DETECTING VARIABILITY IN MASSIVE ASTRONOMICAL TIME-SERIES DATA. II. VARIABLE CANDIDATES IN THE NORTHERN SKY VARIABILITY SURVEY

MIN-SU SHIN[1], HAHN YI[2], DAE-WON KIM[2,4], SEO-WON CHANG[2], AND YONG-IK BYUN[3]

[1] Department of Astronomy, The University of Michigan, 500 Church Street, Ann Arbor, MI 48109, USA; msshin@umich.edu

[2] Department of Astronomy, Yonsei University, Seoul 120-749, Republic of Korea; yihahn@galaxy.yonsei.ac.kr, dakim@cfa.harvard.edu, seowony@galaxy.yonsei.ac.kr

[3] Department of Astronomy and University Observatory, Yonsei University, Seoul 120-749, Republic of Korea; ybyun@yonsei.ac.kr

ABSTRACT

We present variability analysis of data from the Northern Sky Variability Survey (NSVS). Using the clustering method, which defines variable candidates as outliers from large clusters, we cluster 16,189,040 light curves having data points at more than 15 epochs as variable and non-variable candidates in 638 NSVS fields. Variable candidates are selected depending on how strongly they are separated from the largest cluster and how rarely they are grouped together in eight-dimensional space spanned by variability indices. All NSVS light curves are also cross-correlated with *IRAS*, AKARI, Two Micron All Sky Survey, Sloan Digital Sky Survey (SDSS), and *GALEX* objects, as well as known objects in the SIMBAD database. The variability analysis and cross-correlation results are provided in a public online database, which can be used to select interesting objects for further investigation. Adopting conservative selection criteria for variable candidates, we find about 1.8 million light curves as possible variable candidates in the NSVS data, corresponding to about 10% of our entire NSVS sample. Multi-wavelength colors help us find specific types of variability among the variable candidates. Moreover, we also use morphological classification from other surveys such as SDSS to suppress spurious cases caused by blending objects or extended sources due to the low angular resolution of the NSVS.

Key words: astronomical databases: miscellaneous – methods: data analysis – methods: statistical – stars: variables: general

1. INTRODUCTION

Emerging projects in time-domain astronomy produce a large amount of time-series data and allow discoveries of new variable sources and better understanding of known variability types (see Paczyński 2000; Djorgovski et al. 2001; Bono et al. 2003, for a review). This new era needs a computationally intensive processing of massive time-series data with computational algorithms to recover interesting objects with a broad range of variability types (Eyer 2006).

Investigating time variability of astronomical objects begins with detecting any significant changes in brightness. Detection methods can be optimized to specific variability types. For example, the image subtraction method is commonly used to detect supernovae, microlensing, and other variable sources (e.g., Alard & Lupton 1998; Wozniak 2000; Gössl & Riffeser 2002; Becker et al. 2004; Corwin et al. 2006; Yuan & Akerlof 2008). Instead of detecting variability in images, recognizing variability in light curves can also be tailored to a particular type of variability signal such as a transit (e.g., Protopapas et al. 2005; Renner et al. 2008).

In a previous paper (Shin et al. 2009, hereafter Paper I), we presented a new framework of detecting general variability types in massive time-series data by using a non-parametric infinite Gaussian mixture model (GMM). In our approach, variable objects are considered outliers from non-variable objects that should constitute a dominant fraction of given data in multi-dimensional space spanned by several variability indices. In the results from the infinite GMM where each group is described by a multivariate Gaussian distribution (Robert 1996), we recognize large groups as groups of non-variable objects, tagging outliers from these large groups as possible variable object candidates (see Liao 2005 for other methods of clustering). The strength of our approach is based on the assumption that non-variable objects, which do not have enough signals of variable phenomena in their light curves, represent the dominant fraction of data and share the same systematic effects hidden in the given data such as sampling patterns and noise properties. Therefore, by extracting common properties of dominant non-variable objects from the given data, our approach can be less biased than choosing specific types of variability with assumptions of systematic patterns.

In this paper, the second in a series, we apply our methodology to all data from the Northern Sky Variability Survey (NSVS; Woźniak et al. 2004a). The NSVS catalog includes about 14 million[5] objects with optical magnitudes ranging from 8 to 15.5 and declinations higher than $-38°$. These objects are so bright that deeper imaging surveys such as Pan-STARRS (Kaiser et al. 2010) and the Large Synoptic Survey Telescope (Tyson 2002) cannot produce useful photometric data owing to saturation in their normal observation modes. Therefore, variability analysis of the current NSVS data will still be useful for the investigation of bright stars even after deeper images are acquired in future surveys. Moreover, the NSVS data have not been fully exploited because several previous trials have been focused on particular types of variable objects with specific criteria for variability signals (e.g., Woźniak et al. 2004b; Nicholson et al. 2005; Kinemuchi et al. 2006; Wils et al. 2006; Kelley & Shaw 2007; Kiss et al. 2007; Hoffman et al. 2008, 2009; Usatov 2008; Dimitrov 2009; Schmidt et al. 2009).

---

[4] Also at Harvard Smithsonian Center for Astrophysics, Cambridge, MA 01238, USA.

[5] Because some NSVS objects are included in multiple separate light curves from different observation fields, the total number of light curves is larger than that of the total number of objects.

Table 1
Variability Indices

Index	Definition
$\frac{\sigma}{\mu}$	$\dfrac{\sqrt{\sum_{n=1}^{N}(x_n-\mu)^2/(N-1)}}{\sum_{n=1}^{N} x_n/N}$
γ_1	$\dfrac{\sqrt{N(N-1)}}{N-2}\dfrac{\sum_{n=1}^{N}(x_n-\mu)^3/N}{\sqrt[3]{\sum_{n=1}^{N}(x_n-\mu)^2/N}}$
γ_2	$\dfrac{N-1}{(N-2)(N-3)}\left\{(N+1)\left(\dfrac{\sum_{n=1}^{N}(x_n-\mu)^4/N}{(\sum_{n=1}^{N}(x_n-\mu)^2/N)^2}-3\right)+6\right\}$
Con	$1+\dfrac{1}{N-2}\sum_{n=1}^{N-2}\begin{cases} 1 & \text{if } (x_n-\mu_0)>2\sigma \text{ and } (x_{n+1}-\mu_0)>2\sigma \text{ and } (x_{n+2}-\mu_0)>2\sigma \\ 1 & \text{if } (x_n-\mu_0)<-2\sigma \text{ and } (x_{n+1}-\mu_0)<-2\sigma \text{ and } (x_{n+2}-\mu_0)<-2\sigma \\ 0 & \text{otherwise} \end{cases}$
η	$\dfrac{\sum_{n=1}^{N-1}(x_{n+1}-x_n)^2/(N-1)}{\sigma^2}$
J	$\sum_{n=1}^{N-1}\text{sign}(\delta_n\delta_{n+1})\sqrt{\lvert\delta_n\delta_{n+1}\rvert}$
K	$\dfrac{1/N\sum_{n=1}^{N}\lvert\delta_n\rvert}{\sqrt{1/N\sum_{n=1}^{N}\delta_n^2}}$
AoVM	The maximum value of the analysis of variance (ANOVA) statistic (Schwarzenberg-Czerny 1996)

Notes. σ, μ, γ_1, γ_2, and μ_0 are standard deviation, average, skewness, kurtosis, and median of N magnitudes x_n in each light curve, respectively. δ_n is $\sqrt{N/(N-1)}(x_n-\mu)/e_n$, where e_n is a photometric error for each data point. $\text{sign}(\delta_n\delta_{n+1})$ is the sign of $\delta_n\delta_{n+1}$.

Most NSVS objects are already included in other previous surveys. In particular, the NSVS catalog can cover most of the optical magnitude range of objects detected in *IRAS* (Clegg 1980), Two Micron All Sky Survey (2MASS; Kleinmann 1992), and *GALEX* (Bianchi & The GALEX Team 1999) observations. Since colors can be used to identify the particular types of spectral energy distributions, associating variability analysis with colors can be an effective way to discover new variable sources at a specific phase of stellar evolution such as Mira-type variables (e.g., Pojmanski & Maciejewski 2005) or extragalactic variable objects such as quasars (e.g., Bianchi et al. 2007). Associating the NSVS objects with other optical surveys is also critically important. Since the NSVS data do not have a high angular resolution with small telescopes, a large fraction of objects might be affected by blending, poor positioning, and incorrect identification of extended objects as stars in the NSVS data. Our new variability analysis of all the NSVS data will help others find specific variable sources by combining their data and these other multi-wavelength surveys.

This paper is organized as follows. In Section 2, we explain the application of the infinite GMM to the NSVS data. In Section 3, we describe how to extract variable candidates by using the results from the GMM. In Section 4, we explain the Web database of our variability analysis. Properties of variable candidates are explained in Section 5, considering their associations with archival data. Finally, a summary and discussion are given in the last section.

2. METHOD

2.1. Data and Variability Indices

We extract 16,189,040 light curves from the NSVS data (Woźniak et al. 2004a) by limiting our selection to samples that have more than 15 good photometric data points. The good photometric points are defined as not having bad photometry flags: SATURATED, NOCORR, LONPTS, HISCAT, HICORR, HISIGCORR, and RADECFLIP (see Woźniak et al. 2004a for meanings of the flags). Since one object can be included as several light curves in different observation fields in the NSVS catalog, the number of objects is slightly smaller than that of the

light curves in our sample. Here we consider each light curve as a separate entity. Among 644 observation fields of the NSVS data, 6 (123c, 145c, 146c, 147c, 156c, 156d) do not have any light curves having more than 15 good photometric data points.

In addition to the six variability indices used in Paper I, we also derive skewness and kurtosis from each light curve. The definitions of these indices are summarized in Table 1. σ/μ, γ_1, and γ_2 are not sensitive to the structure of light curves. However, estimating them is computationally cheap, describing simple low-order patterns of light curves easily. We note that our definitions of skewness and kurtosis are not the same as the traditional measures (see Joanes & Gill 1998 for a discussion). The other five indices describe more complex patterns in light curves. Con represents how many sets of three consecutive data points are at least 2σ fainter or brighter than the median magnitude, tracing continuous variations in light curves (Wozniak 2000). η measures the ratio of the mean square successive difference from the sample variance (von Neumann 1941). J and K have been commonly used for multi-band light curves although these can be estimated for single-band light curves (Stetson 1996). Here we use only single-band light curves with a slightly modified definition that uses sequential pairs of data points. Finally, we also use the analysis of variance (ANOVA) statistic, which is useful for identifying periodic signals (Schwarzenberg-Czerny 1996; Shin & Byun 2004). The maximum value of the ANOVA (AoVM) is used to measure the strength of periodicity. Even with an incorrect period of the light curve, AoVM can be a valuable quantity that infers periodicity (Shin & Byun 2007).

2.2. GMM and Results

We follow the same procedure for the infinite GMM as in Paper I. The GMM is derived for each NSVS observation field with the eight variability indices. Even though the GMM converges much earlier than 100 iterations, we conduct 100 iterations as shown in Paper I. Since finding large groups is our main goal for detecting non-variable objects from the samples, the number of iterations is not an important factor that affects how many small groups can be recovered. Each GMM
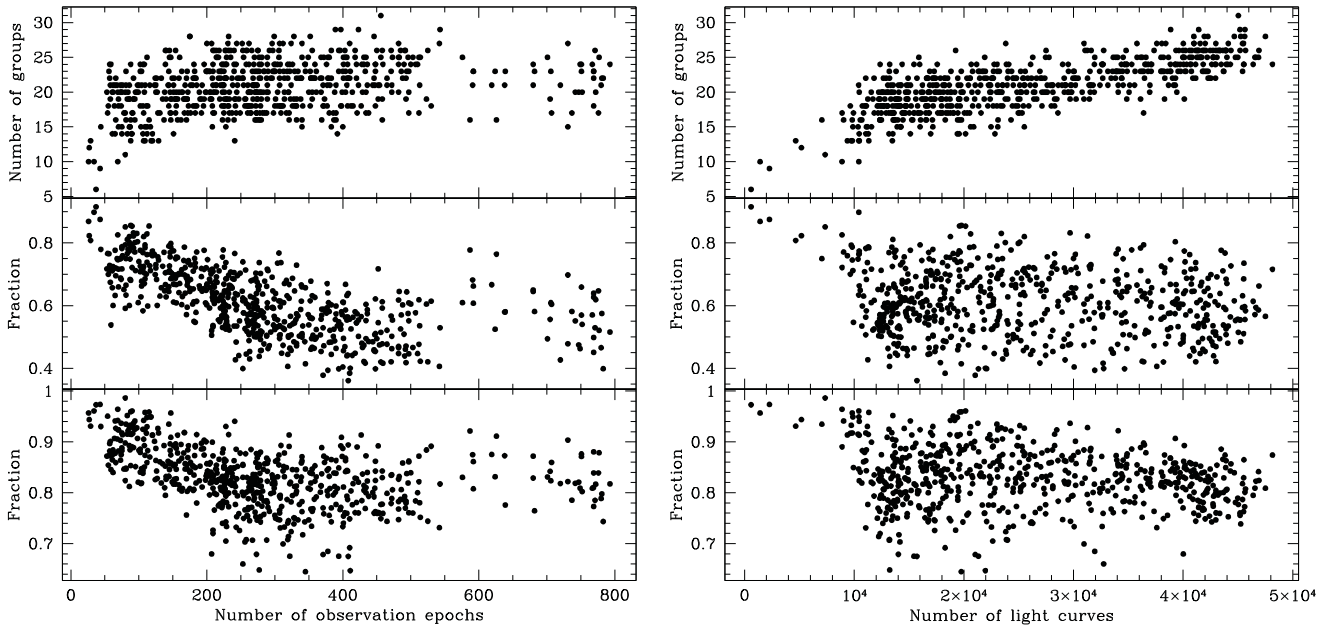
Figure 1. GMM results with respect to the number of observation epochs (left) and the number of light curves (right). From top to bottom, each panel shows the number of groups, the fraction of the largest group, and the fraction of the top three largest groups, respectively.

component[6] is described by a multivariate Gaussian distribution with its mean, i.e., center, and covariance matrix:

$$p_m(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{m}})^T \Sigma_m^{-1}(\mathbf{x} - \mu_{\mathbf{m}})\right), \quad (1)$$

where m is an index of a mixture component, $\mathbf{x} = (\sigma/\mu, \gamma_1, \gamma_2, \mathrm{Con}, \eta, J, K, \mathrm{AoVM})$ is a vector of parameters, and D is the number of parameters (in our case $D = 8$). Furthermore, $\mu_{\mathbf{m}}$ is a vector of mean values (i.e., mixture centers) and Σ_m is the covariance matrix of the Gaussian distribution.

