Application of the trend filtering algorithm to the MACHO database[*]

J. Szulágyi[1,2], G. Kovács[2], and D. L. Welch[3]

[1] Eötvös University, Department of Astronomy, PO Box 32, 1518 Budapest, Hungary
  e-mail: szulagyi@konkoly.hu
[2] Konkoly Observatory, PO Box 67, 1525 Budapest, Hungary
  e-mail: kovacs@konkoly.hu
[3] Department of Physics and Astronomy, McMaster University, Hamilton, Ontario, L8S 4M1, Canada
  e-mail: welch@physics.mcmaster.ca

ABSTRACT

Aims. Because of the strong effect of systematics (or trends) in variable star observations, we apply the Trend Filtering Algorithm (TFA) to a subset of the MACHO database and search for variable stars.
Methods. TFA has been applied successfully in planetary transit searches, where weak, short-duration periodic dimmings are sought in the presence of noise and various systematics (due to, e.g., imperfect flat-fielding, crowding, etc.). These latter effects introduce colored noise in the photometric time-series that can completely overwhelm the signal. By using a large number of available photometric time-series of a given field, TFA utilizes the fact that the same types of systematics appear in several or many time-series of the same field. As a result, we attempt to reproduce each target time-series by a linear combination of templates, optimized by least-squares. After a signal has been identified in the residuals between the original time-series and the systematics computed by TFA, we reconstruct the signal by employing the full model, including the signal, systematics, and noise.
Results. We apply TFA to the brightest ∼5300 objects from subsets of each of the MACHO Large Magellanic Cloud fields #1 and #79. We find that the Fourier frequency analysis performed on the original data detect some 60% of the objects as trend-dominated. This figure decreases to almost zero after using TFA. In total, we detect 387 variables in the two fields, 183 of which would have remained undetected without using TFA. Where possible, we provide preliminary classification of the variables found.

Key words. methods: data analysis – stars: variables: general – galaxies: Magellanic Clouds

1. Introduction

Microlensing survey projects initiated in the 1990s yielded unprecedented amounts of data for variable star research. Subsequent analyses and follow-up observations focused on certain types of variables and largely ignored the several millions additional lower-amplitude variable stars spanning the HR-diagram, including those not suspected of belonging to existing classes of stellar variability. Since then, a new class of intensive photometric survey has emerged, motivated by the search for transiting extrasolar planets. In this case, because of the unexpected nature of various extrasolar planets, the analysis is not restricted to a subsample of the collected time-series. It quickly became evident from these surveys that the expected faint signals are usually overwhelmed by the much larger systematic effects (or trends) caused by imperfect conditions, hardware and photometric data reduction. Until recently, image and photometric reduction methods were unable to identify most of the variability information in these data. To address this deficiency, methods have been developed such as that of Kruszewski & Semeniuk (2003), SysRem by Tamuz et al. (2005), TFA by Kovács et al. (2005, hereafter KBN), and PDT by Kim et al. (2009) to remove systematics in a post-processing phase. The method of

Kruszewski & Semeniuk (2003) is devoted to reconstructing the signal of known period. This includes a combined fit of a Fourier series (with the signal period) and a polynomial with the hour angle. On the other hand, SysRem and TFA are capable of searching for and reconstructing signals by using the temporal characteristics of the objects observed in the field to identify the part of the target time-series that is produced by the systematics. In PDT a classification of systematics is made to aid a lower parameter model of the time-series.

Attempts have been made to determine the most easily identified systematics and correct for them. For example, the method of External Parameter Decorrelation (EPD, see Bakos et al. 2009) is, for constant stars, based on the correlation between certain image properties (e.g., position, PSF width and elongation) and the brightness deviations from the average value. The method of Differential Image Analysis (DIA, see Alard & Lupton 1998; and for a summary Bramich 2008) has also been used successfully in variable star searches over the past ten years. None of these methods produce photometric time-series that are completely free of systematics. This is because they use only the spatial information in the images and do not include the temporal information, which is available in great abundance from long-term observations.

We briefly describe our method in Sect. 2. In Sect. 3, in preparation for the data analysis, we perform various tests to optimize our search for variables in Sect. 4. Our discussion and summary of the results are given in Sect. 5.

---

[*] Full Tables 1 and 2 are only available in electronic form at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via
http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/500/917

2. The trend-filtering method

Various aspects of the trend-filtering method have been already described in several other papers (e.g., KBN, Kovács & Bakos 2008). Here we briefly summarize the basic steps and refer the reader to these papers for further details.

The basic idea behind TFA is that any systematics (or trends – we use these terms interchangeably) contaminating the target signal can be linearly decomposed with the aid of some subset of the photometric signals (or light-curves, LCs for simplicity) of the objects observed in the given field. This template light-curve set is expected to exhibit various types of brightness variations that are common in many stars (constant or variable) in the field. There are two steps in the analysis:

- Assuming that the target signal is trend- and noise-dominated, we fit the target signal with a linear combination of a pre-selected template LC set. This fit yields a residual time-series that is supposedly trend-free and contains only the signal (with some distortion – see below) and noise, which is nearly white. This filtered time-series is frequency-analyzed and examined for the presence of dominant frequency components.
- After the frequencies have been identified, one may *restore* the signal, since in the above procedure we assumed that there are only systematics and noise. Since some part of the original signal may have been fitted and subsequently subtracted (unintentionally) from the original time-series, we need to restore the signal by employing a *full* model that contains the *signal+trend+noise*. Once the constituent frequency components are known, we can perform this reconstruction by iteratively approximating the signal and trend part of the observed signal. For periodic signals, the method has been described in KBN, for multiperiodic ones we refer to Kovács & Bakos (2008).

We note the following. The template set is selected in a nearly uniform (X, Y) grid covering the full field. This distribution of templates ensures that the selection of close pairs is fairly unlikely, thereby decreasing the chance of eliminating blended variables that could be further studied by other means (e.g., by large-telescope follow-up observations). A lower brightness limit is also set, since very faint targets are dominated mostly by random noise and are therefore less valuable to trend filtering (see however Mazeh et al. 2007, for possible systematics of faint stars). We do not omit variables from the template, because: (i) variables are relatively rare (e.g., some 5% of the Galactic field stars are variable – see Kovács & Bakos 2008); (ii) we do not know a priori which stars are variable and; (iii) we do not generate additional signal in the target time-series by including variables in the template set (see KBN for the discussion of the effect of a template with sinusoidal variation on a target of pure noise).

For a time-series with a limited number of data points, large number of templates (comparable with the number of data points) will generate correlated noise in the filtered time-series. Although this is a general feature of any type of data fitting, the effect is likely to remain tolerable compared with the level of correlation caused by the systematics in the original time-series. To avoid overfitting, we compute the unbiased estimate of the standard deviation of the residuals and select a template number for which this standard deviation starts to level off. As an additional test, we also demonstrate the insufficiency of the use of a single, but optimally selected template as a substitute of multiple templates (see Sect. 3.2).

The multiple template requirement is consistent with the manner in which the photometry in the MACHO Project database was originally reduced. Alcock et al. (1999) described reduction and calibration of the MACHO Project photometry, noting that each image was divided into 512×512 pixel subimages for separate, parallel and more efficient reduction. Furthermore, a focal plane-wide pattern of zero-point residuals was found, probably caused by variations in the large, dichroic beamsplitter.

3. Preparation of the analysis

Before applying TFA to the MACHO photometry, it was important to investigate the basic properties of the data, because we wanted to use the method in an optimal way, by filtering out systematics at the lowest possible filter order (i.e., template number N_{TFA}).

Because our earlier experiments with the HATNet[1] data demonstrated that fairly large number of templates (on the order of several hundred, see KBN) are necessary to reach the desired white noise level, we needed to investigate this issue in the particular case of MACHO time-series photometry to avoid overfitting the data.

In Sect. 3.1, we describe the data used in this paper. In Sect. 3.2, we present several tests, including those of finding the optimal template number, demonstrating the improvement in the distribution of the frequencies of the main harmonic components of the TFAd time-series, and showing the inadequacy of obtaining the best fit for a single template. We also estimate the false alarm level based on Gaussian time-series simulations.

