Detecting variability in astronomical time series data

: applications of clustering methods in cloud computing environments

Min – Su Shin[1], Yong – Ik Byun[2], Seo – Won Chang[2], Dae – Won Kim[2,3], Myung – Jin Kim[2,4]
Dong – Wook Lee[2], Jaegyoon Ham[5], Yong – Hwan Jung[5], Junweon Yoon[5], Jae – Hyuck Kwak[5], Joo Hyun Kim[5]
[1] University of Michigan, [2] Yonsei University, [3] CfA, [4] KASI, [5] KISTI

We present applications of clustering methods to detect variability in massive astronomical time series data. Focusing on variability of bright stars, we use clustering methods to separate possible variable sources from other time series data, which include intrinsically non-variable sources and data with common systematic patterns. We already finished the analysis of the Northern Sky Variability Survey (NSVS) data, which include about 16 million light curves, and present candidate variable sources with their association to other data at different wavelengths. We also apply our clustering method to the light curves of bright objects in the SuperWASP Data Release 1 (DR1). This study is conducted in cloud computing environment provided by the KISTI Supercomputing Center. We explain our experience of using the cloud computing test bed.

1. Introduction

Detecting variability in astronomical time series data can be understood as a problem of outlier or anomaly detection in statistics and machine learning. The main assumption is that the time series data of detectable variable objects is considerably different from the rest of data. In usual data sets of astronomical time series data, there are considerably more normal objects, i.e. non-variable objects, than abnormal objects, i.e. variable objects. When we can describe the data properties of non-variable objects well in given data sets, we can detect variable objects easily and completely.
The common types of anomaly detection methods are graphical and statistical-based, proximity/distance-based, density-based, and clustering-based. Traditionally, astronomers have used statistical-based anomaly detection methods, while generally ignoring any systematic patterns of time-series data or erasing the systematic patterns with empirical approaches. Here, we apply clustering methods, which can consider objects affected by common systematic patterns as ordinary types, and can separate peculiar objects as variable candidates.

2. Methods : clustering with variability indices

Multiple variability indices

In order to detect various patterns of light variation, we adopt the following variability indices which are derived with light curves x_n.

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

Note. — 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. $sign(\delta_n\delta_{n+1})$ is the sign of $\delta_n\delta_{n+1}$.

Infinite Gaussian Mixture Model

Gaussian Mixture Model (GMM) is commonly used as a density estimator and a clustering method by describing the distribution of data as mixtures of multivariate Gaussian distributions. Because we do not know how many clusters exist in our data, we use an Infinite Gaussian Mixture Model with Dirichlet Process which allows the model to have infinitely many mixture components:

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

where m is an index over M, $\mathbf{x}$ is a 8-dim vector of parameters, and γ is the number of parameters (in our case $\gamma = 8$). The final distribution of all objects is given by:

$$p(\mathbf{x}) = \sum_{m=1}^{M} p_m(\mathbf{x}) w_m,$$

where w_m is the fraction of each mixture component.

3. Variable candidates in the NSVS

Clustering results

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.