Figure 1 shows how many groups are recovered with the infinite GMM and how many light curves are included in large groups. Since a small number of observation epochs provide poor sampling of light curves and have a low probability of detecting variability, fields having few observation epochs (i.e., few observed frames) are expected to have a small number of groups, while large groups dominate the whole population of light curves. In fields with a large number of light curves, the infinite GMM recovers many groups because these fields are likely to include various kinds of variable light curves. However, the fraction of data included in the large groups does not change simply as a function of the number of light curves because the total number of light curves in the field does not affect the probability that a single light curve represents a non-variable object.

We measure the Davies–Bouldin (DB) index in each field to check systematic differences of the GMM results. The DB index is commonly used to measure the compactness of clusters and separations among them (Davies & Bouldin 1979; Vendramin et al. 2010). The index is defined as

$$\mathrm{DB} = \frac{1}{n_c} \sum_{i=1}^{n_c} \max_{j=1,\ldots,n_c, i\neq j} \left(\frac{s_i + s_j}{d_{ij}}\right), \quad (2)$$

where n_c is the number of clusters, s_i is the average distance of data points included in the cluster i with respect to its center, and d_{ij} is the distance between the centers of two clusters i and j. Distance is defined as the L^2 norm, which is one of the simple measurements (see Yu et al. 2006, for a discussion). As each cluster is compact and well separated from others, this DB index decreases.

Strong systematic differences in the DB index do not appear among the 638 NSVS fields as shown in Figure 2. In most fields, the groups in the GMM results have a DB index smaller than 10. We do not find any strong systematic dependence of the DB index on the number of frames and the number of light curves in each field. The highest DB index ∼29 is found in the field 100c, which has a small number of frames and light curves. Moreover, in that field the number fraction of the top three groups is highest among the sample fields.

2.3. Largest Cluster

The center coordinates of the largest cluster show a significant variation for different fields. Since each NSVS field has different characteristics, it is not surprising to see this variation shown in Figure 3. Importantly, γ_1 and γ_2 are not close to zero, which is in contrast to the expectation assuming a normal distribution for light curves. These non-zero γ_1 and γ_2 imply that a dominant fraction of light curves is not well described by a normal distribution owing to sampling or systematic observational effects. The same implication is also found with η, which has an expected average ∼2 for a normal distribution (Williams 1941). Figure 3 proves the importance of including several variability indices that catch different features of light curves with different sensitivities as we showed in Paper I.

We also examine covariance matrices of the largest cluster in each field. Comparing the absolute values of the covariance matrix elements, the variances of J or AoVM have maximum values in the largest cluster of all 638 fields. In 604 fields, the variances of J range from about 20 to 800, but the variances of AoVM in the rest of the fields range from about 20 to 400.
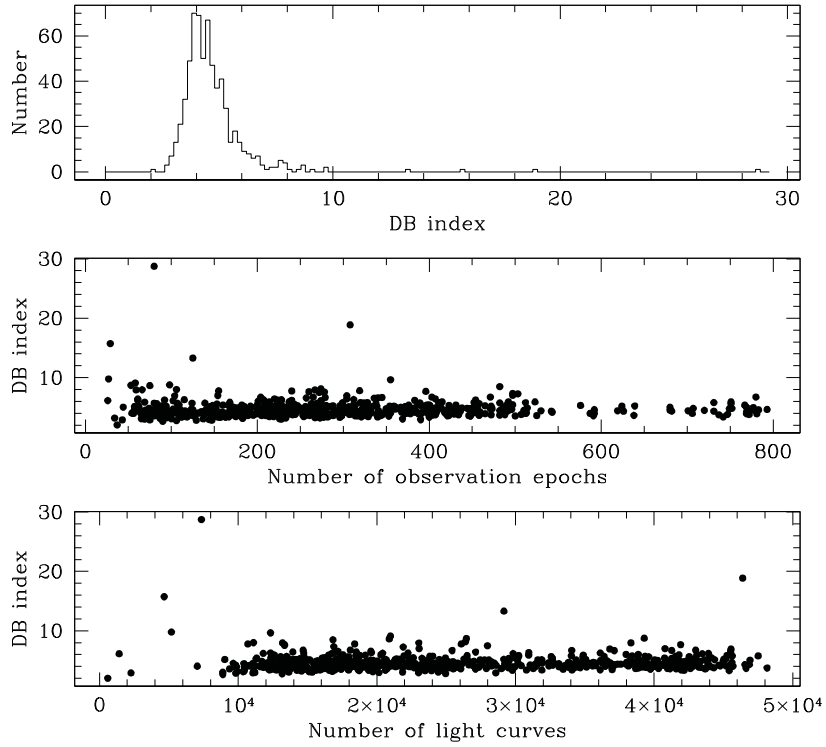
---

[6] We use *component* interchangeably with the terms *cluster* and *group* in this paper. *Component* is a specific term for the GMM.

Figure 2. Distributions of the DB index. When the DB index is low, groups are compact and well separated from others. The NSVS field 100c shows the highest DB index ~29.
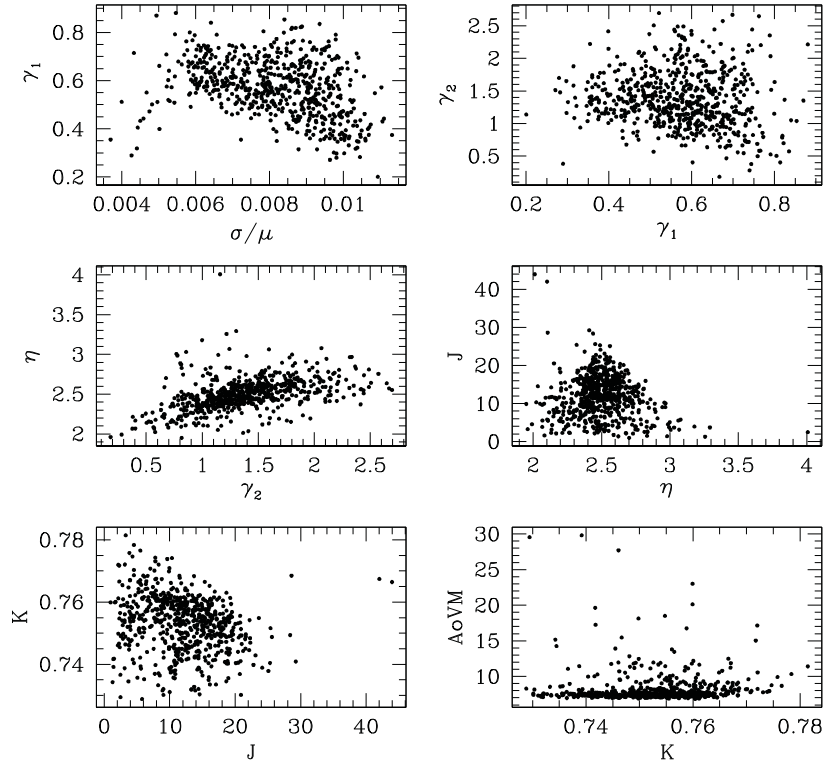

Figure 3. Center coordinates of the largest cluster. Since Con is 1 in all fields, we do not show its distribution here. The distribution shows that there are field-by-field variations of systematic effects that produce variations in the largest cluster's central position in eight-dimensional space.

The largest clusters in the remaining 34 fields also generally have a higher central value of AoVM than that of other fields. In the covariance matrix of the largest cluster, the covariance between η and J has the smallest negative value from about -6 to 0. Only a minor fraction of fields shows the smallest negative covariance between η and AoVM. As shown in Paper I, a substantial number of light curves exhibit negative correlations between η and J, or between η and AoVM in the NSVS data, implying systematic properties of the NSVS data.
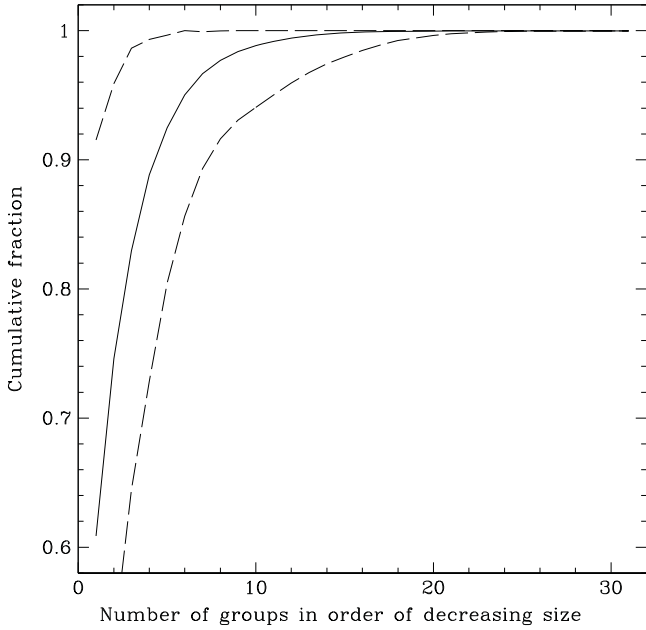
Figure 4. Cumulative fraction of objects included in groups. The median fraction as a function of the number of groups is presented with the solid line, while dashed lines describe the minimum and maximum fractions. For example, the top four largest groups explain about 89% of light curves as the median fraction. The median fraction of about 99% is found within the top 10 largest groups.
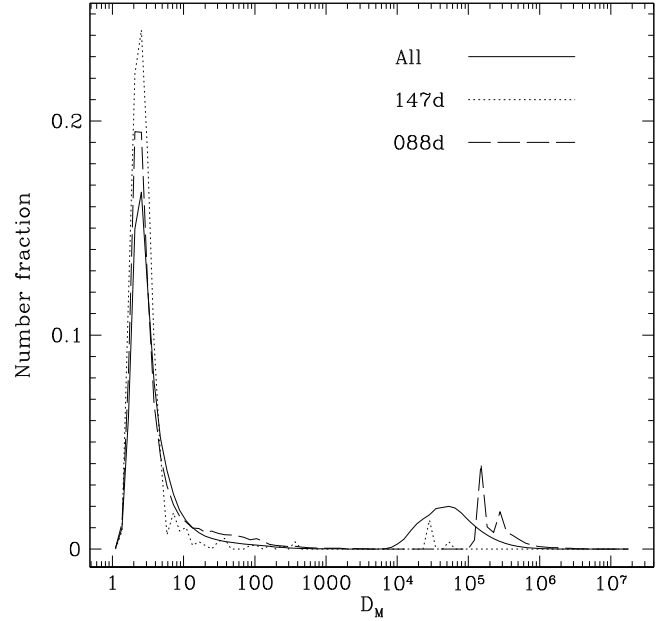

Figure 5. Distributions of D_M. The peak of the distributions around $D_M \sim 2$ corresponds to a mode value of D_M for a multivariate normal distribution (Ververidis & Kotropoulos 2008). Compared with the distribution for all objects (solid line), the field 147d (dotted line), which has the smallest number of light curves, and the field 088d (dashed line), which has the largest number of light curves, in our samples are more dominated by objects around the largest cluster in eight-dimensional space.

3. SEPARATION OF VARIABLE CANDIDATES

One disadvantage of clustering algorithms is that there is no useful validation process for clustering results as there is in other unsupervised learning methods. Our approach using the infinite GMM shares this problem. Therefore, there is no one absolutely right way to define a boundary between variable and non-variable objects in multi-dimensional space using the clustering results. In many situations, the selection rule of variable objects can be limited by practical issues such as the number of objects that can be investigated further in follow-up studies. Here we suggest three possible ways to separate variable objects from non-variable objects by using clustering results.