3.1. The data

We selected subsamples from two neighboring LMC fields (#1 and #79) to create sample databases containing some 5300 stars from the bright tail of the magnitude distribution of each field. Field #79 is densely populated at the Northern edge of the LMC bar, containing in total 532 264 objects. In practice, a MACHO project "object" is a defined location in the focal plane where a sum of a small number of images with higher-than-average image quality exceed the threshold of a the star-finding algorithm. An object may constitute a single star but may also be the combined light of several stars in a single MACHO project resolution element.

Field #1 overlaps #79 and is situated further west at the end of the bar of the LMC. This is a somewhat sparser field containing 503 227 objects detected by the MACHO survey. Both data sets cover a long and continuous time interval of 7.5 years, starting in 1992. We used the MACHO "b" instrumental magnitudes, because of the lower relative noise level judged from our earlier works on RR Lyrae stars (Alcock et al. 2003).

To confirm the unbiased nature of our samples in terms of colors (and implicitly physical properties), we plot the color−magnitude diagrams (CMDs) in Fig. 1. To limit the sample size, exclude a large number of noisy targets, and to reach the overall luminosity level of the RR Lyrae stars, we applied a cut at $V = 20$ mag.

The transformation of the instrumental magnitudes to the (Cousins) R and (Johnson) V magnitudes were discussed by Alcock et al. (1999). We used the formulae given by Alcock et al. (1997) for the LMC (their Eq. (1)). The figure shows that

---

[1] Hungarian-made Automated Telescope Network, see http://cfa-www.harvard.edu/~gbakos/HAT/
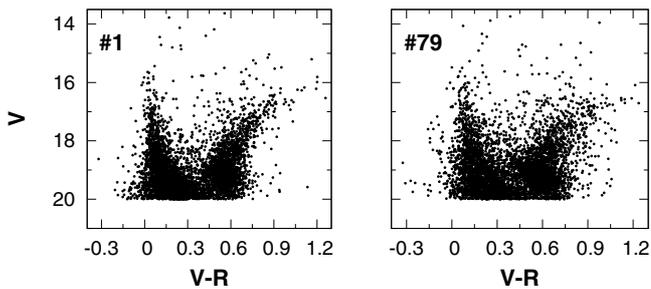
Fig. 1. Color−magnitude diagrams for the stars analyzed in this paper in the LMC fields #1 and #79. Data were cut at $V = 20$ mag.
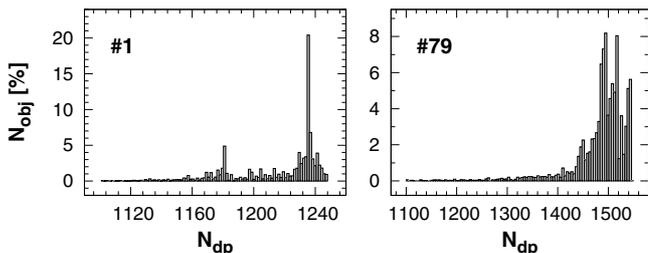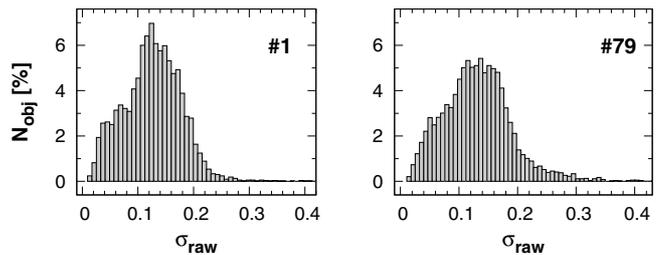

Fig. 2. Distribution of the number of constituent data points in the light curves of the stars analyzed in this paper in the LMC fields #1 and #79.

Fig. 3. Distribution of the standard deviations of the light curves analyzed in this paper from the LMC fields #1 and #79.

the sample used in our analysis covers the expected full color range of objects observed by the project (see the CMD of Alves et al. 1999, for nine million objects from the LMC).

Because TFA requires that all LCs correspond to identical observation epochs, it is important to check how well this uniform sampling condition is satisfied. Figure 2 shows the distribution of the number of data points for each LC. As expected, #79 exhibits a distribution different from that of #1, because of the more frequent observations of the LMC bar field in the case of #79. We note that 80% of the objects have a sizable number of photometric measurements in the ranges of 1180−1250 and 1460−1550 for #1 and #79, respectively. All stars in the TFA template sets are brighter than $V = 18.6$ mag and none of the objects in either field in the above samples have fewer than 1000 data points (objects containing fewer data points were not selected – we see from the figures that this cut is justified). This allowed the application of TFA to all stars, special attention being paid to the low-N cases, when the empty values are replaced by the average of the time-series. We note that this mandatory extension of the target time-series has little effect on the final result, because: (i) there are only a few stars with significantly fewer data points than average; (ii) we used a common sampling time base only for frequency search; and (iii) the final result (i.e., the reconstructed LC) is based on "target-adapted" data points, i.e., we sample the templates exactly on the time base of the target.

It is useful to check the distribution of the standard deviations, since the selection of outliers and other processes affecting the final data quality depend on this quantity. Figure 3 shows that both fields have similar error distributions, perhaps with a slight surplus of objects with errors larger than ∼0.2 mag in field #79. This suspected difference can be attributed to the ∼6% higher surface density of objects in #79 relative to #1. We note that in the analyses presented in this paper, we employed a fairly liberal upper limit of 0.4 mag for selecting erroneous data points based on the errors derived from the images by the MACHO pipeline. In addition, further data points were discarded by an iterative 5σ-clipping based on the standard deviation of the time-series.

The overall value of the standard deviations of the TFA-filtered LCs toward the faint limit of $V = 19.0$ mag is about 0.15 mag. In assessing the minimum detectable amplitude in this regime, we assumed that the LCs are mostly dominated by white noise. From the approximate $A/(\sigma\sqrt{2/N}) > 7$ condition (valid in the high signal-to-noise ratio [SNR] limit – see, e.g., Kovács 1980), for field #79 with an average number of data points of $N = 1400$, we derived a 7σ lower limit of ∼0.04 mag to the amplitudes of periodic signals detectable even toward this faint end of our sample.

3.2. Tests for optimum analysis

First, we checked the TFA template number (N_{TFA}) dependence of σ_{TFA}, which is the unbiased estimate of the standard deviation of the residuals remaining after subtracting the systematics from the original time-series. The lack of consideration of the further decrease in the standard deviation caused by signal reconstruction is reasonably justified, since fewer than 5% of the stars are variables, where reconstruction really matters. The unbiased estimate was computed as derived in Appendix A.

Figure 4 shows the variation in σ_{TFA} for 100 stars randomly selected from each field, which exhibit integer d^{-1} systematics. For comparison, we also plot the results obtained for pure Gaussian white noise generated on the time base of the same stars. We see a decrease in σ_{TFA} with the increase in N_{TFA} for the observed data. The expected constant level of the test data is also clearly visible. From the individual results (thin lines), we see that there are fairly large differences among the various objects with respect to the variation of σ_{TFA}. For "well-behaved" systematics and extensive data sets, we expect a steadily decreasing σ_{TFA} from low template numbers to some optimum value that then levels off to an almost constant value. This pattern is present approximately in only a few cases. In general, we see a nearly monotonic decrease in σ_{TFA} for all template numbers tested. We suspect that this behavior is related to the large variety of systematics that still cannot be determined even with the largest number of templates tested. Based on the above test and the one described below, we fix the template number to 600 and 500 for fields #79 and #1, respectively, as a compromise between overall data number and trend filtering level.

