Future of Photometric, Spectrophotometric and Polarimetric Standardization*, ed. C. Sterken (San Francisco, CA: ASP), 165
- Ivezić, Ž., Tyson, J. A., Acosta, E., et al. 2008b, *arXiv:0805.2366*
- Jeon, Y.-B., Lee, M. G., Kim, S.-L., & Lee, H. 2004, *AJ*, **128**, 287
- Lomb, N. R. 1976, *ApS&S*, **39**, 447
- Lupton, R. H., Ivezić, Z., Gunn, J. E., et al. 2002, *Proc. SPIE*, **4836**, 350
- MacLeod, C. L., Ivezić, Ž., Kochanek, C. S., et al. 2010, *ApJ*, **721**, 1014
- MacLeod, C. L., Ivezić, Ž., Sesar, B., et al. 2012, *ApJ*, **753**, 106
- Oosterhoff, P. T. 1944, *BAN*, **10**, 55
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *JMLR*, **12**, 2825
- Pojmański, G. 2002, *AcA*, **52**, 397
- Poleski, R. 2013, *arXiv:1309.1168*
- Prša, A., & Zwitter, T. 2005, *ApJ*, **628**, 426
- Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, *ApJ*, **733**, 10
- Reimann, J. D. 1994, PhD thesis, Univ. of California, Berkeley
- Rodríguez, E., & Breger, M. 2001, *A&A*, **366**, 178
- Ruan, J. J., Anderson, S. F., MacLeod, C. L., et al. 2012, *ApJ*, **760**, 51
- 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., Durevich, O. V., et al. 2009, *yCat*, **1**, 2025
- Scargle, J. D. 1982, *ApJ*, **263**, 835
- Sesar, B., Ivezić, Ž., Grammer, S. H., et al. 2010, *ApJ*, **708**, 717
- Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, *AJ*, **134**, 2236
- Sesar, B., Ivezić, Ž., Stuart, J. S., et al. 2013, *AJ*, **146**, 21 (Paper II)
- Sesar, B., Stuart, J. S., Ivezić, Ž., et al. 2011, *AJ*, **142**, 190 (Paper I)
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, **131**, 1163
- Smolčić, V., Ivezić, Ž., Knapp, G. R., et al. 2004, *ApJL*, **615**, L141
- Spano, M., Mowlavi, N., Eyer, L., et al. 2011, *A&A*, **536**, A60
- Soszyński, I., Dziembowski, W. A., Udalski, A., et al. 2011, *AcA*, **61**, 1
- Soszyński, I., Udalski, A., Szymański, M. K., et al. 2009, *AcA*, **59**, 239
- Stellingwerf, R. F. 1978, *ApJ*, **224**, 953
- Taylor, M. B. 2005, in *ASP Conf. Ser.* 347, *Astronomical Data Analysis Software and Systems XIV*, ed. P. Shopbell, M. Britton, & R. Ebert (San Francisco, CA: ASP), 29
- Tisserand, P. 2012, *A&A*, **539**, A51
- Tisserand, P., Clayton, G. C., Welch, D. L., et al. 2013, *A&A*, **551**, 77
- Tu, X., & Wang, Z. 2013, *RAA*, **13**, 323
- VanderPlas, J., Connolly, A. J., Ivezić, Z., & Gray, A. 2012, in *Proc. Conf. on Intelligent Data Understanding (CIDU)*, ed. K. Das, N. V. Chawla, & A. N. Srivastava, 47
- Watson, C. L. 2006, *SASS*, **25**, 47
- Whitelock, P. A., Feast, M. W., Marang, F., & Groenewegen, M. A. T. 2006, *MNRAS*, **369**, 751
- Wood, P. R., & Sebo, K. M. 1996, *MNRAS*, **282**, 958
- Woźniak, P. R., Vestrand, W. T., Akerlof, C. W., et al. 2004, *AJ*, **127**, 2436
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2012, *AJ*, **140**, 1868
- Wyrzykowski, L., Udalski, A., Kubiak, M., et al. 2003, *AcA*, **53**, 1
- Yan, L., Donoso, E., Tsai, C.-W., et al. 2013, *AJ*, **145**, 55
- York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, *AJ*, **120**, 1579
- Zechmeister, M., & Kürster, M. 2009, *A&A*, **496**, 577