As suggested in Paper I, first, we can select objects that are not included in large groups as variable candidates.[7] Figure 1 shows that the largest group in each field includes a different number of objects. Therefore, defining objects in the largest group as non-variable objects produces different numbers of variable candidates in each field. Moreover, if there were any systematic patterns in light curves, or if the majority of non-variable light curves was not well described by a multivariate Gaussian distribution, data in each field would make the second or third largest groups include non-variable objects or objects affected by the systematic patterns in their light curves. Basically, the existence of small clusters, i.e., the exact number of clusters, is not an important factor in selecting variable candidates.

Figure 4 presents the change in cumulative fractions of data included in groups as the number of large groups increases. Since we find only six clusters in the field 147d, the fraction of data included in the top four groups is about 99%. However, the median fraction of data included in the top four largest groups is about 89%. To select variable candidates from minor groups in

each field, one could adopt a 90% cut of the cumulative fraction and avoid light curves included in large groups.

Another possible approach is using distances of objects from the center of the largest group in the eight-dimensional space spanned by the variability indices. In Paper I, we introduced the Mahalanobis distance from the largest cluster:

$$D_M = \sqrt{(\mathbf{x} - \mathbf{\mu_0})^{\mathbf{T}} \mathbf{\Sigma_0^{-1}} (\mathbf{x} - \mathbf{\mu_0})}, \tag{3}$$

where the center μ_0 and covariance matrix Σ_0 of the largest cluster are used with the position of an individual object $\mathbf{x}$ in the eight-dimensional space. Unlike a commonly used Euclidean distance, D_M depends on Σ_0, which describes how broadly the objects in the largest cluster disperse (Sharma & Johnston 2009). Objects with large distances can be considered variable objects because they have strong dissimilarity from the dominant fraction of light curves (e.g., Jolion et al. 1991). In Figure 5, the distribution of D_M for all objects with respect to the largest cluster in each field shows a peak of the distribution around $D_M \sim 2$ corresponding to the mode value of the beta distribution, which is expected for the distribution of D_M (Ververidis & Kotropoulos 2008). It also shows another concentration of objects between $D_M \sim 10^4$ and 10^6, implying the possible signature of a large number of variable candidates or any systematic patterns hidden in the NSVS light curves. This feature varies strongly in each field as shown for the fields 088d and 147d in Figure 5.

We can derive the cut of D_M that includes a specific fraction of members in the largest cluster by integrating the multivariate Gaussian distribution with the largest cluster's center μ_0 and covariance matrix Σ_0. For example, the b% cut of D_M can be

---

[7] We warn that this approach does not work if the most dominant fraction of light curves corresponds to variable sources.
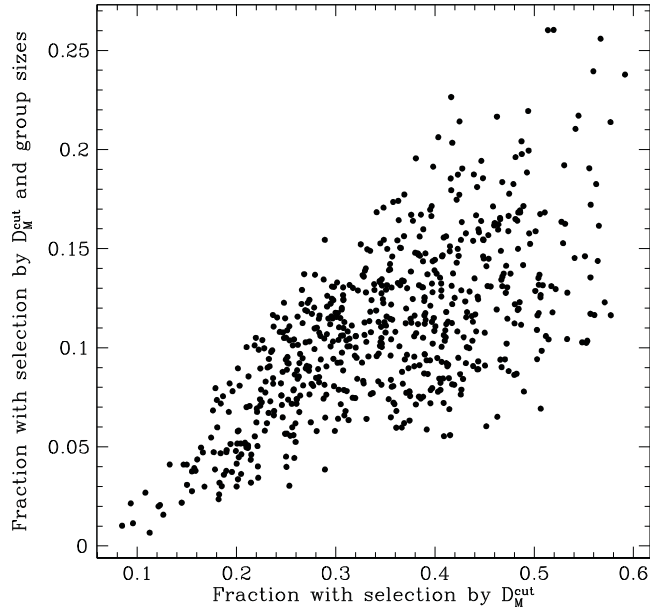
Figure 6. Fraction of objects selected as variable candidates. As found in the different ranges of the horizontal axis and the vertical axis, the size of variable candidates can be decreased significantly by combining the constraints of D_M^{cut} and group sizes together.

found with

$$\int_{\mathbf{x}:D_M(\mathbf{x}) < D_M^{cut}} p(\mathbf{x}|\mu_0, \Sigma_0)\mathbf{dx} = \mathbf{b}, \qquad (4)$$

where $p(\mathbf{x})$ is a multivariate Gaussian distribution. Practically, the integration can be estimated by using the Monte Carlo method (Chen et al. 2006). Each field has a different value of D_M^{cut} for the same probability cut. For example, the 99% cut in the field 088d is $D_M^{cut} = 5.634$, while $D_M^{cut} = 5.105$ corresponds to the 99% cut in the field 147d. Since the largest cluster has a sharper concentration of D_M in the field 147d than in the field 088d (see Figure 5), D_M^{cut} in the field 147d is smaller than that in the field 088d.

These two approaches have their own different problems. Avoiding large groups does not guarantee that minor groups are well separated from the largest group. In a multi-dimensional space, a few minor groups can be close to the large groups with a small D_M. When selecting objects with $D_M > D_M^{cut}$ as variable candidates, those objects with a large D_M can form large groups representing systematic observational features in each field. Although Figure 5 implies that members of most large groups have a small D_M, using only one method can cause more contamination of non-variable objects or objects dominated by systematic patterns in each observation field.

We can define variable candidates conservatively by combining both methods. For example, we find objects having a larger D_M than the D_M^{cut} of the 99% cut and not being included in the top four largest groups. Figure 6 represents the fraction of objects selected by this conservative method compared with results of selecting objects with the D_M^{cut} of the 99% cut. We use this conservative selection of variable candidates in Section 5. Because the top four largest groups are generally located closely in the eight-dimensional space, excluding the members of the top four largest groups from the variable candidates usually determines the size of the variable candidates.

4. DATABASE

We provide the results of our variability analysis and clustering for all sample light curves online. The database stores the eight variability indices, the cluster identification number in each field, D_M from the largest cluster for every light curve, and the basic information of the light curves such as the NSVS object ID and coordinates. The database is supplemented with the number of light curves analyzed, the number of groups found by clustering, the fractions of group members, and the D_M cuts of 99%, 95%, and 90% from the largest group in each field. Therefore, users of the database can select variable candidates by using these clustering results with their own selection rules for variable candidates.

The database is also supplemented by association with other astronomical catalogs. All objects analyzed are cross-matched to the SIMBAD database, 2MASS All-Sky Catalog of Point Sources (Skrutskie et al. 2006), the photometric catalog of the Sloan Digital Sky Survey (SDSS) Data Release 7 (Abazajian et al. 2009), and the photometric catalog of the *GALEX* GR4/GR5 Data Release (Morrissey et al. 2007) with the NSVS coordinates and a search radius of 6″. When multiple objects are matched within the search radius, the nearest match is associated with the NSVS object. In addition to these catalogs, we also present matching results of the NSVS coordinates with *IRAS* (Helou & Walker 1988; Moshir et al. 1990) and *AKARI* (Murakami et al. 2007) catalogs as separate files online. This association with other catalogs can be used to identify the morphology of the NSVS objects as galaxies or stars, to find astronomical types of known objects, to check blending effects with neighbor objects, and to estimate colors of objects as we show in the following section. In particular, because the spatial resolution of the NSVS data is much worse than that of the data in the SDSS and 2MASS catalogs, the morphological information from these catalogs can help database users avoid blended objects and less precise photometry in the NSVS catalog.

The database can be accessed online.[8] Searching the light curves and their variability analysis is possible with equatorial coordinates and a search radius given by users. In particular, the simple cone search interface[9] is also provided for compatibility with the Virtual Observatory environment (Williams et al. 2008). We plan to provide the basic components of the database in Vizier[10] too.

5. PROPERTIES OF VARIABLE CANDIDATES

In this section, we examine properties of variable candidates that are not included in the top four groups and have D_M larger than the 99% cut. With these conservative selection criteria, we find 1,840,310 light curves that are possible variable candidates. We emphasize that this selection of variable candidates is highly conservative. When we select variable candidates simply with D_M larger than the 95% or 99% cut, the total number of variable candidates is 6,640,387 or 5,826,587, which is about 3.6 or 3.1 times more than the number of variable candidates selected with the conservative definition, respectively. By comparison, when we select objects not included in the top four groups as variable candidates, the number of candidates is 1,918,580, which is similar to the result of the conservative selection. Recall, however, that this selection of variable candidates is still affected

---

[8] http://stardb.yonsei.ac.kr
[9] http://stardb.yonsei.ac.kr/conesearch/nsvs_conesearch.php
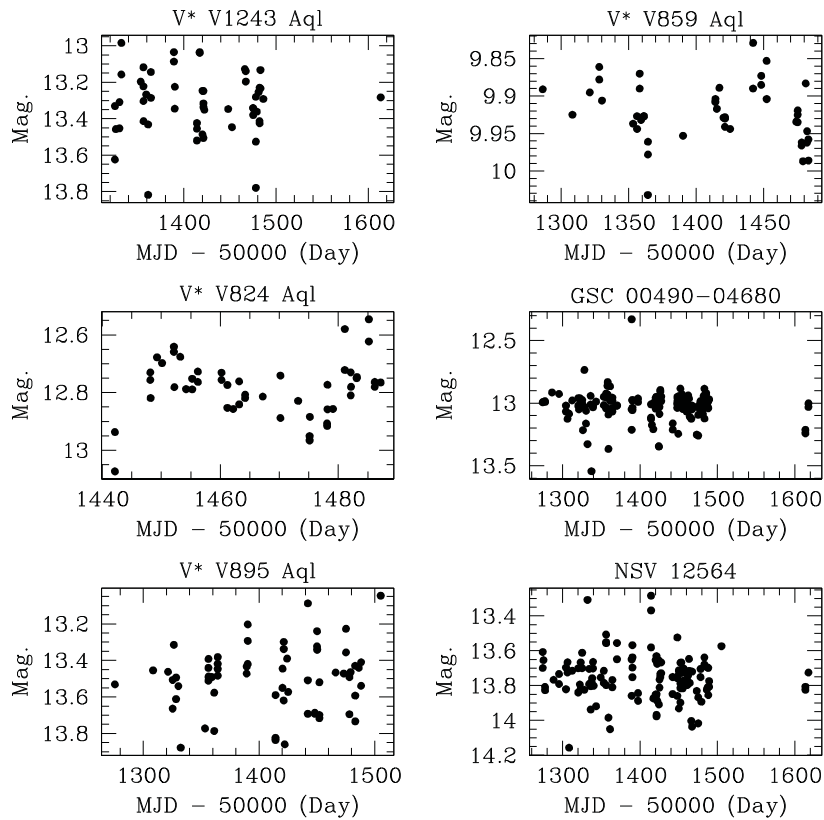[10] http://vizier.u-strasbg.fr/viz-bin/VizieR

Figure 7. Example light curves of known or suspected variable objects that are not selected by our conservative variability detection. The SIMBAD names of the objects are given in the top of each panel.

by the intrinsic limits of the NSVS data such as blending effects and completely different uncertainty properties of photometry for extended objects compared with stellar objects.

5.1. Known Objects and New Candidates

We search all NSVS samples with the SIMBAD database to recognize any known objects and their basic properties such as well-known names. As explained in Section 4, the nearest SIMBAD object around the coordinates of the NSVS objects is retrieved with a search radius of $6''$. Since the NSVS data do not have information about morphological classification such as galaxy and star, the auxiliary information of the SIMBAD database is useful for sorting out spurious variability of extended objects. About 6% of the NSVS objects are matched with at least one SIMBAD object, but this fraction is affected by the precision of the NSVS coordinates.

Based on the object classification given in the SIMBAD database,[11] we find that 16,061 NSVS light curves correspond to known or suspected variable stars. Considering only variable candidates with our conservative selection criteria, the number of known or suspected variable stars is 11,080 among our variable candidates. Finally, excluding known or suspected variable stars as well as known galaxies in the SIMBAD database, the number of new variable candidates is 1,824,123, with our conservative selection. Hereafter, our investigation of variable candidates is limited to these 1,824,123 NSVS objects. We note that uncertain coordinates in either NSVS or the SIMBAD databases can cause us to miss some known galaxies and known or suspected variables. The classifications in the

SIMBAD databases might not be as complete as other catalogs of variable stars such as the AAVSO International Variable Star Index (VSX)[12] (Watson 2006) as we discuss in the Appendix.

Figure 7 shows light curves of six sample NSVS objects that are not included in our conservative selection of variable candidates, but which are known or suspected variables in the SIMBAD database. These examples are among 11 objects, which are not included in our candidates, in NSVS field 112a corresponding to a part of the constellation Aquila. If we selected variable candidates as objects with D_M larger than the 95% cut, four objects, including GSC 00490-04680 (Bernhard & Lloyd 2000) and NSV 12564 (Kinnunen & Skiff 2000) in Figure 7, would be included in variable candidates among the 11 objects. The field 064d misses the largest number of known or suspected variable stars (236 objects) with our conservative selection of candidates. However, its fraction is only 0.6% with respect to the total number of objects that are analyzed by our clustering method. We note that most of these missing objects are identified as variable objects in the *Kepler* field by HATNET, which uses the image subtraction method (Hartman et al. 2004).

Figure 8 presents example light curves of known or suspected variable stars matched with the SIMBAD database and included in our conservative selection of variable candidates for NSVS field 112a. We find 88 known or suspected variable stars among our 6389 variable candidates in field 112a, corresponding to about 1%.