In a subsequent test, we showed that the above choice of template number is adequate for filtering out prominent systematics. We followed KBN, where the filtering capability was tested by the comparison of the occurrence rates of the peak frequencies found by the box-fitting least squares (BLS) analysis (see Kovács et al. 2002) before and after the application of TFA. We then performed the same type of test but using discrete Fourier transformation (DFT, see e.g., Deeming 1975), since we were mostly interested in the oscillation components of the data. The resulting distribution functions are shown in Figs. 5 and 6. We
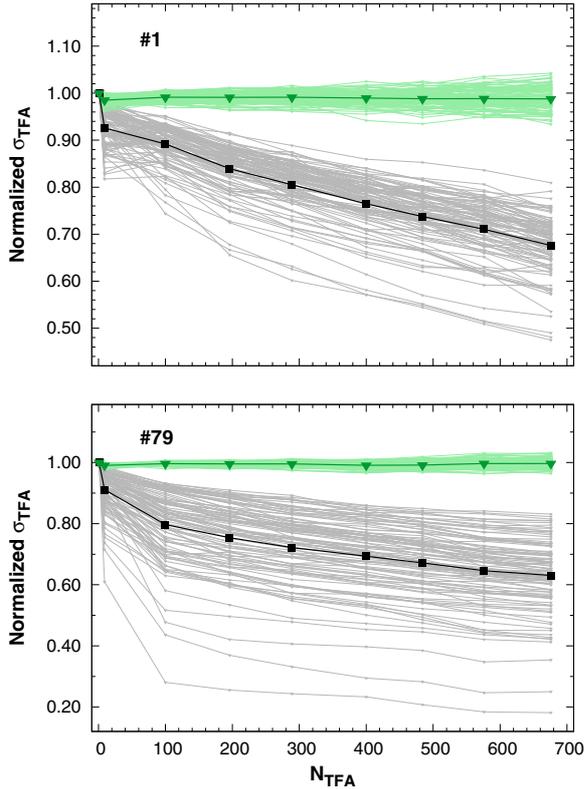
Fig. 4. Dependence of σ_{TFA} (see Appendix A) on N_{TFA}. Near horizontal lines are for pure white noise test data. Thin lines are for the individual objects. Thick lines show the average values computed from the 100 objects. Each function is normalized to 1.0 at $N_{\mathrm{TFA}} = 0$, i.e., for the standard deviation of the original (non-TFAd) data.

Fig. 5. Distribution of the peak frequencies for the full sample of ~5300 objects in field #1. The DFT analysis was performed in the frequency range of $[0.0, 10.0]$ d^{-1}. *Upper panel*: original (RAW, non-TFAd) data, *lower panel*: TFAd data with $N_{\mathrm{TFA}} = 500$.

Fig. 6. As in Fig. 5, but for field #79. The TFAd data were obtained with $N_{\mathrm{TFA}} = 600$.

see that the original (RAW) data exhibit very strong systematics associated with various forms of the common 1 d^{-1} effects (e.g., change in the point spread function, and color-dependent absorption, etc.). All of these are absent in the TFA-filtered data and the frequency distributions are nearly flat, as expected for a randomly selected sample of stars. At the low-frequency end, part of the remaining objects are real (long-period) variables as we discuss them in Sect. 4. In addition to the cyclic systematics, TFA is capable of handling also outlier data points that originate in various transients (e.g., incorrectly subtracted cosmic ray hits). This property is obviously not exhibited in Figs. 5 and 6, but can be seen in several phase-folded light curves (in Sect. 4, we show an example of this).

The current implementation of TFA is not optimal. We selected a large sample of templates, which were used for all targets in the field. Since we did not know a priori which template members are "valuable", we needed to select a sufficiently large set to be more confident that the necessary templates for any given target were included. In keeping the "most relevant" templates (or the linear combinations of those), some additional selection could be incorporated by using principal component decomposition, in a similar way to that of the SysRem algorithm of Tamuz et al. (2005). However, applying this or any other similar method, we had to make decision where to cut the main principal components. Certain systematics (e.g., transients) may have escaped our attention, since they were not part of the main principal components: this was because the statistics we used (e.g., the eigenvalues) in the cut may be insensitive to the given systematics due to their relatively low incidence rates, low amplitudes, or short durations.
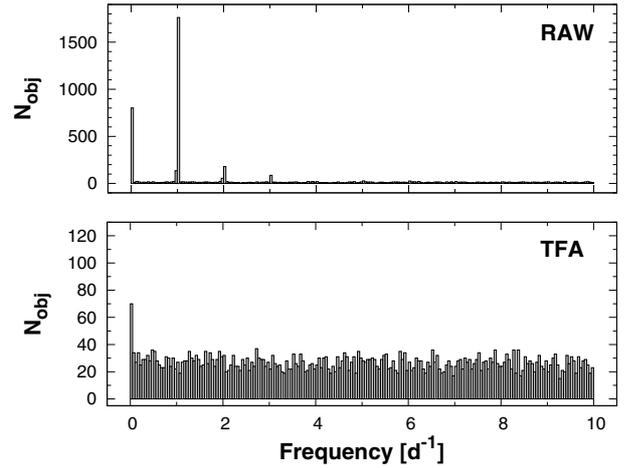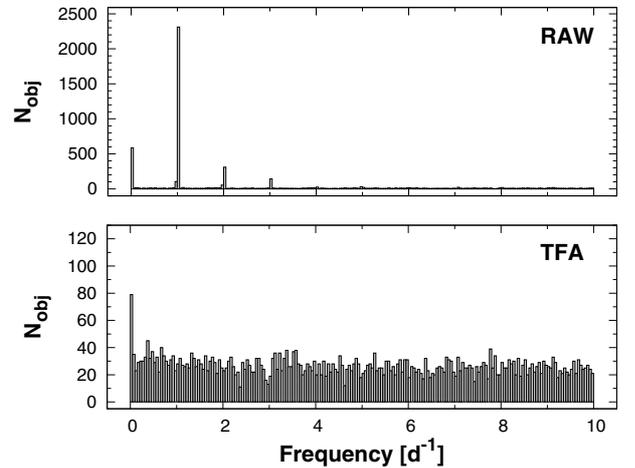
In another effort to identify a more economical selection of templates, we tested whether the best-fit single template selected from the standard TFA template set provided similar results to the algorithm used in the original framework devised by KBN. By using the large template sets chosen above in this section, we computed the unbiased estimates of the standard deviations (see Appendix A) for all stars in the two fields.

For each object, we selected the single template yielding the lowest σ_{TFA} value. In Fig. 7, we plot the ratio of these standard deviations versus the standard deviations computed within the original framework of KBN. It is clear that the optimal single template selection *always performs less effectively* than the standard multi-template method. There is a dependence of the gain of the multiple-template implementation on the LC scatter. This effect is due to the increase in the random noise of the light curves that cannot be filtered out by TFA. Although the peak frequency distribution for the above optimum single-template fit is similar to that of the RAW data, we also note that there is a visible decrease in the number of stars with integer d^{-1} frequencies. These all imply that the multi-template implementation of TFA performs more effectively by a significant amount than its optimized single-template variant.
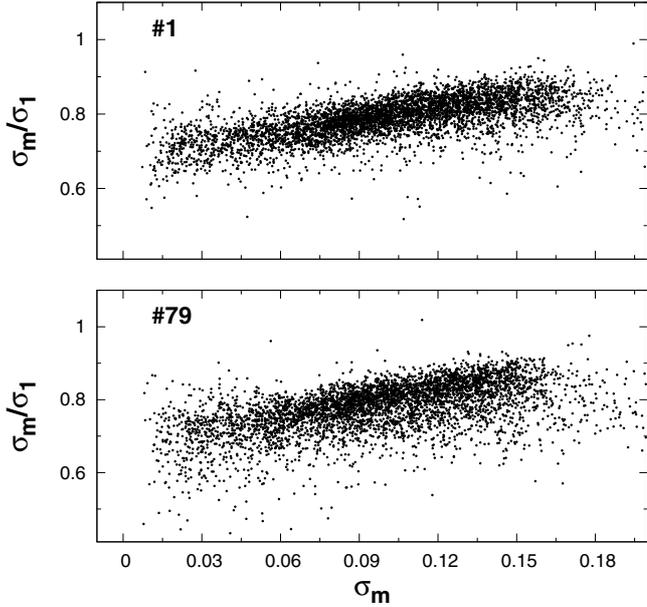
Fig. 7. Unbiased estimates of the standard deviations obtained in the multi-template (σ_{m}) and optimum single-template (σ_1) applications of TFA.