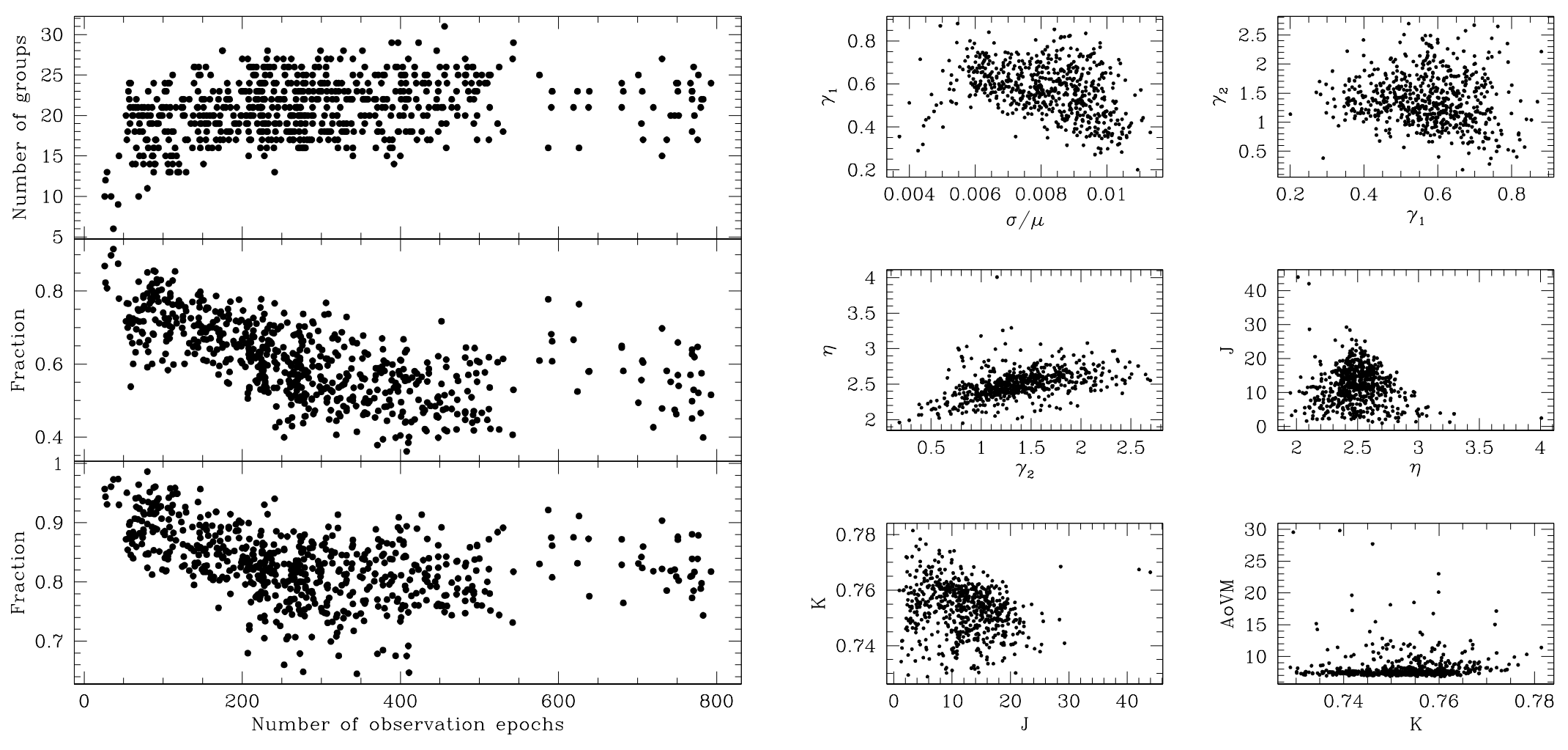


Figure - GMM results with respect to the number of observation epochs *(left)* and center coordinates of the largest cluster *(right)*. The distribution shows that there are field-by-field variations of systematic effects which produce variations of the largest clusters central position in the eight-dimensional space.