In Figure 9, we show 12 examples of new variable candidates that are matched to any kind of known objects except variable stars in the SIMBAD database. CCDM J19302+0219AB is not

---

[11] http://simbad.u-strasbg.fr/simbad/sim-display?data=otypes
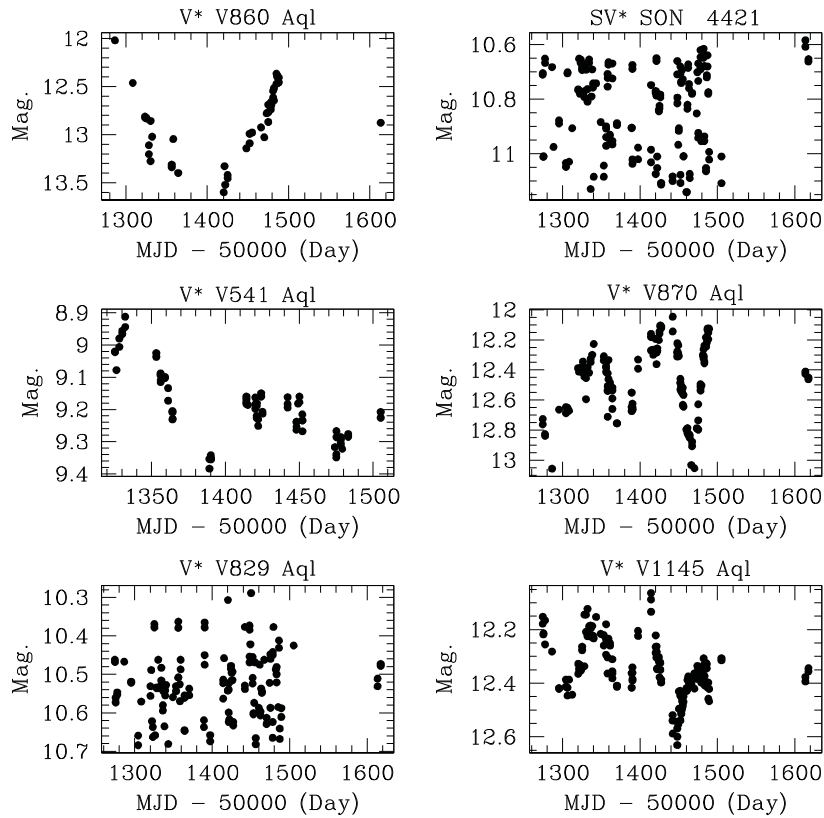
[12] http://www.aavso.org/vsx/

Figure 8. Example light curves for known or suspected variable objects included in our conservative variable candidates. The SIMBAD names of the objects are given in the top of each panel.
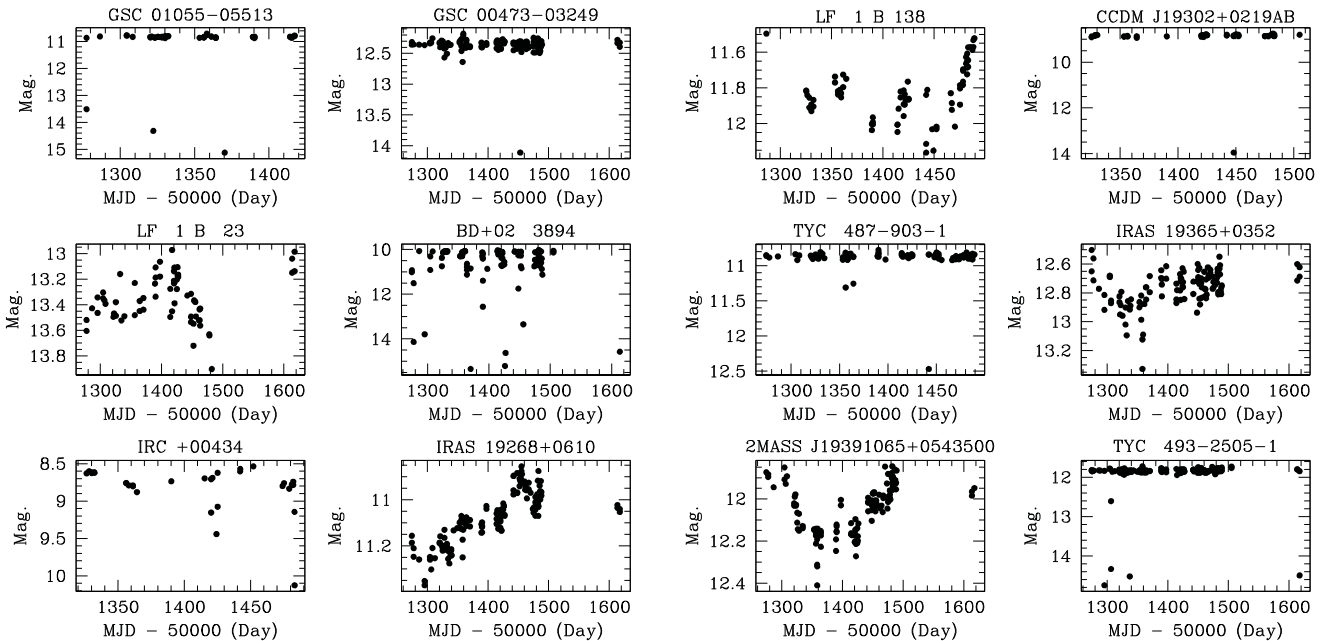

Figure 9. Example light curves of variable candidates matched to SIMBAD objects of non-variable stars. Each light curve corresponds to the SIMBAD object with the name given in the top of each panel. Although IRC +00434 and 2MASS J19391065+0543500 are not found as known variable sources in the SIMBAD database, these objects are found in the VSX catalog (Pojmanski 2002; Usatov & Nosulchik 2008) as we explain in the Appendix.

listed as a variable star in the SIMBAD database, but this object is a known system of double stars (Dommanget & Nys 1994) that might have been affected by blending in the NSVS data. 2MASS J19391065+0543500 is also not included as a variable

star in the SIMBAD database, but Usatov & Nosulchik (2008) suggest that this object is a red asymptotic giant branch (AGB) variable star, supporting the idea that the light curve shown in Figure 9 exhibits true light variation. These examples show that

our conservative selection of variable candidates can catch real variable objects.

We also compare our variable candidates with those found by others using the NSVS data. For example, 785 RR Lyrae candidates were previously reported (Wils et al. 2006). Of these, 781 objects are included in our conservatively selected candidates, but all 785 objects are found when we select variable candidates as objects having larger than the 99% D_M cut. When we select variable candidates with the 95% D_M cut, all new β Lyrae and Algol-type variable candidates from Hoffman et al. (2008) are recovered with our method if they are included in our original NSVS samples. We recover 95% of them with the conservative selection of variable candidates. About 4700 variable candidates of other kinds such as δ Scuti stars and Cepheid objects were also studied with the NSVS data (Hoffman et al. 2009). Again, our selection with the 95% D_M cut recovers most variable candidates except for objects that have different light curves owing to different definitions of good photometric data and different systematic patterns in light curves (Hoffman et al. 2009). These other studies use simple rules such as 0.1 mag dispersion of light curves, which are conservative selection methods for specific types of variable candidates. Therefore, the number of variable candidates is much larger in our approach than in others.

5.2. IRAS Sources

For the conservative selection of variable candidates, we find *IRAS* sources that are spatially matched to the NSVS coordinates with a search radius of 6″ in the *IRAS* point-source catalog (Helou & Walker 1988) and the *IRAS* faint-source catalog (Moshir et al. 1990). Among all NSVS samples, 31,852 light curves have matching *IRAS* sources. Considering only the *IRAS* sources for our conservative variable candidates, the number is 12,987, which is about 41%.

We derive two colors of the *IRAS* sources by using the *IRAS* photometric flux at 12, 25, and 60 μm ($F_{12\,\mu m}$, $F_{25\,\mu m}$, and $F_{60\,\mu m}$). The conventional definition of the *IRAS* colors (e.g., van der Veen & Habing 1988; Olivier et al. 2001; Sevenster 2002) is

$$C_{12/25} = 2.5\log\left(\frac{F_{25\,\mu m}}{F_{12\,\mu m}}\right); \quad C_{25/60} = 2.5\log\left(\frac{F_{60\,\mu m}}{F_{25\,\mu m}}\right),$$

(5)

where we do not apply any color corrections to the fluxes.

The *IRAS* colors have been commonly used for classification of infrared sources. In particular, a two-color diagram like Figure 10 helps us understand what kind of variable candidates show variability that is relevant to the late stage of stellar evolution such as AGB stars with evolved circumstellar dust (Zuckerman 1987; van der Veen & Habing 1988; Kwok et al. 1997; Ramos-Larios et al. 2009). In Figure 10, we plot the colors of *IRAS* sources with the quality number $Q = 3$ at all three wavelengths 12, 25, and 60 μm (Helou & Walker 1988). AGB stars dominate colors of $C_{25/60} < -0.3$, while planetary nebulae and young stellar objects dominate $-0.3 < C_{25/60} < 0.4$ and $0.4 < C_{25/60}$, respectively (Jackson et al. 2002).

In Figure 11, we show example light curves of variable candidates that have corresponding *IRAS* identifications. Although these *IRAS* sources are not known variable stars in the SIMBAD database, some of them have been investigated in various ways without variability information. IRAS 03534+6945 (NSVS 513536) and IRAS 23400+6320 (NSVS 1487299) were found to be Hα-emitting stars in Stephenson (1986) and Coyne
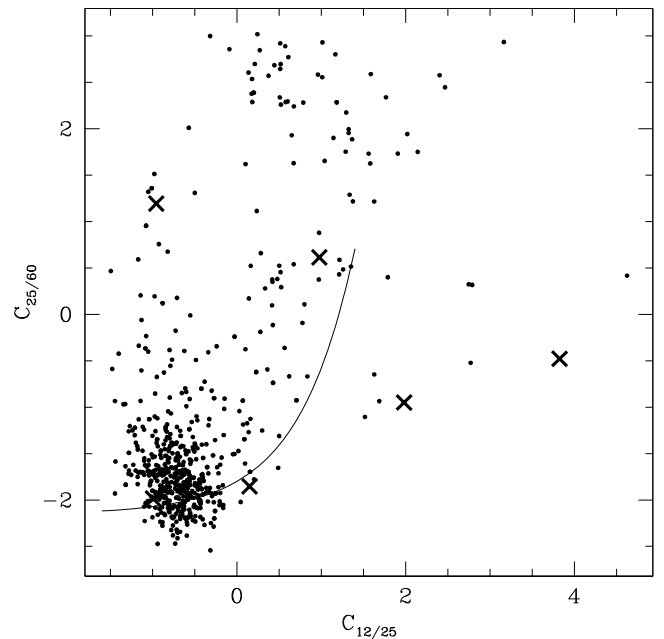

Figure 10. *IRAS* two-color diagram of variable candidates. The solid line represents the color of oxygen-rich Mira variables and variable OH/IR stars from van der Veen & Habing (1988). Cross symbols correspond to objects shown in Figure 11.

& MacConnell (1983), respectively. These sources were also identified as possible variable sources in the VSX catalog (see the Appendix). IRAS 17203-1534 (NSVS 16483061) and IRAS 01005+7910 (NSVS 262162) are post-AGB stars that are sub-classified as hot post-AGB stars and high galactic latitude supergiants (Szczerba et al. 2007), respectively. IRAS 01005+7910 has also been observed with the *Hubble Space Telescope*, which found nebulae around it (Siódmiak et al. 2008).

5.3. AKARI Sources

Bright objects included in this paper are expected to be included in *AKARI* observations, which have been conducted with two instruments: the Far-Infrared Surveyor (FIS) and Infrared Camera (IRC; Murakami et al. 2007). Both instruments produced all-sky source catalogs that are much deeper and spatially better resolved than *IRAS* (Oyabu et al. 2010).

We match our NSVS samples to the *AKARI*/IRC All-Sky Point Source Catalog (Version 1.0; Ishihara et al. 2010a, 2010b) and *AKARI*/FIS All-Sky Survey Bright Source Catalog (Version 1.0; Yamamura et al. 2010), with a search radius of 6″. The numbers of our NSVS samples matching to the *AKARI* objects are 267,732 and 8742 for the IRC and FIS catalogs, respectively. This matching rate with the IRC catalog is much higher than that with the *IRAS*.