Fig. 8. False-alarm probability (FAP) diagram for the data used in this paper. Inset shows the close neighborhood of the cutoff signal-to-noise ratio (*SNR*) of 7, above which FAP is lower than 1%.

We tested the signal-to-noise ratio above which one can declare a signal to be a significant detection. Although the statistical properties of DFT for continuously sampled data are well known (e.g., Foster 1996, and references therein), for both gapped and randomly sampled datasets the situation could be more complicated. Therefore, we estimate the confidence level by following the same method as described in Nagy & Kovács (2006). Over a thousand time-series were generated by using pure Gaussian white noise on the time base of randomly selected objects in each field. These time-series were treated in the same way as the true observational data, and TFA/DFT analyses were performed in the [0.0, 10.0] d^{-1} frequency range (with 500 and 600 templates in fields #1 and #79, respectively). After finding the peak frequency, we computed the signal-to-noise ratio, defined as follows

$$SNR = \frac{A_{\mathrm{peak}} - \langle A_\nu \rangle}{\sigma_{A_\nu}}, \tag{1}$$

where A_{peak} is the amplitude at the highest peak in the spectrum, $\langle A_\nu \rangle$ is the average (over the frequency values $\{\nu\}$ of the spectrum in the passband of the analysis) and σ_{A_ν} is its standard deviation, computed by an iterative 5σ clipping. As follows from the above definition, we note that for colored noise, *SNR* is a function of the frequency range used in the analysis. In the most common situation of decaying low-frequency noise (the so-called "red noise" – see, e.g., Pont et al. 2006), by increasing the frequency range, *SNR* also increases. Here we assume that in the frequency interval used, the data were well-filtered from systematics (corresponding to the main source of red noise) and the flat (white) noise spectra generated in the simulations represent the true distribution closely enough.

The distribution functions of the *SNR* values derived in this way are displayed in Fig. 8. It is remarkable that the pattern of the false alarm probability (FAP) is very similar for both fields (the small differences can be attributed to the finite sample size). Based on this diagram, we placed our cutoff value at *SNR* = 7.0, corresponding to ~1% FAP. Although this limit suggests a low rate of false detection, in practice the rate of rejected
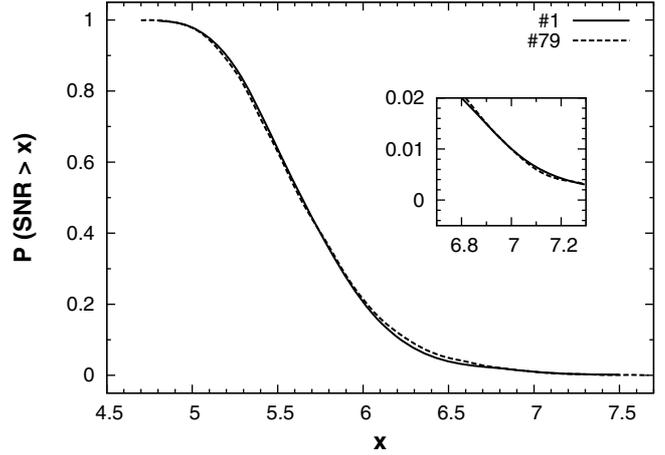
detections was much higher, because visual inspection and other criteria (e.g., blend situation) can reveal additional signatures of false detection. From the originally selected 450 variables satisfying the *SNR* > 7.0 criterion, only slightly greater than 80% therefore survived and entered our variable list. Decreasing the *SNR* cutoff to 6.5, we would have been able to select 1015 variables, for an estimated very high true false-alarm rate, since the formal rate is already ~4% at this *SNR* value. On the other hand, a more stringent criterion of *SNR* > 7.5 would have caused the loss of some of the interesting variables, because this criterion would have been satisfied by only 330 objects, that is 10% lower than the number of the finally selected variables, based on the DFT analysis.

4. Frequency analysis

We performed DFT and BLS frequency analyses on both fields by using the optimal template numbers of 500 and 600 for fields #1 and #79, respectively. To avoid faint, noise-dominated stars, the template members were selected from stars brighter than $V = 18.6$ mag. We also performed the same analyses on the RAW data, to estimate the effectiveness of TFA. The frequency bands of the searches were different for the BLS and DFT runs, since the line profiles in the BLS spectra are much narrower that those in the more traditional Fourier spectra (see Kovács et al. 2002). For the BLS analysis, we considered the [0.01, 1.0] d^{-1} range, thereby covering most of the orbital periods of eclipsing binaries, the prime targets of the BLS search. For the DFT spectra, we have a wider range of [0.0, 10.0] d^{-1}, since we wished to study a wide class of variables from long-period red variables to the much shorter period δ Sct stars (including at least those with sufficiently long periods to correspond to the above frequency range). In both cases, we used 3×10^5 frequency steps.

Variable star candidates were selected in three main steps. First, we selected targets with spectra that exceeded a given *SNR* value. This cutoff value was set to be 7.0 for the DFT spectra and to ~9 for the BLS spectra (this latter value was selected approximately on the basis of our earlier experiments that showed the BLS spectra to be more vulnerable to false alarms). In this way we selected some 200 stars from each of the two fields (many of the variables were common among the DFT and BLS candidates). All variables were then visually inspected and further filtered for possible false alarms. In the course of this, we

Table 1. Variable stars in the MACHO field #1.

ID	V	$V - R$	$f[\mathrm{d}^{-1}]$	SNR	Type	TFA
05011976−6921083	16.889	0.101	1.80367	14.5	B	1
05011992−6919188	17.585	0.049	0.37573	10.8	B	0
05011998−6917495	18.066	0.430	0.252205	16.9	EB	0
05012021−6920315	18.940	0.582	7.82853	7.0	d Sct?	1
05012029−6920248	19.615	0.113	0.02328	13.9	LPV	0

Notes:
− The object identification follows the notation of the 2MASS catalog, namely it is simply the abbreviated J2000 (RA, Dec) coordinates of the object. For example, 05011976−6921083 corresponds to RA = 05:01:19.76, Dec = −69:21:08.3.
− The frequencies for type EB variables are given by a larger number of digits, because of the higher sensitivity of the eclipse shape to the precision of the orbital period.
− The definition of the signal-to-noise ratio (SNR) and classification scheme are described in the text.
− Single digits under the column "TFA" mean exclusive TFA detections if "1" and pre-TFA detections if "0".
− The excerpt shown is part of the full table, which is only available at the CDS.

also assigned some preliminary classification to the variables. In a final step, the independent classifications of the authors were discussed and the final list of variable star candidates was selected. We also performed some supplementary analyses. For example, we compared in more detail the long-period variables in both the RAW and the TFA-filtered time-series and investigated the effect of target-specified time base on the signal detection (an effect that could be important in the case of targets with a smaller number of data points than the average). We show the coordinates and the most important parameters of the variables obtained in this way in Tables 1 and 2[2].

We caution that our classification is both *preliminary* and *approximate*. Except for the trivial cases (e.g., typical fundamental mode RR Lyrae stars), variable star classification (i.e., the first step toward estimating the physical parameters) is difficult for heterogeneous stellar systems. In our simple classification scheme, we considered: (i) the period; (ii) the approximate brightness in V and color $V - R$; and (iii) the light-curve shape. These helped us to identify 11 classes, including the wide group of long-period variables (LPVs), objects on the AGB and more advanced evolution stages and, of course, miscellaneous variables, which exhibit definite variability but indefinite classification based on the parameters listed above. The broad definitions of the variable classes are as follows.