Variable candidates

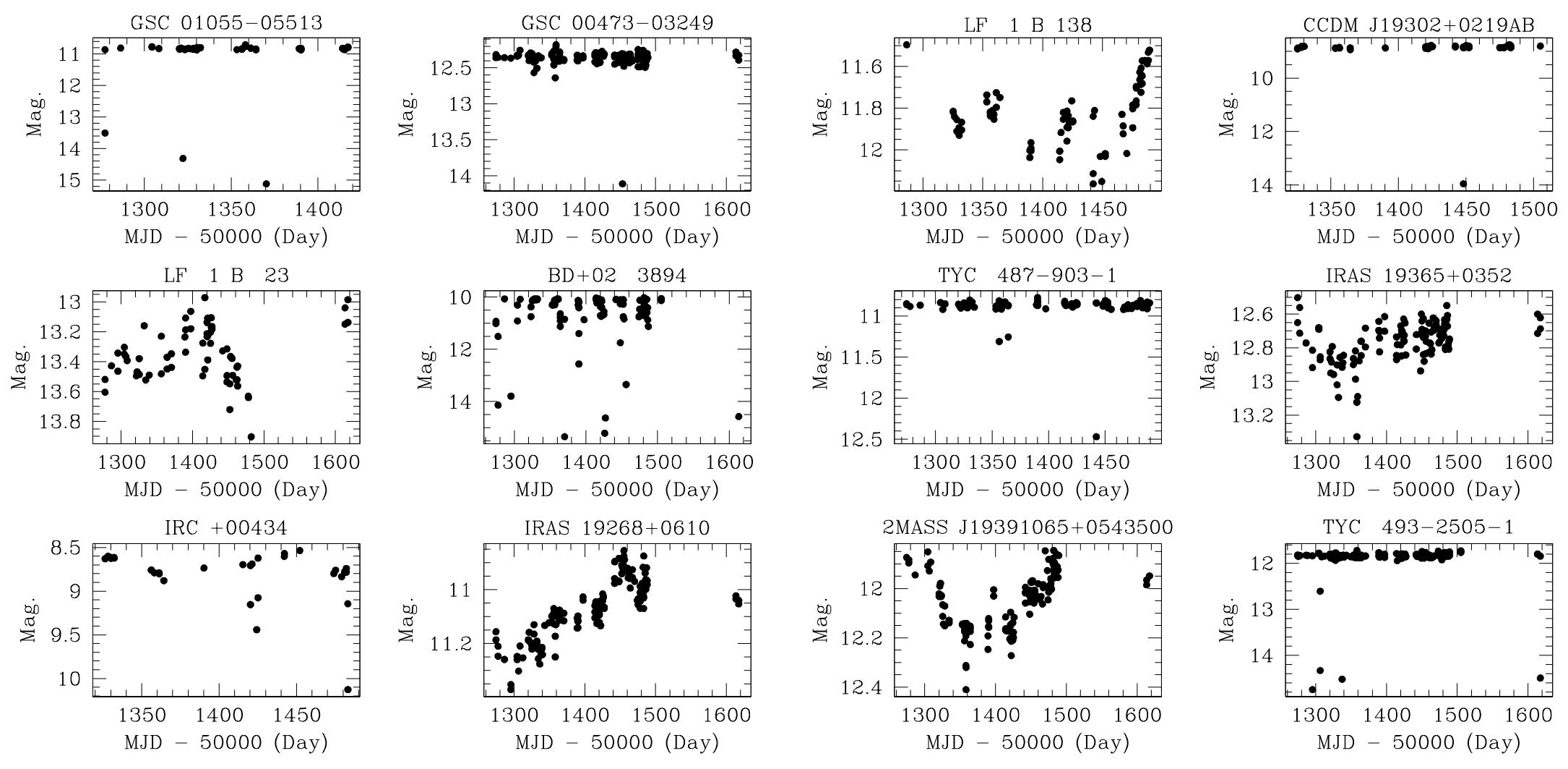


Figure - Example light curves of variable candidates matched to SIMBAD objects of non-variable stars.