Figure 12 shows a color–color diagram of *AKARI* fluxes similar to Figure 10. In the plot, we show objects identified as point sources with only good photometric observations of *AKARI* IRC at 9 and 18 μm, and FIS at 65 μm, which are found with IRC photometric flags of conditions q_S09 = 3, f09 = 0, X09 = 0, q_S18 = 3, f18 = 0, X18 = 0, and with FIS photometric flags of conditions q_S65 > 1. The number of NSVS objects with good photometric data is 374, while only 54 *AKARI* point-source objects with good photometric data correspond to the variable candidates selected conservatively.
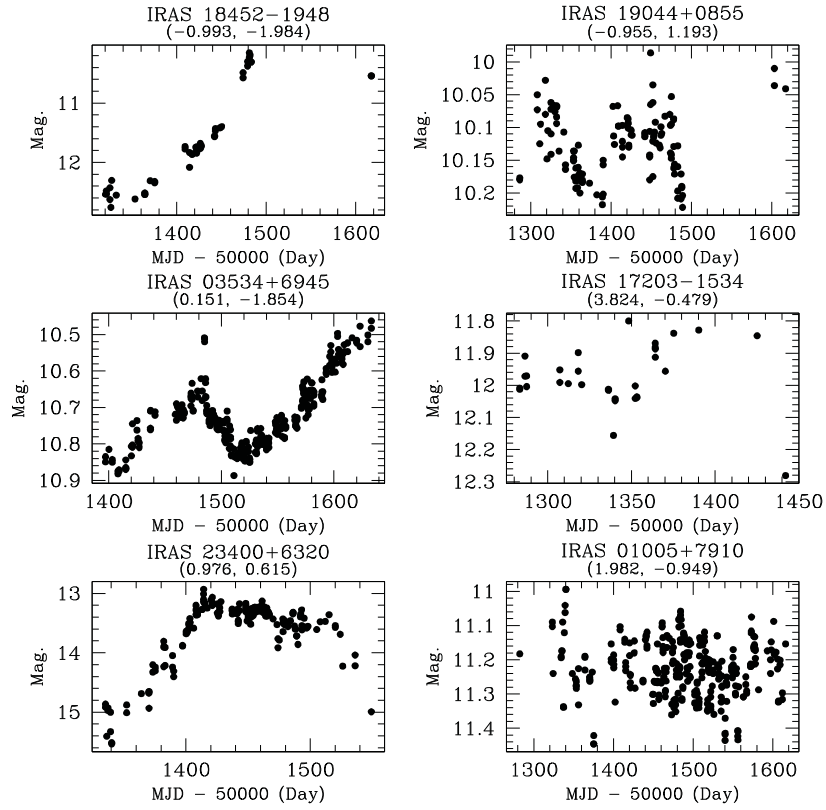
Figure 11. Example light curves of variable candidates that are *IRAS* sources. The *IRAS* designations and $(C_{12/25}, C_{25/60})$ are presented in the top of each panel. Objects in the left column have colors near the curve presented in Figure 10.
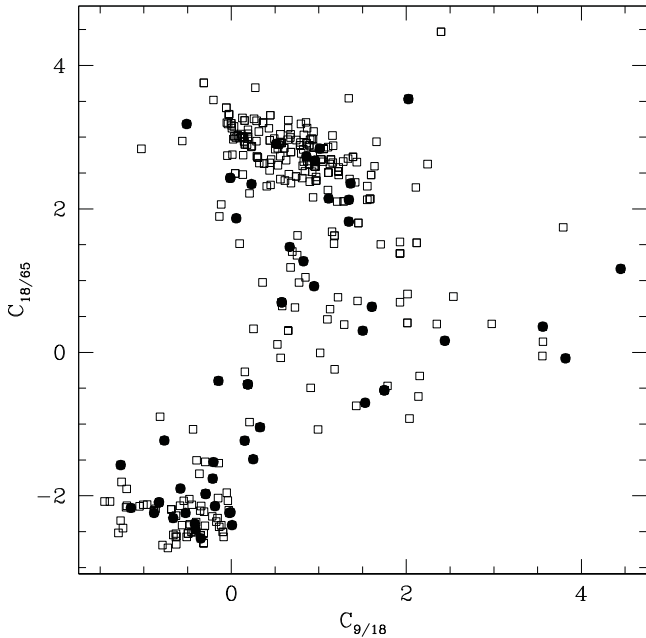

Figure 12. *AKARI* two-color diagram of variable candidates and other objects. Following the definition of the *IRAS* colors given in Equation (5), *AKARI* colors are defined by using fluxes at IRC 9, 18 μm and FIS 65 μm. Filled circles represent new variable candidates selected with our conservative selection, while empty squares correspond to non-variable objects or objects that are known variable stars or galaxies in the SIMBAD database.

The distribution of *AKARI* colors is similar to that of *IRAS* colors because of the similar wavelength ranges of the observation bands.

As with the *IRAS* colors, the variable candidates with corresponding *AKARI* objects might be long-period late-type stars (Ita et al. 2010). Figure 13 shows example light curves of the variable and non-variable candidates in the SIMBAD database. *AKARI* IRC 200011367 (NSVS 1713088) corresponds to TYC 3668-112-1, which is also an *IRAS* object. Even though this light curve does not have many observed data points, it is selected as a possible variable candidate in the NSVS field 013b. *AKARI* IRC 200843875 (NSVS 3393876) is close to IRAS 22164+6427, which might be the same object. The light curve of *AKARI* IRC 200752804 does not have many observed data points, but its variation seems reasonable because of the fact that it is a post-AGB star or a protoplanetary nebula, corresponding to HD 331319 and IRAS 19475+3119. The light curves of the other three objects presented in Figure 13 do not correspond to any known objects in the SIMBAD database and are not selected as variable objects.

5.4. 2MASS and SDSS Photometry

Near-infrared (NIR) colors are also commonly used to identify basic properties of stars and to separate non-stellar objects such as quasars. We match all variable candidates to the 2MASS All-Sky Catalog of Point Sources (Skrutskie et al. 2006) with a search radius of $6''$. A total of 1,439,381 variable candidates have corresponding 2MASS sources. Hereafter, 2MASS photometry data are given in the Vega system.

Figure 14 shows colors of the matched 2MASS objects with unblended, unsaturated, and accurate photometry that is described by read flag Rflg = 2, blend flag Bflg = 1, and contamination and confusion flag Cflg = 0 in all three bands of 2MASS data (Covey et al. 2007), and with a separation between the NSVS and 2MASS positions less than $1''$. Most variable
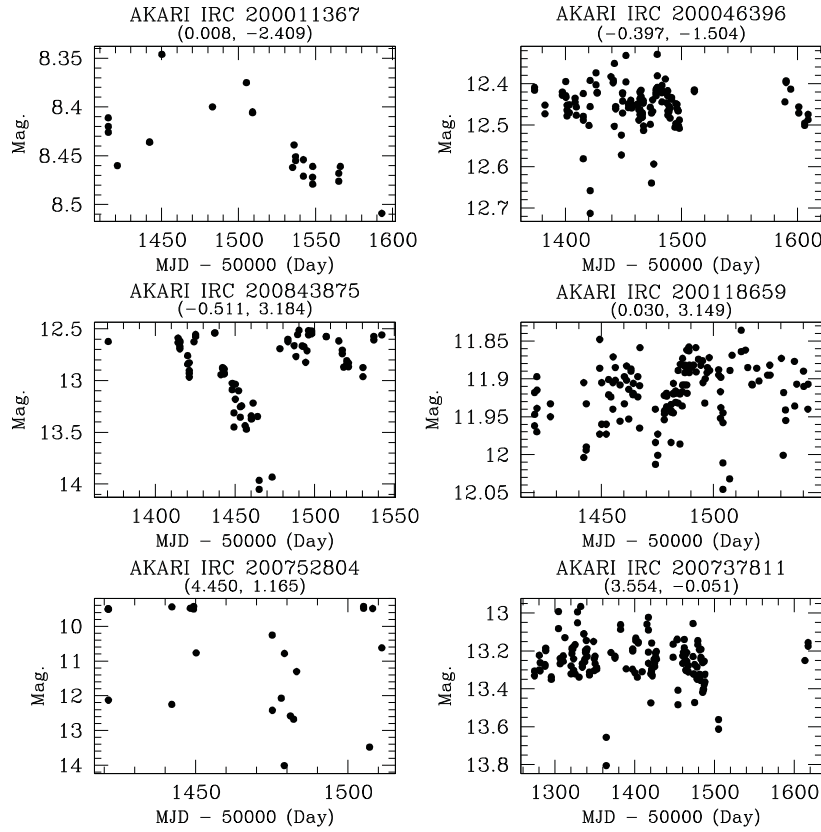
Figure 13. Example light curves of variable candidates (left) and non-variable objects (right) that are *AKARI* sources. The *AKARI* designations in the IRC catalog and $(C_{9/18}, C_{18/65})$ are presented in the top of each panel.
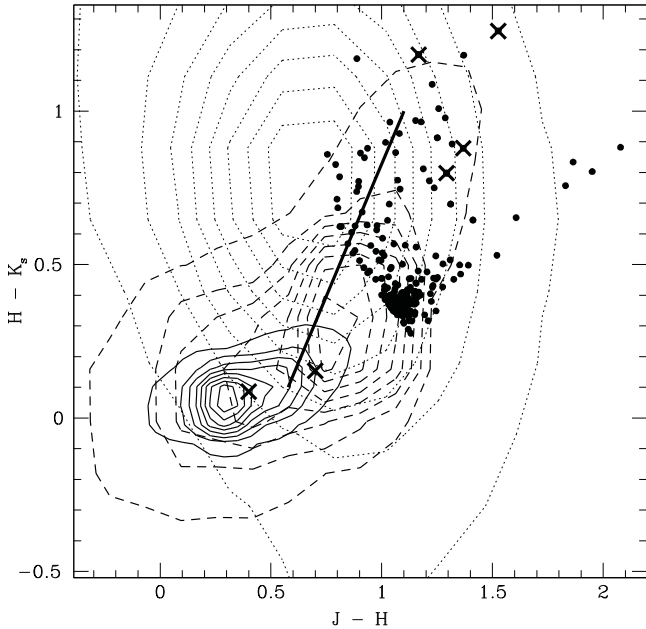

Figure 14. 2MASS color–color diagram of variable candidates. The color distribution of variable candidates (solid-line contours) is similar to the overall color distribution of ordinary stars (Covey et al. 2007), while observed colors of 6658 QSOs with redshifts >0.3 from Véron-Cetty & Véron (2006, dotted-line contours) are not similar to those of variable candidates. The color distribution of known pulsating variable stars in LMC and SMC (dashed-line contours, Ita et al. 2004) implies that variable candidates redder than $(J - K_s) = 1.4$ (dots) might be variable carbon-rich stars. The loci of classical T Tauri stars with de-reddened colors are presented as a thick solid line (Meyer et al. 1997). Cross symbols correspond to objects shown in Figure 15. The Galactic extinction is not considered here.

candidates have colors similar to those of normal stars in our Galaxy, which are mainly $0 < J - H < 1$ (e.g., Finlator et al. 2000; Zoccali et al. 2003). Although quasars have distinctive colors in $J - H$ versus $H - K_s$ where most variable candidates are not found (see Chiu et al. 2007 and Kouzuma & Yamaoka 2010 for discussions), some fraction of variable candidates might be quasars with colors around $J - H \sim 0.9$ and $H - K_s \sim 0.3$.

Late-type variable candidates such as red giant branch (RGB) and AGB stars can be identified more reliably with 2MASS colors. As shown in Figure 14, the dominant NIR color of pulsating variable stars found in the Magellanic Clouds (Ita et al. 2004) is different from the major colors of our variable candidates and known quasars. In particular, $J - K_s \sim 1.4$ is the boundary between oxygen-rich and carbon stars (Nikolaev & Weinberg 2000; Cole & Weinberg 2002; Kiss & Bedding 2003). Therefore, our variable candidates with colors similar to those of known variables might be pulsating RGB and AGB stars (Ita et al. 2004; Kouzuma & Yamaoka 2009).

By using the NIR color–color diagram given in Figure 14, we can also investigate whether our variable candidates include possible obscured young stars with or without disks such as T Tauri stars (Meyer et al. 1997; Tsujimoto et al. 2002; Ozawa et al. 2005). If reddening is significant in some of our variable candidates, their colors might be consistent with those of young stars.