LPV: not eclipsing and period longer than ∼10 days and brighter than the HB level of $V \sim 19$ mag.
FU Cep: fundamental mode Cepheid.
FO Cep: first overtone mode Cepheid.
SO Cep: second overtone mode Cepheid.
RRab: fundamental mode RR Lyrae.
RRc: first overtone mode RR Lyrae.
EB: eclipsing binaries of any type.
B: ∼B-type pulsators, $V \lesssim 18.0$, $V - R \lesssim 0.1$.
δ Sct: frequency is greater than ∼5 d^{-1} and $0 \lesssim V - R \lesssim 0.3$.
Hump: hump/eruption in the light curve.
Misc: anything that cannot be classified.
The following comments apply to the variable star inventories provided in Tables 1 and 2.

Table 2. Variable stars in the MACHO field #79.

ID	V	$V - R$	$f[\mathrm{d}^{-1}]$	SNR	Type	TFA
05122727−6920387	16.948	0.907	0.04885	8.9	LPV	1
05122740−6919308	16.270	0.091	1.31131	11.0	B	1
05122742−6903459	18.708	0.552	7.12253	7.2	d Sct	1
05122749−6919426	18.831	0.424	2.76320	26.5	RRab	0
05122754−6918475	18.626	0.046	0.47837	7.8	Misc	0

Note: see notes to Table 1.

– Magnitudes were derived by using Eq. (1) of Alcock et al. (1997). We note that these magnitudes are approximate and can only be used as a rough guide to variable classification.
– Variables without a dominant frequency component (leading to high – above the noise level – SNR values) are not included in our lists.
– Blends are defined as stars with positions and periods close to those of the target but which appear to have lower SNR values (and concomitantly lower amplitudes). Except for a few cases, the distinction between the blend and the source was unique. These stars are indicated together with other peculiarities in Appendix B. Blends are not reported in Tables 1 and 2.
– Frequencies can be ambiguous within integer fractions for EB stars.
– There are some near integer d^{-1} variables that may well be LPVs, but this is impossible to decide from the present data. In particular, some of the red, short-periodic Misc variables might be LPVs.
– The slowly varying LPVs – those that change their brightness over the timescale of the observational time span – may be subject to "signal killing". That is, since for long-period targets it is easier to find templates with similar long-term changes, it is also easier to subtract the signal by the proper combination of these templates. However, we checked the original (non-TFAd) LCs of all objects, and we found that only a handful of them disappeared after TFA-filtering. Further examination of these cases revealed that they all had rather large scatter and variations reminiscent of systematics, which we declared them to contain.
– There are some stars classified as RRab?, because the period, color and brightness are in the appropriate ranges. Otherwise the shape of the LC is not of a typical RRab and the amplitude is rather low. This is also true for other types of variables.
– Although the prime selection was made on the basis of low false-alarm probability, there is still some chance that a small fraction (maybe 5−10%) of the variables with low SNR are, false alarms.
– TFA detections are those with one of the following properties: (i) the RAW data show typical systematics at integer d^{-1} peak frequencies that disappear in the TFA spectra with the concomitant appearance of significant peak(s) at non-integer d^{-1} frequencies; (ii) as in (i) but the RAW frequencies are at non-integer d^{-1} frequencies. We also note that when the SNR of the TFA spectrum was higher but the RAW spectrum had the same peak frequency, albeit at a lower SNR level (but above the reasonable detection limit), we considered this to be a non-TFA detection.

In Table 3, we provide the statistics of the variables detected in the two fields. The large number of *exclusive* TFA detections is striking (since these are the signals that did not reach the detection limit in the RAW data). Interestingly, the relative number

**Table 3.** Variables in the samples of fields #1 and #79.

Type	#1		#79	
	ND_{RAW}	ND_{TFA}	ND_{RAW}	ND_{TFA}
LPV	18	24	34	32
FU Cep	1	1	3	0
FO Cep	3	1	0	1
SO Cep	1	1	1	0
RRab	10	4	16	3
RRc	3	3	1	1
EB	27	7	30	21
B	11	8	12	14
δ Sct	1	9	4	5
Hump	1	0	1	0
Misc	9	23	17	25
Sum	85	81	119	102

Notes: ND_{RAW} = number of detections in the original (RAW) data; ND_{TFA} = number of *exclusive* detections due to TFA-filtering.

of TFA detections in both fields is nearly the same. By applying TFA, we have *doubled* the number of detected variable stars. The variability rate (number of variables versus total number of stars analyzed) is 3−5% in the two fields, which is similar to the rate derived from the HATNet database in selected Galactic fields (Kovács & Bakos 2008). We also note that TFA is useful even if the signal is detected already in the RAW data. If systematics are present, then signal reconstruction with the aid of TFA helps to obtain a more accurate signal, and eventually, more accurate stellar parameters.

To exhibit the signal detection and reconstruction power of TFA, we show three examples. In Fig. 9, we plot the DFT frequency spectra of 05011976-6921083 from field #1, a short-periodic sinusoidal variable, classified as type B (blue, relatively bright, presumably pulsating variable). The total (peak-to-peak) amplitude is only 0.03 mag in MACHO instrumental "b" passband. In the spectrum of the RAW data, the true signal frequency is hidden in the 1 d^{-1} side lobe of the systematics. This component is clearly evident in the spectrum after applying TFA. Reconstruction of the signal with this period results in a substantial improvement in both the noise level and the shape of the signal (about which no assumption has been made during the reconstruction). However, we note that the reconstructed signal displays a combination of the true signal and the residuals computed by the full signal model of signal+trend+noise, the latter component simply being the residual after subtracting the signal and trend components from the RAW data. Since the procedure involves the least squares fitting of a considerable number of parameters, the residuals will be *biased*, i.e., they show smaller scatter than they should, because of the additional correlation generated among them by the above fit. This effect (mentioned in Sect. 3.2 and discussed in detail in Appendix A), causes the true noise level to be some 30% higher than shown in the figure. Nevertheless, even with this caveat, the application of TFA clearly leads to a cleaner signal shape (which, in turn, might suggest a different classification – e.g., that of EB – but the data are insufficient to be conclusive in this case).

Figures 11 and 12 show a similarly strong TFA detection for the binary star 05163013-6927029 from field #79, showing up in the data with far higher *SNR* by BLS than by DFT. The BLS frequency spectrum of the RAW data clearly shows the peak structure at $(0.5/n)$ d^{-1}, reminiscence of the daily systematics present in the BLS spectra of most of the light curves. The true signal component is visible as a middle-size peak in the forest of the subharmonics of the daily trend. We note that the
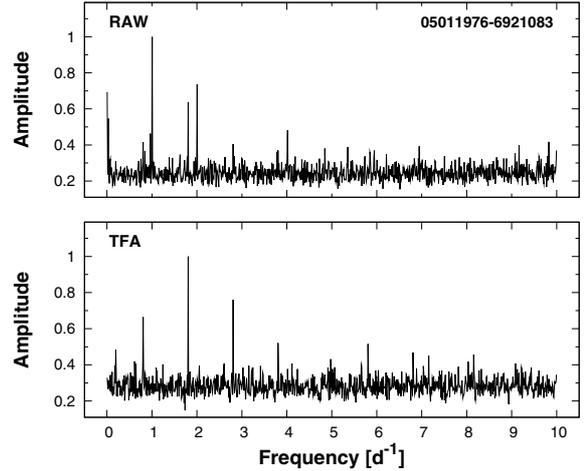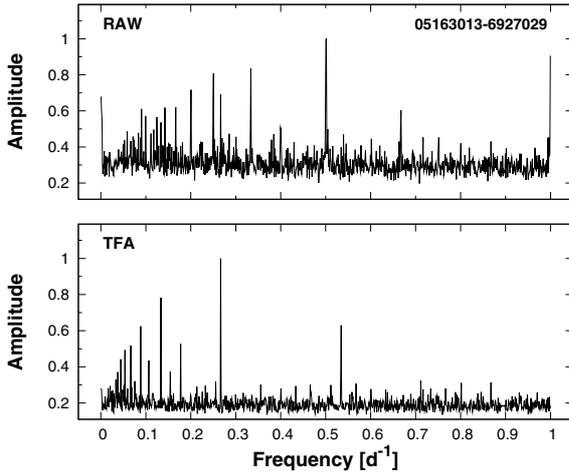