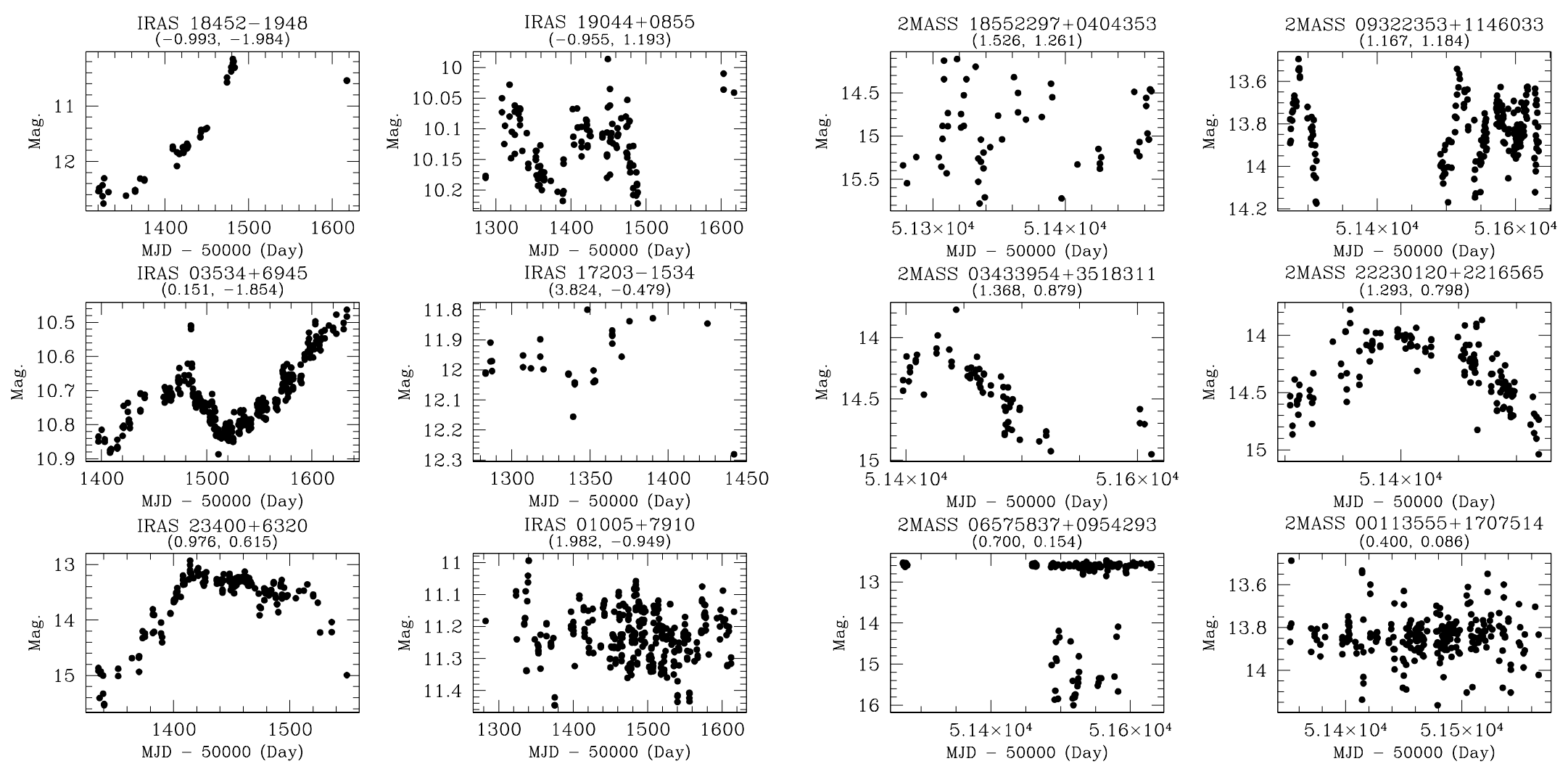


Figure - Example light curves of variable candidates that are IRAS sources *(left)*, and with the reliable 2MASS photometry *(right)*. The IRAS designations and colors ($C_{12/25}$, $C_{25/60}$) are presented in the top of each panel. The 2MASS designations and ($J-H$, $H-K_s$) are given in the top of each panel. 2MASS 18552297+0404353 is also PDS 551 which is a Herbig Ae/Be candidate star.