Figure 15 shows six example light curves of the NSVS objects that have corresponding 2MASS measurements. 2MASS 18552297+0404353 (NSVS 13924374) is a Herbig Ae/Be candidate star (Vieira et al. 2003) with a different name—PDS 551. The source 2MASS 09322353+1146033 (NSVS 10229563 and IRAS 09296+1159) is a post-AGB star (Blommaert et al. 1993). We find that its variation recorded in
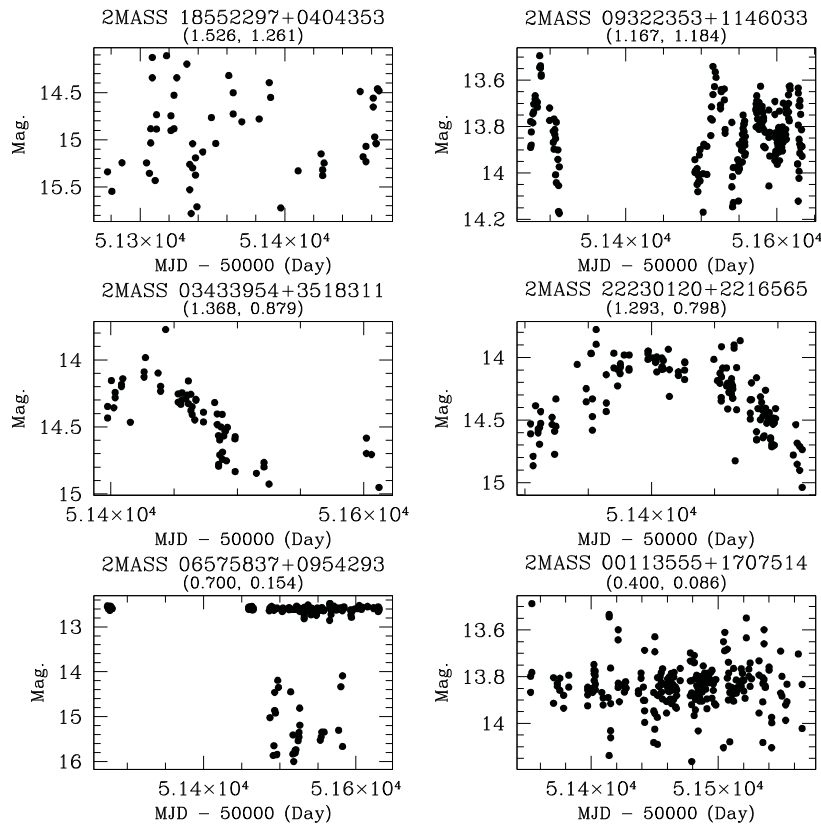
Figure 15. Example light curves of the NSVS objects with reliable 2MASS photometry. The 2MASS designations and $(J - H, H - K_s)$ are given in the top of each panel. 2MASS 18552297+0404353 is also PDS 551, which is a Herbig Ae/Be candidate star (Vieira et al. 2003). 2MASS 09322353+1146033 corresponds to IRAS 09296+1159.

the NSVS light curve is regular with a period of about 46.88 days.[13] 2MASS 22230120+2216565 (NSVS 11767619) is confirmed as a carbon star in spectroscopic observations by Mauron et al. (2007), but other objects presented in Figure 15 have not been assigned a type.

We also match the NSVS coordinates of the variable candidates to SDSS Data Release 7 with a 6″ search radius. The five bands of the SDSS photometric systems have been commonly used to identify stellar and non-stellar sources by using their distinctive colors (e.g., Fan 1999; Fukugita et al. 2011). Combining the 2MASS photometry with the SDSS photometry also helps us precisely determine stellar source types (e.g., Finlator et al. 2000). Among 438,087 variable candidates having corresponding SDSS photometric objects, 406,564 candidates also have corresponding 2MASS sources.

Figure 16 presents the distribution of colors for the variable candidates. We plot only good SDSS photometric data and clean 2MASS photometric data as we explained earlier. In particular, the matched objects with less than 1″ distance are shown in the plot. The good SDSS photometric data are defined as stellar (i.e., unresolved) objects without the SDSS photometric flags EDGE, BLENDED, PEAKCENTER, NO-PROFILE, COSMIC_RAY, SATURATED, NOTCHECKED, DEBLENDED_AS_MOVING, SATUR_CENTER, INTERP_CENTER, DEBLEND_NOPEAK, and PSF_FLUX_INTERP (Stoughton et al. 2002). We check these flags in each SDSS band. Therefore, the number of objects shown in each panel of Figure 16 varies for different color combinations. In all SDSS
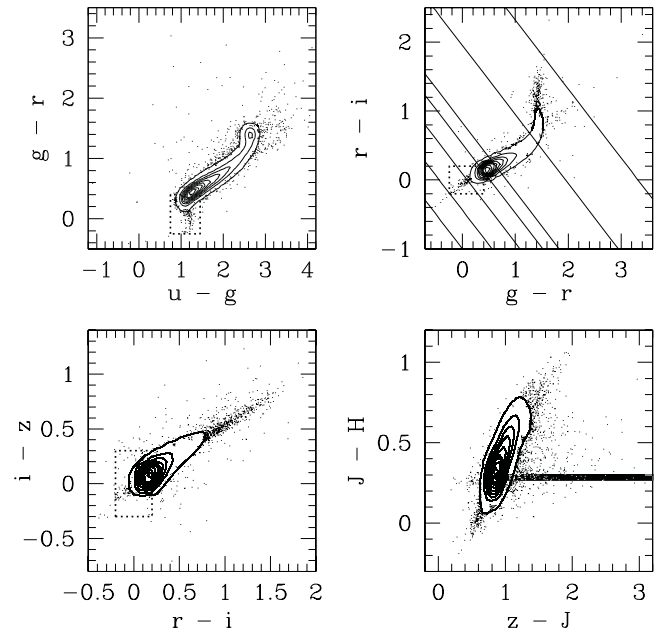
Figure 16. SDSS color–color diagrams of the variable candidates. Boxes represent the ranges of single-epoch colors for RR Lyrae variable candidates from Sesar et al. (2010). Solid lines in the panel of $(g - r)$ and $(r - i)$ colors represent $(g - i)$ colors corresponding to spectral types O5, A0, F0, G0, K0, M0, and M5 from left to right, which are derived from synthesized stellar spectra (Covey et al. 2007). The Galactic extinction is not included here.

photometric data, we use point-spread function magnitudes. We also find SDSS objects with limiting magnitudes $m_u = 22.3$, $m_g = 23.3$, $m_r = 23.1$, $m_i = 22.3$, and $m_z = 20.8$.

_____

[13] This period is found by using the tool provided in http://www.astro.lsa.umich.edu/ msshin/science/code/MultiStep_Period/ (Shin & Byun 2004).
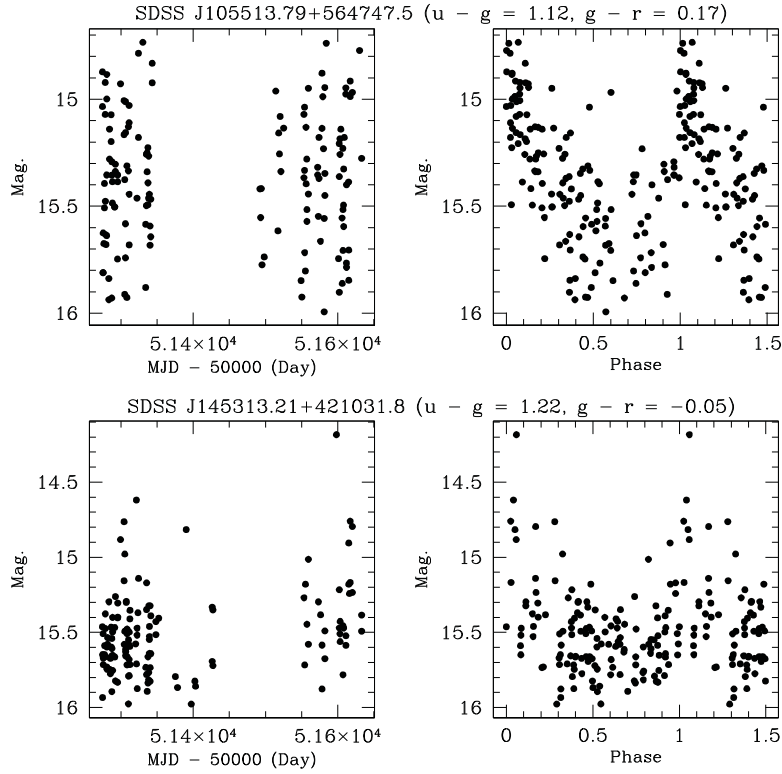
Figure 17. Example light curves of the NSVS objects selected as RR Lyrae variable candidates with the SDSS spectroscopic data. The left column shows the raw NSVS light curves, while the right column presents the light curves folded with approximate periods of 0.541757 (top) and 0.489448 (bottom) days, respectively.

The SDSS color–color diagram can be used to pick out probable RR Lyrae variables that are pulsating horizontal branch stars (Gautschy & Saio 1996). As suggested in the theoretical prediction of colors for RR Lyrae stars in the SDSS photometric system (Marconi et al. 2006), the following color ranges can be used to find RR Lyrae candidates (Sesar et al. 2010):

$$0.75 < u - g < 1.45, \qquad (6)$$

$$-0.25 < g - r < 0.4, \qquad (7)$$

$$-0.2 < r - i < 0.2, \qquad (8)$$

$$-0.3 < i - z < 0.3, \qquad (9)$$

which are shown as boxes in Figure 16. A large number of variable candidates are identified as F-, G-, and K-type stars in the figure.

Among the variable candidates selected by the RR Lyrae color cuts, we present example light curves of two objects, SDSS J105513.79+564747.5 (NSVS 2594623) and SDSS J145313.21+421031.8 (NSVS 5152328), which have also been observed in the SDSS spectroscopy, in Figure 17. Neither object is classified as a variable source in the SIMBAD database. However, SDSS J105513.79+564747.5 is identified as a variable in *GALEX* observations (Welsh et al. 2005) as included in the VSX catalog (see the Appendix). We can estimate approximate periods of these two variables with the NSVS light curves as 0.541757 and 0.489448 days, respectively. These examples clearly show that a low cadence in the NSVS data is not high enough to derive complete light curves of these RR Lyrae variables, which have short periods (Sterken & Jaschek 2005), except for a few NSVS objects with enough data

(Kinemuchi et al. 2006). Therefore, further follow-up observations of interesting NSVS objects will be required to confirm their variability classes. SDSS J105513.79+564747.5 was identified as a probable RR Lyrae by Wheatley et al. (2008), although they could not retrieve a complete light curve from their *GALEX* observations. SDSS J145313.21+421031.8 was also recognized as a blue horizontal branch star in the SDSS observation (Sirko et al. 2004). These examples prove that using colors of objects is complementary to variability analysis to identify object types.

Six other examples of variable candidates are presented in Figure 18 where NSVS objects have corresponding reliable SDSS and 2MASS photometric data. Except for SDSS J021532.23−104029.3, these objects have $(g - i)$ colors without the SDSS bad photometric flags. Spectral types of stars can be described approximately by $(g - i)$ colors where B0, A0, F0, G0, K0, and M0 correspond to $(g - i) \sim -0.94$, -0.44, 0.09, 0.52, 0.83, and 1.95, respectively (Covey et al. 2007). Most variable candidates are close to G5 as shown in Figure 16. Although $(g - i)$ is generally a good proxy of spectral types, SDSS J155325.80+530924.2 (NSVS 5206326) was already spectroscopically confirmed as an M8III star for its $(g - i) = 5.81$. We also note that this object is also included in the VSX catalog as a variable star (see the Appendix). These examples reassert that the low cadence in the NSVS data does not guarantee a certain classification of variable objects, requiring further follow-up observations with different cadences.

5.5. SDSS and GALEX Photometry

Hot stellar objects such as white dwarfs and massive main-sequence stars are generally not detected in NIR, but they can be more easily recognized over UV wavelength ranges, which
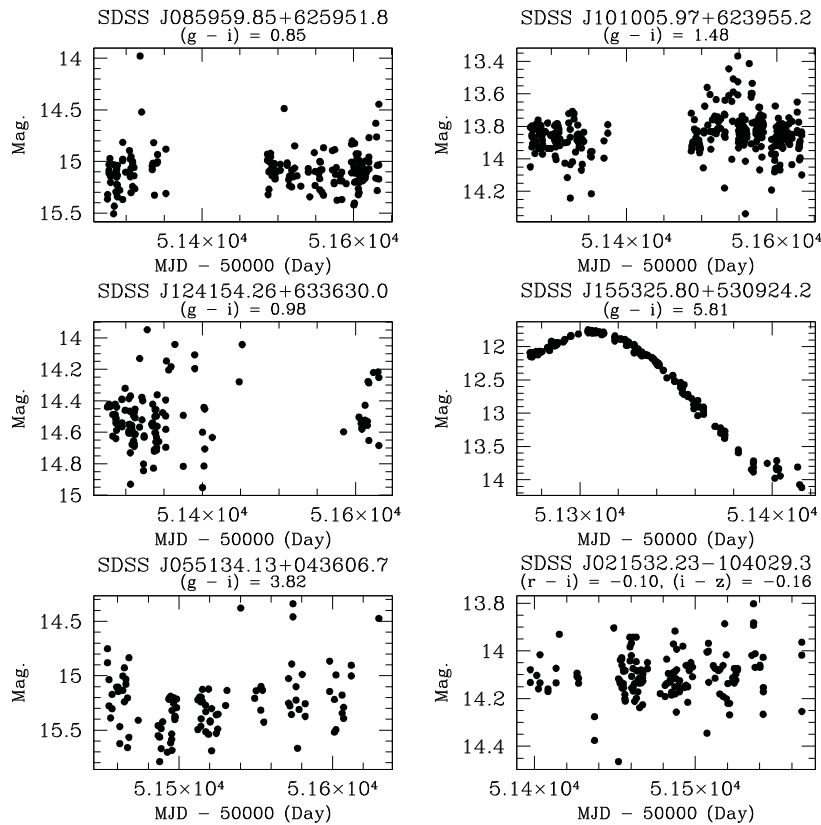
13

Figure 18. Example light curves of variable candidates with reliable SDSS and 2MASS photometric measurements. The top two objects are included in all color–color diagrams shown in Figure 16. SDSS J155325.80+530924.2 is not classified as a known variable star in the SIMBAD database. However, it is a spectroscopically confirmed giant star with another designation, 2MASS J15532581+5309244 (Cruz et al. 2003), and is found to be a variable star in Woźniak et al. (2004b). SDSS J021532.23−104029.3 is identified as a field horizontal branch star BPS CS 22175-0003 by Wilhelm et al. (1999).

cover most stellar flux. In the *GALEX* GR4/5,[14] we find objects matching our NSVS variable candidates' coordinates within 6″. When multiple *GALEX* objects are matched to a single NSVS object, we choose the nearest *GALEX* object as the best match. A total of 739,625 variable candidates, i.e., about 40% of the candidates, correspond to *GALEX* photometric objects. A total of 286,185 candidates have corresponding SDSS objects too.