Fig. 9. Example of the detection of a signal strongly dominated by systematics (see the high peaks at integer d^{-1} frequencies in the RAW spectrum). The variable is from field #1.
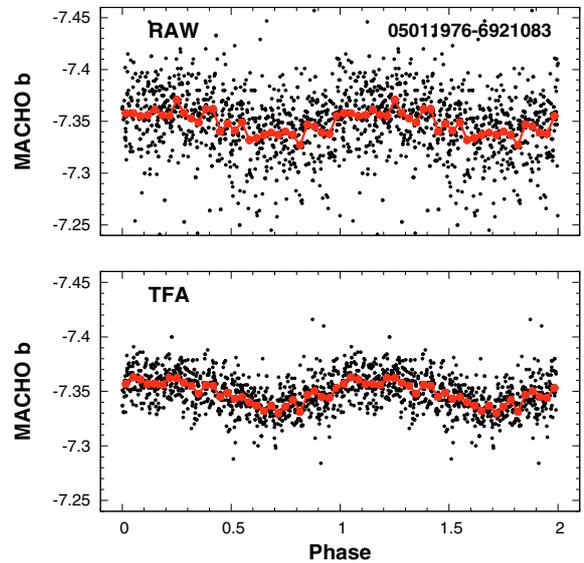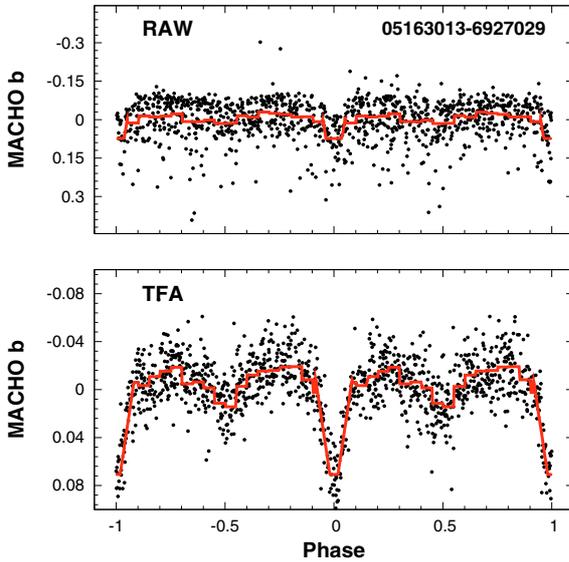

Fig. 10. TFA reconstruction of the B-type variable shown in Fig. 9. Please check the note in the text on the scatter of the TFA-reconstructed light curve. Bin averages are shown by thick dots.

confusion in the case of BLS spectra is exaggerated due to the (sub)harmonic structure inherent to the BLS method (or to any other method, based on period-folding – see Kovács et al. 2002). The reconstructed and folded LC shown in the lower panel of Fig. 12 clearly exhibits a shallow secondary eclipse, which is not evident at all in the RAW LC. The bias factor for the residual scatter is the same here as mentioned in the context of Fig. 10, since for field #79 we have more data points, but, at the same time, we also have more extended TFA templates.

Our third example demonstrates that systematics may not be exhibited in the Fourier domain. As mentioned in Sect. 3.2, transients may not be properly subtracted from the images. Since these phenomena are non-periodic, they usually do not affect the frequency spectra in a significant way. Figure 13 shows the DFT spectra of such an object (an RR Lyrae star). The signal is easily detected in the RAW data. We also see a decrease in the *SNR* in the TFA spectrum, because the observed signal is strong (relative to the systematics), and in the signal search mode, TFA assumes a zero target function. When this assumption is no
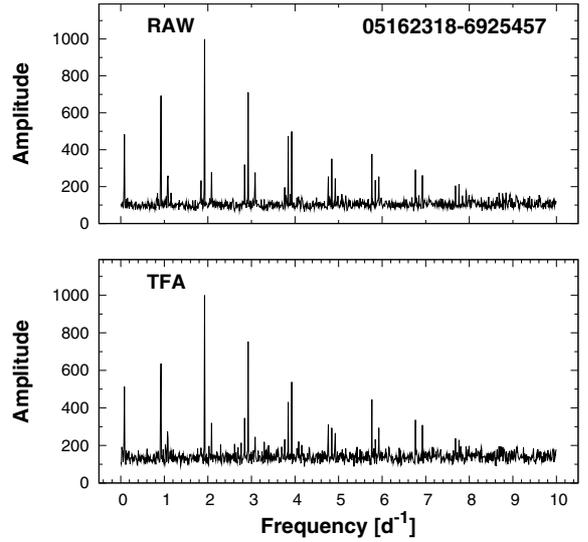
Fig. 11. Example of the detection of a signal nearly completely smeared by systematics in the original (RAW) data. The frequency spectra were computed by the BLS algorithm. The variable is from field #79.
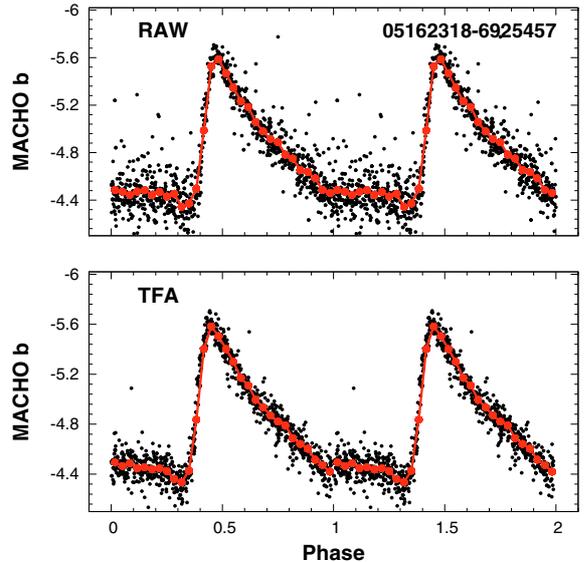

**Fig. 12.** TFA reconstruction of the EB variable shown in Fig. 11.

Fig. 13. Example of the detection of a DFT signal in the RAW data. The particular systematics present in this object do not affect the DFT frequency spectra. The lower frequency sidelobes visible at the alias components are due to the first harmonic (and the aliases) of the fundamental frequency. The variable is from field #79.

Fig. 14. TFA reconstruction of the RR Lyrae variable shown in Fig. 13. Systematics are exhibited as outlying data points, causing only a small effect on the frequency spectrum.

longer made, i.e., a full model is used and a signal reconstruction is performed, we obtain the result shown in Fig. 14. Nearly all outlying data points are successfully corrected. A small change in the LC shape is also observable.

To locate the variables approximately in the physical parameter space, in Fig. 15 we overplot the detected variables on the CMD of #79. We see that the variables basically follow the distribution of the constant stars, apart from the luminous red variables and perhaps some mild surplus of red clump variables centered on $V = 19.2$ and $V - R = 0.6$. The lack of variables on the giant branch between $V = 18$ and 17 mag is striking. As expected from their short evolutionary lifetimes, there are also not many post AGB stars. Although we have a reasonably large number of RR Lyrae variables (41 altogether), they do not seem to form a well-defined separate group at the horizontal branch between $V = 19$ and 20 mag, but they are intermingled with other types of variables. However, the main sequence is nearly uniformly populated, from the bright end to the faintest stars in our sample.

5. Discussion and conclusions

We have employed the Trend Filtering Algorithm (TFA) combined with Fourier (DFT) and box-search (BLS) frequency analyses to explore the variable star content of a subset of the MACHO Project database for the Large Magellanic Cloud. From direct analyses, a large part of the database yields too many false signals because of the strong systematics present in the data. We found that about 50% of detectable variables were not detected by standard strategies in the MACHO Project database for the samples we investigated (~5300 objects in each of the fields #1 and #79, down to $V = 20.0$ mag). About 60% of all stars are dominated by trends.
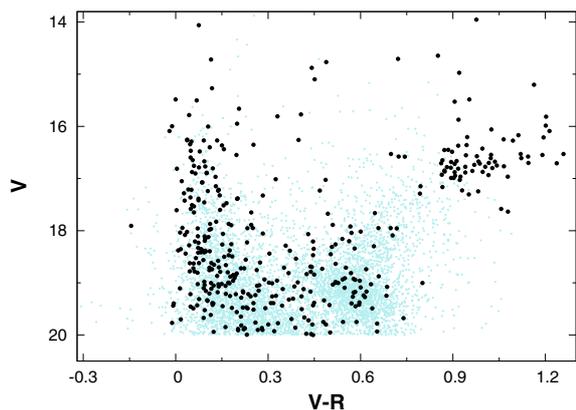
Fig. 15. Color−magnitude diagram of field #79. Overplotted are the 387 variables detected in fields #1 and 79.