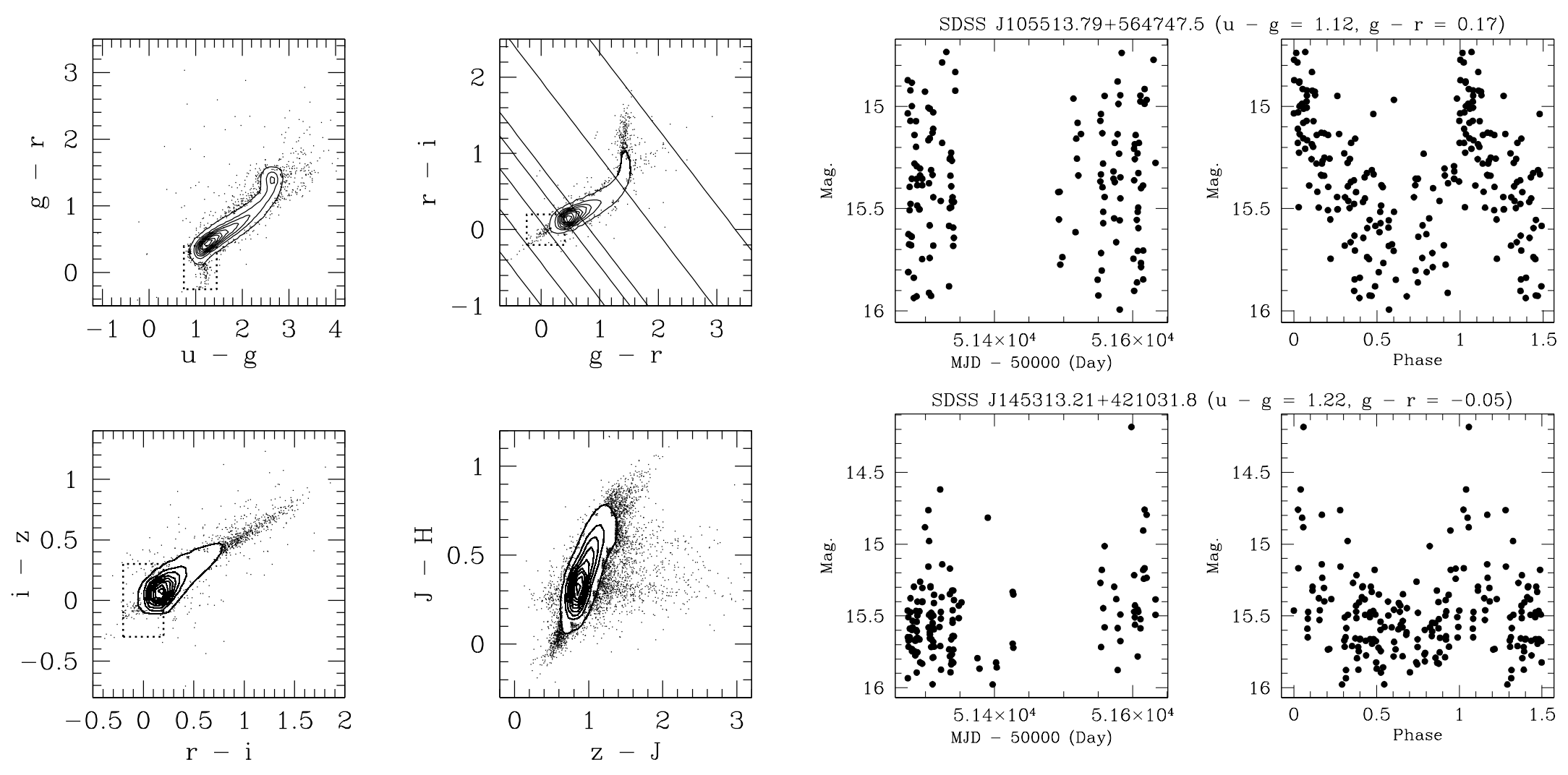


Figure - SDSS color-color diagrams of the variable candidates *(left)*, and example light curves selected as RR Lyrae variable candidates with the SDSS spectroscopic data *(right)*. Boxes represent the ranges of single-epoch colors for RR Lyrae variable candidates. Solid lines in the panel of $(g-r)$ and $(r-i)$ colors represent $(g-i)$ colors corresponding to spectral types O5, A0, F0, G0, K0, M0, and M5 from left to right. The light curves of RR Lyrae variables are folded with approximate periods of 0.541757 (top) and 0.489448 (bottom) days, respectively.