In Figure 19, we present colors of variable candidates with reliable *GALEX* and SDSS photometric measurements. In addition to following the same conditions for the reliable SDSS photometric measurements as in Section 5.4, we require that the *GALEX* objects should have a distance from the center of the *GALEX* field of view <0.°6, and both FUV and NUV magnitudes <25 (Agüeros et al. 2005; Maxted et al. 2009). The FWHM angular resolution is about 6″ in the NUV channel (Morrissey et al. 2005). Considering the combined effects of poor spatial resolution in the NSVS and *GALEX* data, the spatial association among the NSVS, SDSS, and *GALEX* objects needs a careful check when people select interesting NSVS objects with corresponding SDSS and *GALEX* objects together in our database.

Colors of most variable candidates presented in Figure 19 are consistent with the expected colors of normal stars (Seibert et al. 2005; Bianchi et al. 2007, 2009), considering the Galactic extinction that makes overall colors red. The color distribution of known quasars (Trammell et al. 2007) is well separated from that of stars in the diagram of (FUV − g) and (g − i). Because our variable candidates are bright objects, all of them might not be

quasars but stars even though some objects seem to have quasar-like colors. Hot white dwarf candidates can be selected with the color cut of (FUV − NUV) < 0 and (g − r) < −0.2 (Agüeros et al. 2005). However, we must keep in mind that the multiple matches of the NSVS coordinates to both SDSS and *GALEX* catalogs have worse precision than a single match to either the SDSS or *GALEX* catalog. Therefore, the color combining both SDSS and *GALEX* photometric data might not be reliable when the objects are faint or close to neighboring objects in the SDSS and *GALEX* catalogs. Since the precision of the SDSS object coordinates is much better than that of the *GALEX* catalogs, the SDSS colors are more reliable than the *GALEX* colors in the color–color diagram combining both.

Among the NSVS objects with corresponding SDSS and *GALEX* objects, several SIMBAD objects are found with further information about their properties. For example, NSVS 7609761 corresponds to SDSS J115800.38+295731.4 and GALEX J115800.4+295731 with (FUV − NUV) = 1.63 and (g − i) = −0.55. This object is included in Brown et al. (2008) as CHSS 835, which is a star with a spectral type of B8. In the diagram of (FUV − NUV) and (NUV − r), NSVS 4819428 is known as a spectral type B subdwarf, FBS 0839+399 (Wegner & McMahan 1985; Mickaelian 2008), with (FUV − NUV) = −0.35 and (NUV − r) = −1.28. The light curves of these two objects given in Figure 20 do not show distinctive features owing to the poor sampling rate in the NSVS data. Flare-like variation can be presumed from the light curve of NSVS 2744942, which is recognized as an active M dwarf, corresponding to an X-ray object RX J1447.2+5701 (Mochnacki
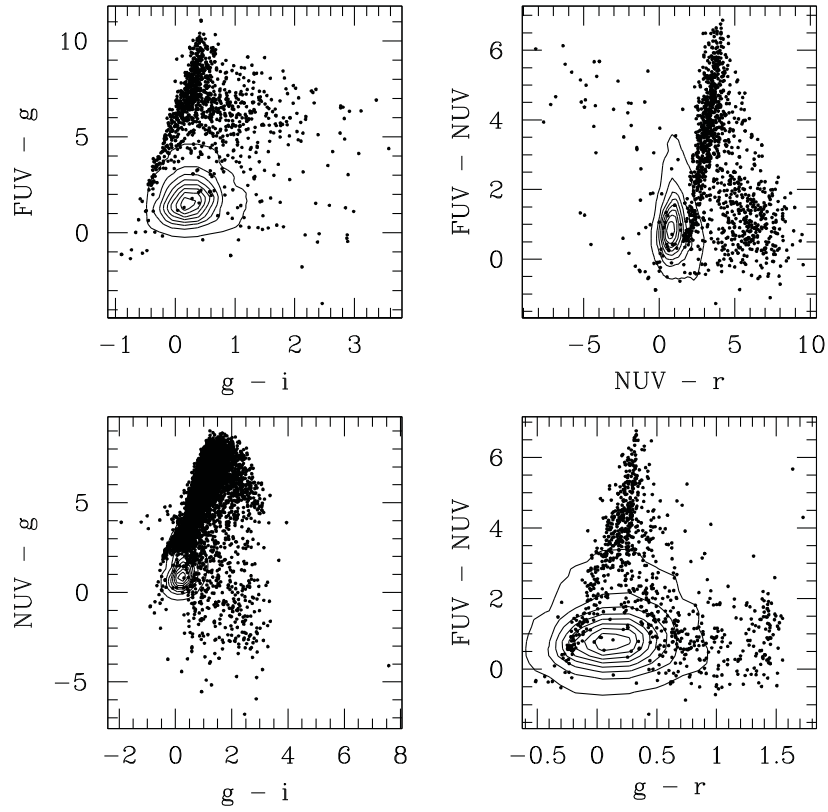
---

[14] http://galex.stsci.edu

14

Figure 19. Color–color diagrams of variable candidates with the SDSS and *GALEX* photometric data. The plots show variable candidates matching the SDSS objects within 1″. Contours correspond to the color distributions of quasars that are detected in both SDSS and *GALEX* (Trammell et al. 2007).
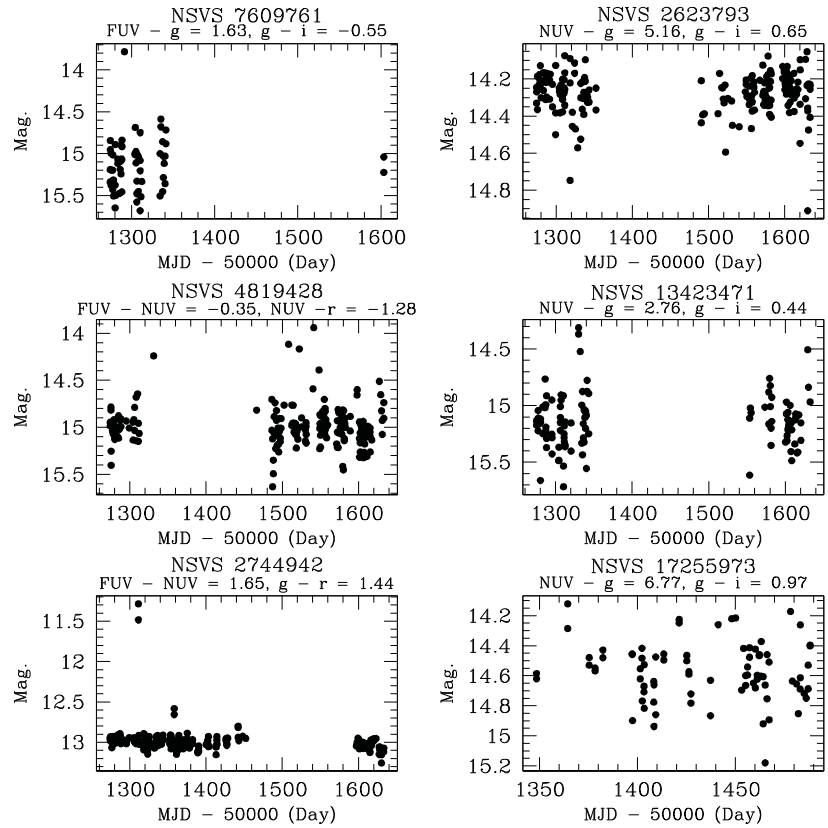

Figure 20. Example light curves of variable candidates with reliable *GALEX* and SDSS photometry. Three objects (left) have further identification in SIMBAD, while the other three objects (right) are selected in order of increasing (NUV − g) color from top to bottom.

et al. 2002). The NSVS light curve of this object shows about 1.5 mag variation even with the poor sampling rate.

Figure 20 also shows the light curves of the three variable candidates that have reliable data for the SDSS g- and i-band photometry as well as the *GALEX* NUV measurement. None of them have different identification in the SIMBAD database. The light variation seems real, but the low cadence in the NSVS data does not produce complete light curves with distinctive types.

6. SUMMARY AND DISCUSSION

A new systematic investigation of variable candidates in the NSVS data was presented with a clustering method for time-series data. Assuming that the dominant fraction of light curves represents non-variable objects, our method finds clusters of light curves with their eight-dimensional features and then finds how many light curves are included in each cluster. When choosing as variable candidates objects that are not included in the top four large clusters and that have $D_M > 99\%$ cut from the largest cluster, the total number of new variable candidate light curves is 1,824,123 in our entire sample of the NSVS data.

The cross-correlation with *IRAS*, *AKARI*, 2MASS, SDSS, and *GALEX* catalogs helps us identify interesting objects with specific spectral types or variability classes (Eyer & Mowlavi 2008). In particular, variable stars over the instability strip can be easily selected from their specific colors (e.g., see Figure 17). We also show examples of long-period variables that can be selected from their *IRAS*, *AKARI*, or 2MASS colors (e.g., see Figures 11, 13, and 15).

Our analysis is presented online with the information on cross-correlations with other catalogs. Because the sampling pattern in the NSVS data is not good enough to identify detailed structures of different variability types, follow-up observations of variable candidates will be necessary to understand these variable sources by including more photometric data points and improving the sampling rates. Moreover, some variable candidates such as *IRAS* sources might be new maser sources, which are interesting objects in the radio region.

Our approach of detecting variable candidates in the NSVS data is supplementary to previous methods of finding variable candidates. We do not claim that this method is the best way in all cases. If observation systems including instruments, environments, and data reduction procedures are previously well known or are well controlled, supervised methods can be superior to unsupervised methods like our approach because supervised methods can simulate observing systems with known variable and non-variable sources to find the best separation between variable and non-variable sources. This separation can be applied to detect new variable candidates in the test data. Therefore, when the data properties, including all kinds of systematic patterns and real variability patterns, are understood and modeled well, detecting variable sources becomes a *classification* problem instead of a *clustering* problem.

Our analysis results can be used with many different methods of selecting variable candidates. In this paper, our conservative selection method is avoiding the top four large clusters and objects with $D_M < 99\%$. However, when people are interested in infrared variable sources, they can choose as variable candidates objects corresponding to *IRAS* objects with $D_M > 90\%$. If known variable objects can produce clusters with reasonable sizes, finding clusters with many known variable objects can be an efficient way to find variable candidates. Unfortunately, this approach is not feasible now because the number of known variable objects is too small in each NSVS observation field.

Several variability surveys cover the same apparent magnitude ranges as the NSVS does. A large fraction of sky has already been observed in the All Sky Automated Survey (ASAS; Pojmanski 1997) and SuperWASP (Street et al. 2003). Our approach of variability detection can be applied to those data sets too. In addition, objects included in both the NSVS and others can be combined to extend the span of time-series data or to complement different sampling patterns. We plan to update the online database with these additional data sets in the future. Furthermore, because almost all objects included in our analysis will be monitored by the *Gaia* mission for five years (Cacciari 2009), our analysis will be combined with the future time-series data and astrometric/kinematic information.

Our method can also be improved to catch much broader types of variable objects and objects with weak variability signals. For this purpose, it is important to include various features of light curves as we emphasized in Section 2. In particular, the usage of AoVM as one feature of light curves is strongly limited because a complete form of a periodogram has more information for light curves. Therefore, it must be useful to develop new features describing periodograms more completely if the new features can be estimated in computationally cheap ways. Moreover, Stetson's *I* index (Stetson 1996) can be included in our method if data include multi-band light curves. For instance, *J* and *K* can be derived in each band, while *I* is estimated with multi-band light curves. Our usage of *J* can also be changed to use all pairs of data points instead of using sequential pairs.

APPENDIX

KNOWN OR SUSPECTED VARIABLES FOUND IN THE VSX

The SIMBAD database is commonly used to identify known objects. We also use the database to find all known variable objects and extragalactic objects as shown in Section 5.1. However, the SIMBAD database is not as complete as other catalogs of variable stars. In particular, the AAVSO International Variable Star Index (VSX; Watson 2006) catalog is frequently updated with new reports of variable stars.

We check how many variable candidates selected by our conservative selection are not classified as variable sources in the SIMBAD database but are included in the VSX catalog. Here, we use the catalog released online on 2009 November 15, including 178,599 stars. Among 1,824,123 variable candidates chosen by our conservative selection and not included in the SIMBAD database, we find that 41,019 objects are included as variable stars in the VSX catalog. These known variable stars in the catalog are largely included with references to the ASAS observations (9999 objects; Pojmanski 1997) and to the NSVS observations (15,958 objects), which we also use here. A total of 1147 objects are matched to suspected variable stars in the catalog. Interestingly, a total of 1106 suspected variables are included with references to the NSVS light curves. Therefore, at least 1,783,104 variable candidates are newly selected in our method. In the online database, we provide links to the most recent VSX catalog for the NSVS light curves we examine.