Due to the relatively small number of data points per object, we investigated the optimum number of TFA templates to be used in the survey. In an ideal case, the representative model of a time-series is reached when the increase in the number of fitting functions does not lead to a further decrease in the unbiased estimate in the standard deviation of the residuals between the input time-series and the model. Unfortunately, for the MACHO project time-series, it seems that the required number of templates exceeds half of the number of data points. This would lead to a prohibitive overuse of the TFA fit and, as a result, to a significant increase in false alarms. Fortunately, the distribution of the peak frequencies of the Fourier spectra of the TFA-filtered time-series – even with sub-optimal template numbers – shows that the most significant systematics exhibited as integer fractions of the diurnal periodicity disappear. We also examined the case of very long-period variables (showing basically a linear trend), where TFA is more likely to distort the light variation and lead eventually to the reduction in the true photometric variation. Although we found that the variations were real in a few cases and TFA distorted the signal considerably, visual inspections of the original (non-filtered) light curves led to the conclusion that, in general, the signals were most likely long-term systematics.

We found altogether 387 variables from the total of ~10 600 objects analyzed in the subsets of the two fields. This variability ratio is very close to the value derived for a subsample of the HATNet database (Kovács & Bakos 2008). This coincidence is most probably accidental, since HATNet surveys the Galactic field stars and also detects many variables in the millimagnitude amplitude regime that, for the majority of objects, is inaccessible to the MACHO project survey. Nevertheless, it is remarkable that with TFA, we are also able to detect many low-amplitude variables in the 10 mmag regime in the MACHO project database. We understand that most of these variables are indeed "normal"-amplitude variables, but due to significant blending they appear as low-amplitude variables.

We found that the application of TFA to variable searches increases our ability to find many low-amplitude variables that might be inaccessible in the raw photometric data. This may lead to exciting new studies in various fields related to variable stars. Here we mention only three topics of immediate interest.

- Low-amplitude pulsating variables across the HR-diagram: there are of course interesting questions in many respects. For example, the excitation map of B-type stars is not well known, because of the delicate dependence on the metal content (Karoff et al. 2008). The limits of variability in the

δ Scuti regime is still to be determined. This would be especially interesting, since linear pulsation theory predicts excitations throughout this part of the HR-diagram (e.g., Breger et al. 2008). There are regions above the Horizontal Branch (HB) that remain to be explored: the red-clump stars, the high-metallicity, low-temperature "cousins" of the HB stars; variables on the giant branch and in the AGB phase. The change of the role of convection in the context of the classical κ-mechanism as we reach higher luminosity levels is not understood. The relatively well-confined instability strip of classical pulsators becomes poorly defined at higher luminosities. Although recent investigations explored many interesting properties of the red giant variables (e.g., Soszyński et al. 2007, and references therein), they also revealed their complexity and mixed physical nature.

- Ultra low-amplitude (ULA) classical pulsators: although the physical mechanism in establishing the borders of the instability strip of Cepheids and RR Lyrae stars is more or less known, it is still unclear how the pulsation stops and if other forms of pulsation at low amplitudes could exist outside the instability strip. From a subsample of the MACHO LMC Cepheids, Buchler et al. (2005) found 14 objects with characteristic amplitudes of 0.005−0.01 mag. Interestingly, these stars apparently follow the PLC relation spanned by the "normal" large-amplitude Cepheids. The above sample of ULA Cepheids was extended by Soszyński et al. (2008). Some theoretical considerations and numerical simulations (Buchler & Kolláth 2002; Bono et al. 1995) suggest that in the case of "soft" bifurcations (when the pulsation growth rates are in the same order as the evolutionary rates), the transition from various pulsation states may last several thousand years, so we may have rare, but observable events.

- Shallow-eclipsing binaries: these are important objects in extending the determination of stellar parameters toward the very low-mass (<0.1 $M_\odot$) regime. Only a handful of objects span the low-mass tail of the mass-radius relation, so discovering additional objects would be important in clarifying the source of the systematic difference between the observations and the currently available theory (Beatty et al. 2007). Depending on the configuration, the depth of the eclipse can be very low, but a typical M star with an F primary yields a ~0.01 mag deep eclipse. It is clear that detection of these events (even if it is periodic) requires good quality data avoiding contamination with systematic effects.

Although the discovery of low-amplitude variables is an important task, it is also a difficult one. We not only have a faint signal but it is also intermingled with the effect of colored noise due to stellar activity and instrumental systematics as discussed above. Furthermore, on the top of these, there is the blending issue. This problem has been highlighted by the search for extrasolar planets by transit methods (Brown & Latham 2008) and problems concerning the measured PL relations in distant galaxies (Vilardell et al. 2007). The problem is quite common in all photometric works (Kiss & Bedding 2005) and can be tackled at a certain degree by various data reduction methods (i.e., image subtraction and profile analysis, see Hoekstra et al. 2005), but in difficult cases, the combined efforts by experts in large telescope photometry, spectroscopy, and stellar modeling are required (see, e.g., Mandushev et al. 2005).

Appendix A: Unbiased estimation of the standard deviation of the residuals of least squares fits to sparsely sampled target functions

Here we derive a formula for the unbiased estimate of the standard deviation of the residuals when the target function is sparsely sampled relative to the fitting functions of a Least Squares (LS) problem. This situation may occur when applying TFA, which adopts a filter with a template set defined on a more extended time base than that of the target (a situation that can be common in photometric databases containing data of diverse quality).

The LS problem is defined in the standard way

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} \left[y(i) - \sum_{j=1}^{m} a_j x_j(i) \right]^2 = \min, \tag{A.1}$$

where $\{y(i)\}$ is the target function (in the following: time-series), $\{x_j(i); j = 1, 2, ..., m\}$ are the fitting functions with the corresponding regression coefficients $\{a_j\}$. We note that the fitting functions might include any functions that contribute to the signal; for example, they might contain Fourier components and TFA templates. At the minimum variance, s stands for the root mean squares (rms) of the residuals. Our assumptions are the following:

- All $\{x_j(i)\}$ are sampled in the same time base $\{t(i)\}$.
- The sampling time base $\mathcal{T}$ of $\{y(i)\}$ is a nonzero subset of $\{t(i)\}$. This part of $\{y(i)\}$ contains n_1 data points.
- $\{y(i)\}$ is assumed to be zero-averaged with "zeros" at the moments where it is not defined (in respect to $\{t(i)\}$).
- The noise $\{\eta(i)\}$ is additive in $\{y(i)\}$ with the following expectation values: $E(\eta(i)) = 0$, $E(\eta(i)\eta(j)) = \delta_{ij}\sigma^2$.
- The noiseless signal can be represented by the assumed model given by the best-fit linear combination of $\{x_j(i)\}$.