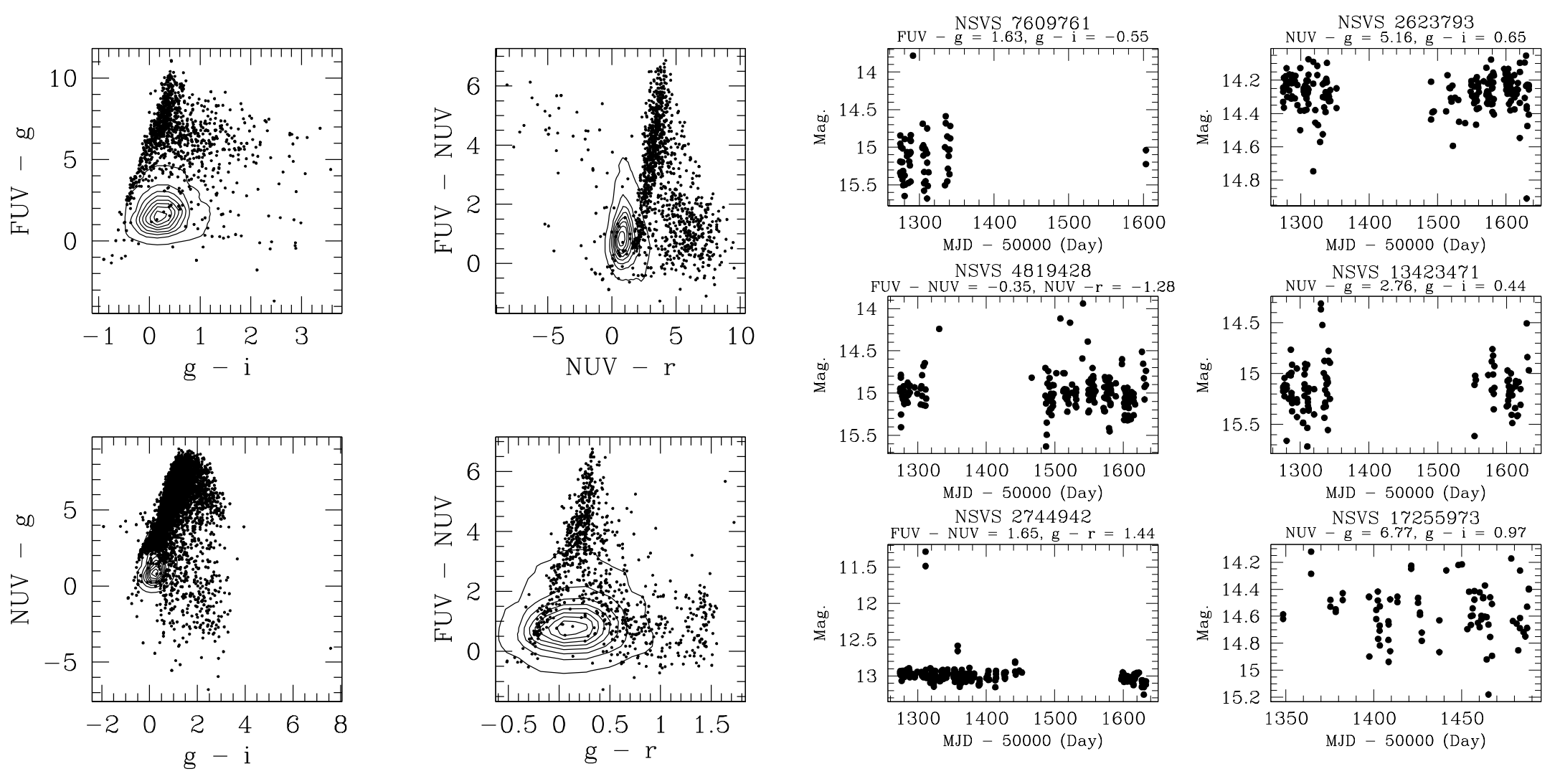


Figure - Color-color diagrams of variable candidates with the SDSS and GALEX photometric data *(left)*, and example light curves of variable candidates with the reliable GALEX and SDSS photometry *(right)*. Contours correspond to the color distributions of quasars which are detected in both SDSS and GALEX.

4. Processing SuperWASP DR1

We also explore the public data release 1 from the SuperWASP project. This data set has about 15 million light curves over both northern and southern skies. Because the SuperWASP DR1 does not have well-defined observation fields, first, we group light curves having similar time-scales such as starting and ending times.

Cloud computing environments

For processing the SuperWASP DR1 data, we try to use a Cloud computing testbed which is deployed by the Korea Institute of Science and Technology Information (KISTI) Supercomputing Center. We test two different system configurations. One system uses Condor as a job management software, and stores data in Lustre distributed file system. Another system adopts Hadoop computing environment with its distributed file system. Both systems are built with virtual machines which are managed by Eucalyptus. Although Hadoop systems does not allow us to use different file systems and is less flexible than the Condor system, its job management considers data locality which can improve performance of parallel distributed processing.