REFERENCES

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Agüeros, M. A., Ivezić, Ž., Covey, K. R., et al. 2005, AJ, 130, 1022
Alard, C., & Lupton, R. H. 1998, ApJ, 503, 325
Becker, A. C., Wittman, D. M., Boeshaar, P. C., et al. 2004, ApJ, 611, 418
Bernhard, K., & Lloyd, C. 2000, IBVS, 4920, 1
Bianchi, L., & The GALEX Team 1999, Mem. Soc. Astron. Ital., 70, 365
Bianchi, L., Hutchings, J. B., Efremova, B., et al. 2009, AJ, 137, 3761
Bianchi, L., Rodriguez-Merino, L., Viton, M., et al. 2007, ApJS, 173, 659
Blommaert, J. A. D. L., van der Veen, W. E. C. J., & Habing, H. J. 1993, A&A, 267, 39
Bono, G., Trevese, D., & Turatto, M. 2003, Mem. Soc. Astron. Ital., 74, 1004
Brown, W. R., Beers, T. C., Wilhelm, R., et al. 2008, AJ, 135, 564
Cacciari, C. 2009, Mem. Soc. Astron. Ital., 80, 97
Chen, T., Morris, J., & Martin, E. 2006, J. Roy. Stat. Soc. Ser. C, 55, 699
Chiu, K., Richards, G. T., Hewett, P. C., & Maddox, N. 2007, MNRAS, 375, 1180
Clegg, P. E. 1980, Phys. Scr., 21, 678
Cole, A. A., & Weinberg, M. D. 2002, ApJ, 574, L43
Corwin, T. M., Sumerel, A. N., Pritzl, B. J., et al. 2006, AJ, 132, 1014
Covey, K. R., Ivezić, Ž., Schlegel, D., et al. 2007, AJ, 134, 2398
Coyne, G. V., & MacConnell, D. J. 1983, Vatican Obs. Publ., 2, 73
Cruz, K. L., Reid, I. N., Liebert, J., Kirkpatrick, J. D., & Lowrance, P. J. 2003, AJ, 126, 2421
Davies, D. L., & Bouldin, D. W. 1979, IEEE Trans. Pattern Anal. Mach. Intell., 1, 224
Dimitrov, D. 2009, Bulg. Astron. J., 12, 49
Djorgovski, S. G., Mahabal, A. A., Brunner, R. J., et al. 2001, in ASP Conf. Ser. 225, Virtual Observatories of the Future, ed. R. J. Brunner, S. George Djorgovski, & A. S. Szalay (San Francisco, CA: ASP), 52
Dommanget, J., & Nys, O. 1994, Catalog of the Components of Double and Multiple Stars, Com. de l'Observ. Royal de Belgique, Vol. 115, 1
Eyer, L. 2006, in ASP Conf. Ser. 349, Astrophysics of Variable Stars, ed. C. Sterken & C. Aerts (San Francisco, CA: ASP), 15
Eyer, L., & Mowlavi, N. 2008, J. Phys. Conf. Ser., 118, 012010
Fan, X. 1999, AJ, 117, 2528
Finlator, K., Ivezić, Ž, Fan, X., et al. 2000, AJ, 120, 2615
Fukugita, M., Yasuda, N., Doi, M., Gunn, J. E., & York, D. G. 2011, AJ, 141, 47
Gautschy, A., & Saio, H. 1996, ARA&A, 34, 551
Gössl, C. A., & Riffeser, A. 2002, A&A, 381, 1095
Hartman, J. D., Bakos, G., Stanek, K. Z., & Noyes, R. W. 2004, AJ, 128, 1761
Helou, G., & Walker, D. W. 1988, Infrared Astronomical Satellite (IRAS) Catalogs and Atlases, Vol. 7 (Washington, DC: NASA)
Hoffman, D. I., Harrison, T. E., Coughlin, J. L., et al. 2008, AJ, 136, 1067
Hoffman, D. I., Harrison, T. E., & McNamara, B. J. 2009, AJ, 138, 466
Ishihara, D., Onaka, T., Kataza, H., et al. 2010a, A&A, 514, A1
Ishihara, D., Onaka, T., Kataza, H., et al. 2010b, VizieR Online Data Catalog, 2297, 0
Ita, Y., Matsuura, M., Ishihara, D., et al. 2010, A&A, 514, A2
Ita, Y., Tanab, T., Matsunaga, N., et al. 2004, MNRAS, 353, 705
Jackson, T., Ivezić, Ž., & Knapp, G. R. 2002, MNRAS, 337, 749
Joanes, D. N., & Gill, C. A. 1998, J. R. Stat. Soc. (Series D): The Statistician, 47, 183
Jolion, J. M., Meer, P., & Bataouche, S. 1991, IEEE Trans. Pattern Anal. Mach. Intell., 13, 791
Kaiser, N., Burgett, W., Chambers, K., et al. 2010, Proc. SPIE, 7733, 77330E-1
Kelley, N., & Shaw, J. S. S. 2007, J. Southeastern Assoc. Res. Astron., 1, 13
Kinemuchi, K., Smith, H. A., Woźniak, P. R., & McKay, T. A. 2006, AJ, 132, 1202
Kinnunen, T., & Skiff, B. A. 2000, IBVS, 4865, 1
Kiss, L. L., & Bedding, T. R. 2003, MNRAS, 343, L79
Kiss, L. L., Derekas, A., Szabó, G. M., Bedding, T. R., & Szabados, L. 2007, MNRAS, 375, 1338
Kleinmann, S. G. 1992, in ASP Conf. Ser. 34, Robotic Telescopes in the 1990s, ed. A. V. Filippenko (San Francisco, CA: ASP), 203
Kouzuma, S., & Yamaoka, H. 2009, AJ, 138, 1508
Kouzuma, S., & Yamaoka, H. 2010, A&A, 509, A64
Kwok, S., Volk, K., & Bidelman, W. P. 1997, ApJS, 112, 557
Liao, T. W. 2005, Pattern Recognit., 38, 1857
Marconi, M., Cignoni, M., Di Criscienzo, M., et al. 2006, MNRAS, 371, 1503
Mauron, N., Gigoyan, K. S., & Kendall, T. R. 2007, A&A, 475, 843
Maxted, P. F. L., Gänsicke, B. T., Burleigh, M. R., et al. 2009, MNRAS, 400, 2012
Meyer, M. R., Calvet, N., & Hillenbrand, L. A. 1997, AJ, 114, 288
Mickaelian, A. M. 2008, AJ, 136, 946
Mochnacki, S. W., Gladders, M. D., Thomson, J. R., et al. 2002, AJ, 124, 2868
Morrissey, P., Conrow, T., Barlow, T. A., et al. 2007, ApJS, 173, 682
Morrissey, P., Schiminovich, D., Barlow, T. A., et al. 2005, ApJ, 619, L7
Moshir, M., et al. 1990, IRAS Faint Source Catalog, Version 2.0
Murakami, H., Baba, H., Barthel, P., et al. 2007, PASJ, 59, 369
Nicholson, M., Sutherland, J., & Sutherland, C. 2005, Open Eur. J. Var. Stars, 12, 1
Nikolaev, S., & Weinberg, M. D. 2000, ApJ, 542, 804
Olivier, E. A., Whitelock, P., & Marang, F. 2001, MNRAS, 326, 490
Oyabu, S., Matsuhara, H., Wada, T., et al. 2010, Proc. SPIE, 7731, 77312P-1
Ozawa, R., Grosso, N., & Montmerle, T. 2005, A&A, 429, 963
Paczyński, B. 2000, PASP, 112, 1281
Pojmanski, G. 1997, Acta Astron., 47, 467
Pojmanski, G. 2002, Acta Astron., 52, 397
Pojmanski, G., & Maciejewski, G. 2005, Acta Astron., 55, 97
Protopapas, P., Jimenez, R., & Alcock, C. 2005, MNRAS, 362, 460
Ramos-Larios, G., Guerrero, M. A., Suárez, O., Miranda, L. F., & Gómez, J. F. 2009, A&A, 501, 1207
Renner, S., Rauer, H., Erikson, A., et al. 2008, A&A, 492, 617
Robert, C. P. 1996, in Mixtures of Distributions: Inference and Estimation, Markov Chain Monte Carlo in Practice, ed. W. Gilks & D. J. elhalter (London: Chapman & Hall), 441

Schmidt, E. G., Hemen, B., Rogalla, D., & Thacker-Lynn, L. 2009, AJ, 137, 4598

Schwarzenberg-Czerny, A. 1996, ApJ, 460, L107

Seibert, M., Budavri, T., Rhee, J., et al. 2005, ApJ, 619, L23

Sesar, B., Ivezić, Ž, Grammer, S. H., et al. 2010, ApJ, 708, 717

Sevenster, M. N. 2002, AJ, 123, 2788

Sharma, S., & Johnston, K. V. 2009, ApJ, 703, 1061

Shin, M.-S., & Byun, Y.-I. 2004, J. Korean Astron. Soc., 37, 79

Shin, M.-S., & Byun, Y.-I. 2007, in ASP Conf. Ser. 362, The Seventh Pacific Rim Conference on Stellar Astrophysics, ed. Y. W. Kang, H.-W. Lee, K.-S. Cheng, & K.-C. Leung (San Francisco, CA: ASP), 255

Shin, M.-S., Sekora, M., & Byun, Y.-I. 2009, MNRAS, 400, 1897 (Paper I)

Siódmiak, N., Meixner, M., Ueta, T., et al. 2008, ApJ, 677, 382

Sirko, E., Goodman, J., Knapp, G. R., et al. 2004, AJ, 127, 899

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163

Stephenson, C. B. 1986, ApJ, 300, 779

Sterken, C., & Jaschek, C. (ed.) 2005, Light Curves of Variable Stars (Cambridge: Cambridge Univ. Press)

Stetson, P. B. 1996, PASP, 108, 851

Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, AJ, 123, 485

Street, R. A., Pollaco, D. L., Fitzsimmons, A., et al. 2003, in ASP Conf. Ser. 294, Scientific Frontiers in Research on Extrasolar Planets, ed. D. Deming & S. Seager (San Francisco, CA: ASP), 405

Szczerba, R., Siódmiak, N., Stasińska, G., & Borkowski, J. 2007, A&A, 469, 799

Trammell, G. B., Vanden Berk, D. E., Schneider, D. P., et al. 2007, AJ, 133, 1780

Tsujimoto, M., Koyama, K., Tsuboi, Y., Goto, M., & Kobayashi, N. 2002, ApJ, 566, 974

Tyson, J. A. 2002, Proc. SPIE, 4836, 10

Usatov, M. 2008, Open Eur. J. Var. Stars, 81, 1

Usatov, M., & Nosulchik, A. 2008, Open Eur. J. Var. Stars, 87, 1

van der Veen, W. E. C. J., & Habing, H. J. 1988, A&A, 194, 125

Vendramin, L., & Campello, R. J. G. B. 2010, Stat. Anal. Data Min., 3, 209

Véron-Cetty, M.-P., & Véron, P. 2006, A&A, 455, 773

Ververidis, D., & Kotropoulos, C. 2008, IEEE Trans. Signal Process., 56, 2797

Vieira, S. L. A., Corradi, W. J. B., Alencar, S. H. P., et al. 2003, AJ, 126, 2971

von Neumann, J. 1941, Ann. Math. Stat., 12, 367

Watson, C. L. 2006, The International Variable Star Index (VSX), Society for Astronomical Sciences Annual Symposium, Vol. 25 (Society for Astronomical Sciences), 47

Wegner, G., & McMahan, R. K. 1985, AJ, 90, 1511

Welsh, B. Y., Wheatley, J. M., Heafield, K., et al. 2005, AJ, 130, 825

Wheatley, J. M., Welsh, B. Y., & Browne, S. E. 2008, AJ, 136, 259

Wilhelm, R., Beers, T. C., Sommer-Larsen, J., et al. 1999, AJ, 117, 2329

Williams, J. D. 1941, Ann. Math. Stat., 12, 239

Williams, R., Hanisch, R., Szalay, A., & Plante, R. 2008, Simple Cone Search Specification Version 1.03, http://www.ivoa.net/Documents/REC/DAL/ConeSearch-20080222.html

Wils, P., Lloyd, C., & Bernhard, K. 2006, MNRAS, 368, 1757

Woźniak, P. R. 2000, Acta Astron., 50, 421

Woźniak, P. R., Vestrand, W. T., Akerlof, C. W., et al. 2004a, AJ, 127, 2436

Woźniak, P. R., Williams, S. J., Vestrand, W. T., & Gupta, V. 2004b, AJ, 128, 2965

Yamamura, I., Makiuti, S., Ikeda, N., et al. 2010, VizieR Online Data Catalog, 2298, 0

Yu, J., Amores, J., Sebe, N., & Tian, Q. 2006, in Proc. ICME 2006, A New Study on Distance Metrics as Similarity Measurement, ed. J. Yu, J. Amores, N. Sebe, & Q. Tian, 533

Yuan, F., & Akerlof, C. W. 2008, ApJ, 677, 808

Zoccali, M., Renzini, A., Ortolani, S., et al. 2003, A&A, 399, 931

Zuckerman, B. 1987, in Cool Stars, Stellar Systems and the Sun, ed. J. L. Linsky & R. E. Stencel (Lecture Notes in Physics, Vol. 291; Berlin: Springer), 351

18