With the above assumptions, the target time-series reads as follows

$$y(i) = \begin{cases} 0 & \text{if } t(i) \notin \mathcal{T} \\ y_0(i) + \eta(i) - \langle y_0(i) + \eta(i) \rangle & \text{if } t(i) \in \mathcal{T} \end{cases}$$

Here $\langle y_0(i) + \eta(i) \rangle = \frac{1}{n_1} \sum_{i \in \mathcal{T}} [y_0(i) + \eta(i)]$, $\{y_0(i)\}$ is the noiseless part of the target function. With $g_{kl} = \sum_{i=1}^{n} x_k(i)x_l(i)$ and $b_k = \sum_{i=1}^{n} x_k(i)y(i)$ the solution of the LS problem becomes

$$a_j = \sum_{i=1}^{m} h_{ij} b_i, \tag{A.2}$$

where $\{h_{ij}\}$ is the inverse of $\{g_{ij}\}$ ($\sum_{j=1}^{m} h_{ij} g_{jk} = \delta_{ik}$). With these, the expectation value of the square of rms can be written in the following form

$$E(s^2) = \frac{1}{n} \sum_{i=1}^{n} y^2(i) - \sum_{j,k=1}^{m} h_{jk} b_j b_k. \tag{A.3}$$

To compute the unbiased estimate of σ^2, we need to evaluate the expectation value of s^2. This is a straightforward computation, but we need to consider that the lack of data points in $\{y(i)\}$ as given above, will introduce extra correlation. Finally, we end up with the following expression

$$n \frac{E(s^2)}{\sigma^2} = n_1 - 1 - \sum_{j,k=1}^{m} h_{jk}(C_{jk} - S_j S_k), \tag{A.4}$$

where $C_{jk} = \sum_{i \in \mathcal{T}} x_j(i)x_k(i)$ and $S_j = n_1^{-\frac{1}{2}} \sum_{i \in \mathcal{T}} x_j(i)$. It is easy to see that the above expression reduces to the familiar formula for the unbiased estimate of the variance if all items of the target function have time values matching those of the fitting functions ($C_{jk} = g_{jk}$) and if the latter are also zero-averaged ($S_j = 0$).

Appendix B: Comments to the variable selection of Sect. 4

Notes on #1:

- 05021002-6850573 (*SNR* = 33.1, Typ = RRc?) is a close companion of 05021018-6850566 (*SNR* = 25.9, Typ = B?). Inspection of the LCs does not support an obvious blend scenario although their proximity strongly suggests that this is the case.
- 05022450-6850457 has features that may indicate that this is a ~9 d^{-1} (!) EB.
- Stars such as 05022832-6853097 with *SNR* = 7.7 have been omitted, because their LCs were ragged. In this particular case, the alias components at ~1 d^{-1} are similarly strong, but we did not explore the possibility that this was the true frequency.
- Some of the red, short-periodic Misc candidates might actually be LPVs, but we cannot be certain because of alias problems. For example, 05055618-6844166 was classified earlier as "Misc", but: (i) the difference between the heights of the long- and the short-periodic peaks is less than 5%; (ii) the long-periodic LC indeed shows long-term features; and (iii) the color is very red. So it has been re-classified as "LPV". The frequency and *SNR* given in Table 1 are those of the RAW analysis. Also, 05014177-6920207 has a frequency of nearly ~1 d^{-1}. The frequency close to zero is also viable. The color is very red, implying that this could be an LPV.

Notes on #79:

- Stars 05122749-6919426 and 05122739-6919434 are blends of each other. Since 05122749-6919426 has a slightly greater *SNR*, we left this object as a variable in the database. However, this choice is not well justified. Only further observations with high resolution images can decide which of these stars is the true source of the variation. SuperMACHO Project images go deeper, have higher image quality, and could resolve this ambiguity.
- 05123435-6904427: TFA yields a bit higher power at the 1 d^{-1} alias. After visual inspection of the unfolded time-series, we classified this variable as LPV.
- 05125486-6918537: the RAW light curve is noisy and ragged with a long shallow dimming in the first half of the observational time span. After TFA, the signal characteristics of *SNR* = 7.3 and $f_{\text{peak}} = 1.00329$ d^{-1} did not seem to be convincing enough, to consider the variability as real.

- 05130892-6919317: both RAW and TFA yield long periods, but it is actually aperiodic within the observational time span. Therefore, we assigned the RAW period and *SNR* to this star.
- 05161575-6926342: some part of the long-term variation has been suppressed by TFA.
- 05162206-6925526: there is a relatively strong signal deformation by TFA but the LPV classification is secure.
- 05162318-6925457: is a heavily blended RRab (the three close companions are 05162314-6925413, 05162251-6925440 and 05162257-6925491).
- 05162865-6926167: TFA eliminates both the long-term drift and the short-term variation (approximately on a time scale of one year with amplitudes of $0.05-0.1$ mag). It is suspected that these variations are related to some systematics.
- 05162730-6926236: LPV (non-periodic). TFA yields a slightly higher peak at $\sim 1\,\mathrm{d}^{-1}$, that could lead to a misclassification. Frequency, *SNR* entries are those of the RAW data.
- 05163808-6926589: TFA/BLS yields *SNR* = 7.7, therefore, due to the *SNR* cutoff of ~ 9 employed on the BLS spectra, this star did not enter in our prime selection. Visual inspection of the LC folded by the BLS peak frequency of $0.5746619\,\mathrm{d}^{-1}$ suggests an EB classification.

References

Alard, C., & Lupton, R. H. 1998, ApJ, 503, 325
Alcock, C., Allsman, R. A., Alves, D., et al. 1997, ApJ, 486, 697
Alcock, C., Allsman, R. A., Alves, D., et al. 1999, PASP, 111, 1539
Alcock, C., Allsman, R. A., Alves, D., et al. 2003, ApJ, 598, 597
Alves, D., Basu, A., Cook, K. H., Welch, D. L., et al. 1999, IAU Symp., 190, 517
Bakos, G. Á., Torres, G., Pál, A., et al. 2009, ApJ, accepted [arXiv:0901.0282]
Beatty, T. G., Fernandez, J. M., Latham, D. W., et al. 2007, ApJ, 663, 573
Bono, G., Castellani, V., & Stellingwerf, R. F. 1995, ApJ, 445, L145
Bramich, D. M. 2008, MNRAS, 386, L77
Breger, M., Lenz, P., & Pamyatnykh, A. A. 2008, MNRAS, in press [arXiv:0812.0856v1]
Brown, T. M., & Latham, D. W. 2008, PASP, submitted [arXiv:0812.1305v1]
Buchler, J. R., & Kolláth, Z. 2002, ApJ, 573, 324
Buchler, J. R., Wood, P. R., Keller, S., & Soszyński, I. 2005, ApJ, 631, L151
Deeming, T. J. 1975, Ap&SS, 36, 137
Faccioli, L., Alcock, C., & Cook, K. 2008, ApJ, 672, 604
Foster, G. 1996, AJ, 111, 541
Hoekstra, H., Wu, Y., & Udalski, A. 2005, ApJ, 626, 1070
Karoff, C., Arentoft, T., Glowienka, L., et al. 2008, MNRAS, 386, 1085
Kim, D.-W., Protopapas, P., Alcock, C., Byun, Y.-I., & Bianco, F. 2009, MNRAS, submitted [arXiv:0812.1010v3]
Kiss, L. L., & Bedding, T. R. 2005, MNRAS, 358, 883
Kovács, G. 1980, Ap&SS, 69, 485
Kovács, G., & Bakos, G. Á. 2008, CoAst, 157, 82
Kovács, G., Zucker, S., & Mazeh, T. 2002, A&A, 391, 369
Kovács, G., Bakos, G. Á., & Noyes, R. W. 2005, MNRAS, 356, 557
Kruszewski, A., & Semeniuk, I. 2003, Acta Astr., 53, 241
Mandushev, G., Torres, G., & Latham, D. W. 2005, ApJ, 621, 1061
Mazeh, T., Tamuz, O., & Zucker, S. 2007, ASP Conf. Ser., 366, 119
Nagy, A., & Kovács, G. 2006, A&A, 454, 257
Pont, F., Zucker, S., & Queloz, D. 2006, MNRAS, 373, 231
Soszyński, I., Dziembowski, W. A., Udalski, A., et al. 2007, Acta Astron., 57, 201
Soszyński, I., Poleski, R., Udalski, A., et al. 2008, Acta Astron., 58, 163
Tamuz, O., Mazeh, T., & Zucker, S. 2005, MNRAS, 356, 1466
Vilardell, F., Jordi, C., & Ribas, I. 2007, A&A, 473, 847