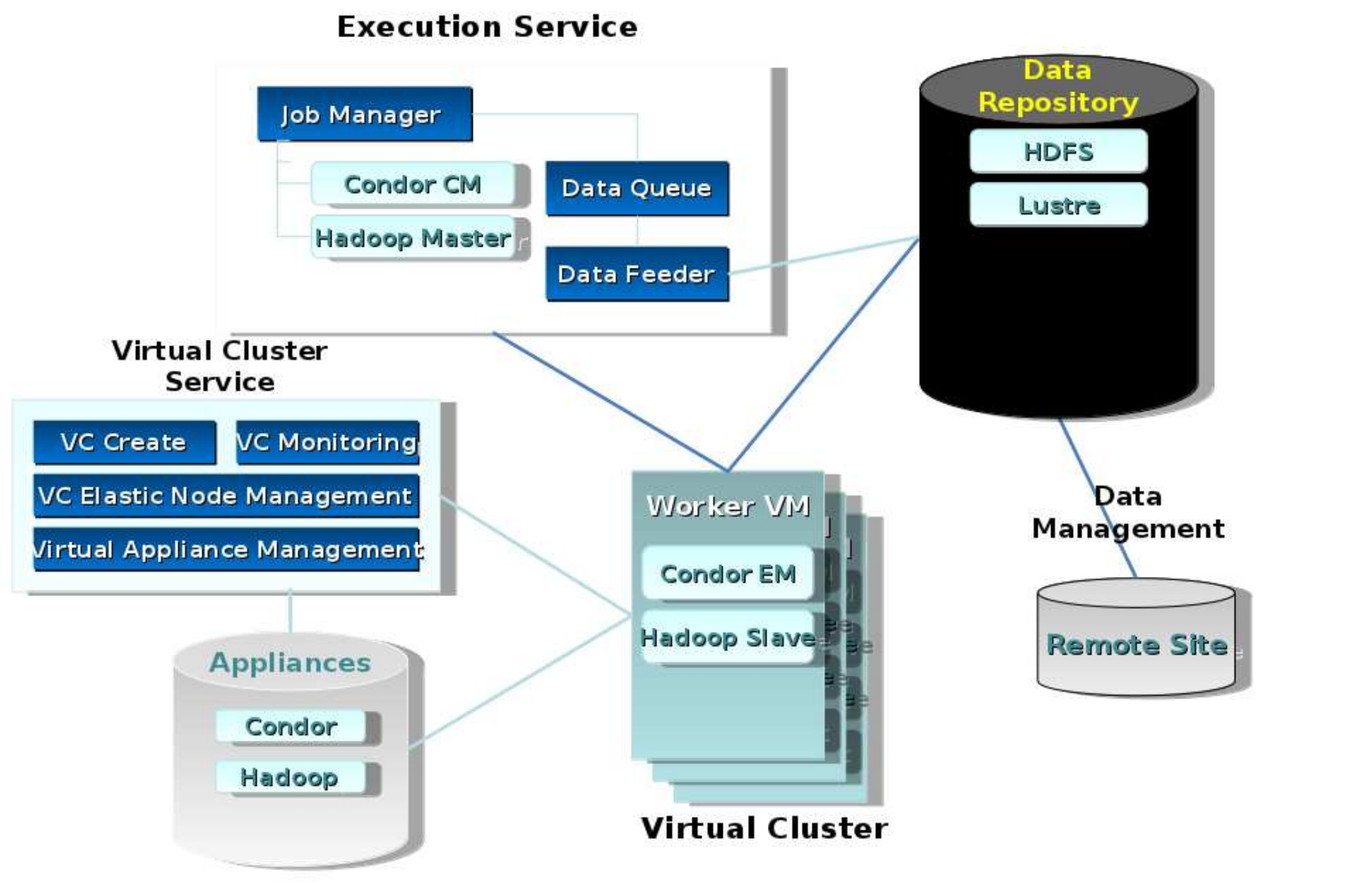


Figure - Cloud computing environments developed by the KISTI for data-intensive computing. In this initial configuration, Condor and Hadoop systems will be equipped with maximum 300 virtual machines, which can be elastically configured and deployed by users. The current test bed is made of 13 and 20 computing virtual machines for the Condor and Hadoop systems, respectively.

The Hadoop system shows a slightly better performance than the Condor system when the input light curve is large due to its usage of data locality. Because the Hadoop system requires a Java program to exploit its general support of the Map-Reduce approach, we use Hadoop streaming to run programs written in C or C++. Therefore, it is more attractive to use the Condor system for the whole procedure of processing the SuperWASP DR1.

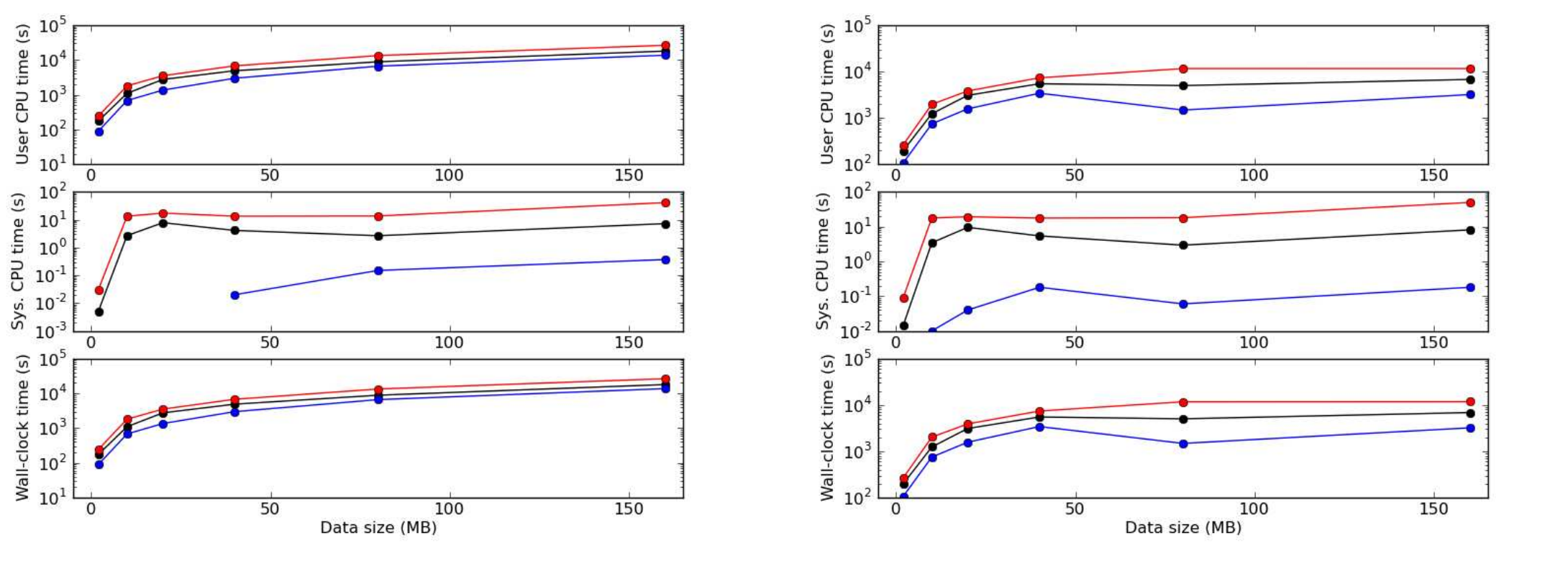
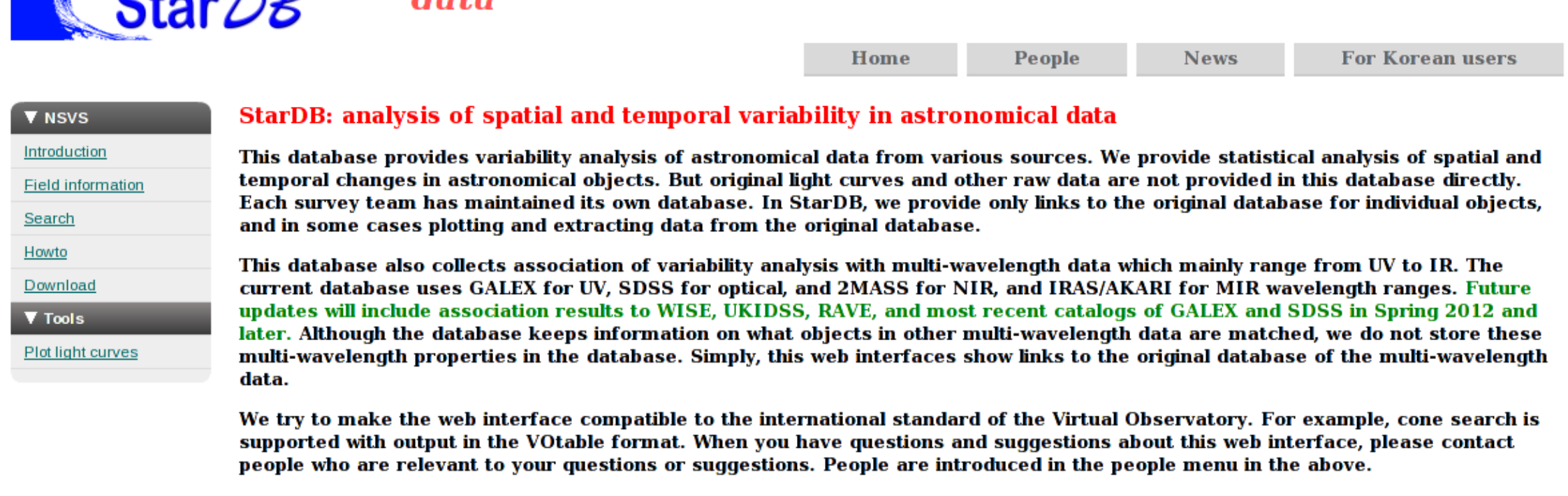


Figure - Performance tests of deriving variability indices from light curves combined to the specific sizes with the Condor *(left)* and Hadoop *(right)* systems. From top to bottom, the plots show user CPU time, system CPU time, and wall-clock time. Red, black, and blue lines correspond to maximum, average, and minimum measurements.

5. Conclusion

We are processing all SuperWASP DR1 light curves to find new variable candidates by using the KISTI Cloud computing environment. The most parts of the processing will be done with the Condor system. The analysis results of the NSVS and SuperWASP DR1 will be released to public on our web site (http://stardb.yonsei.ac.kr) soon.



Analysis of spatial and temporal variability in astronomical data

StarDB: analysis of spatial and temporal variability in astronomical data

This database provides variability analysis of astronomical data from various sources. We provide statistical analysis of spatial and temporal changes in astronomical objects. But original light curves and other raw data are not provided in this database directly. Each survey team has maintained its own database. In StarDB, we provide only links to the original database for individual objects, and in some cases plotting and extracting data from the original database.

This database also collects association of variability analysis with multi-wavelength data which mainly range from UV to IR. The current database uses GALEX for UV, SDSS for optical, and 2MASS for NIR, and IRAS/AKARI for MIR wavelength ranges. Future updates will include association results to WISE, UKIDSS, RAVE, and most recent catalogs of GALEX and SDSS in Spring 2012 and later. Although the database keeps information on what objects in other multi-wavelength data are matched, we do not store these multi-wavelength properties in the database. Simply, this web interfaces show links to the original database of the multi-wavelength data.

We try to make the web interface compatible to the international standard of the Virtual Observatory. For example, cone search is supported with output in the VOtable format. When you have questions and suggestions about this web interface, please contact people who are relevant to your questions or suggestions. People are introduced in the people menu in the above.