Mon. Not. R. Astron. Soc. **358,** 30–38 (2005) doi:10.1111/j.1365-2966.2005.08651.x

Automated classification of variable stars for All-Sky Automated Survey 1–2 data

L. Eyer[1,2]★ and C. Blake[1]

[1]*Princeton University Observatory, Ivy Lane, Princeton, NJ 08544, USA*
[2]*Observatoire de Genève, CH-1290 Sauverny, Switzerland*

Accepted 2004 November 17. Received 2004 November 5; in original form 2004 June 25

ABSTRACT

With the advent of surveys generating multi-epoch photometry and the discovery of large numbers of variable stars, the classification of these stars has to be automatic. We have developed such a classification procedure for about 1700 stars from the variable star catalogue of the All-Sky Automated Survey 1–2 (ASAS 1–2) by selecting the periodic stars and by applying an unsupervised Bayesian classifier using parameters obtained through a Fourier decomposition of the light curve. For irregular light curves we used the period and moments of the magnitude distribution for the classification. In the case of ASAS 1–2, 83 per cent of variable objects are red giants. A general relation between the period and amplitude is found for a large fraction of those stars. The selection led to 302 periodic and 1429 semiperiodic stars, which are classified in six major groups: eclipsing binaries, 'sinusoidal curves', Cepheids, small amplitude red variables, SR and Mira stars. The type classification error level is estimated to be about 7 per cent.

Key words: astronomical data bases: miscellaneous – catalogues – surveys – Cepheids – stars: variables: other.

1 INTRODUCTION

Knowledge of the variability of bright objects is relatively poor. Paczyński (2001) denounces this situation vigorously: 'I think this ignorance is inexcusable and embarrassing to the astronomical community.' Yet, at magnitude 12, it is estimated that 90 per cent of the variables are unknown (Paczyński 2000).

In recent years, only the *Hipparcos* satellite has carried out a multi-epoch photometric all-sky survey with an associated analysis aimed at systematically detecting variability. The main mission survey has a mean of 110 measurements per star over 3.3 yr. It goes down to V magnitude 7.3/9.0 depending on the colour of the star and its ecliptic latitude β. The result of this analysis for the main mission is published in volumes 11 and 12 of ESA (1997) (see also the flags H6 and H49–H53 of the main catalogue) and concerns about 11 500 variable stars that comprise about 10 per cent of all catalogue entries. *Hipparcos* enables the description of the behaviour of individual stars by giving mean colour and magnitude, parallax, period, amplitude and epoch of the minimum or maximum light, but also for the stellar variability across the Hertzsprung–Russell (HR) diagram (Eyer & Grenon 1997) to be described globally. It is worth noting that the data classification in the different variable types was 'manual' and incomplete. The Tycho photometric data, a deeper survey using the *Hipparcos* star mappers, caused more problems in its variability analysis and gave disappointing results even after efforts to censor poor quality data points (Piquard et al. 2001).

There are photometric optical surveys which have specific goals, such as the Optical Gravitational Lensing Experiment (OGLE), Expérience pour la Recherche d'Objets Sombres (EROS) and the Massive Compact Halo Object Project (MACHO) for detecting microlensing events, and the Robotic Optical Transient Search Experiment (ROTSE) and the Livermore Optical Transient Imaging System (LOTIS) for detecting gamma ray bursts. We do not discuss these targeted surveys here and refer to Paczyński (2000). They were exploited for other purposes and global variability analyses have been performed (Belokurov, Evans & Du 2003) or are underway (Woźniak et al. 2001; Marquette, private communication).

There are many ways to conduct a survey, because several competing parameters cannot all be maximized. Four primary questions are in competition: how deep, how frequently sampled, how wide, and how precise is a survey? Consequently, different astronomical subjects/objects will be explored/discovered depending on these choices.

It is important to consider this work in view of the general research trend. With surveys from the ground and from space becoming wider and deeper, the number of objects is increasing dramatically and the handling of data is becoming more difficult. With exponential data growth, there is a clear need for automated algorithms.

Here, we analyse the All-Sky Automated Survey (ASAS) survey data obtained during its test implementation (phases 1 and 2). ASAS is described in the next section. An extraction of the variable

★E-mail: laurent.eyer@obs.unige.ch

objects was performed by Pojmański (2000) and we address here the question of classification of the stars of this selection. The number of objects that are considered is about 3900 variable objects, among them about 400 periodic variable objects (Pojmański 2000).

Pojmański (2002, 2003, 2004) has continued to develop the ASAS, and has presented an analysis of the third phase of ASAS. About 1.3/3.2 million stars brighter than $V = 15$ were measured, 3126/10453 variable stars were extracted and classified as eclipsing (1046/1718), regular pulsating (778/731), Mira (132/849) and 'other' (1170/7155), mostly SR, infrared (IR) and long period variable (LPV) stars (the numbers refers to the year up to 2003/2004). The regular pulsating stars have been separated into δ Scuti stars, RRab and RRc stars, δ Cep stars (fundamental and first overtone pulsators). Pojmański's classification methods differ from the method presented here. In his article of 2003, he used carefully selected two-dimensional projections of the Fourier coefficient space where the separation of variable types are pre-analysed and well distinguishable. This method was used by Ruciński (1993, 1997) to distinguish between contact binaries and detached ones. In the study of Pojmański & Maciejewski (2004), they added Two-Micron All-Sky Survey (2MASS) and *IRAS* photometry, using two additional parametric planes (H–K, log P) and (H–K, J–H).

The methodology presented here is divided in three main objectives: (i) to search for periodicity; (ii) to model the light curve and characterize the data by a set of parameters, and to remove dubious cases; (iii) to determine the variability types.

2 DATA DESCRIPTION

The ASAS is a photometric survey. Its goal is to regularly monitor the sky so as to detect any variable phenomenon. In its testbed, during the years 1997–2000, ASAS 1–2 repeatedly measured 50 fields (2×3 deg^2 each) and obtained photometry for about 140 000 stars in the I band with a 135-mm photolens and a 768×512 pixel CCD camera. Among the 140 000 observed stars, a set of 3890 variable objects was extracted by Pojmański (2000).

The limiting magnitude is about $I = 13$ and saturation occurs for stars brighter than $I = 7$. The precision of the measurements is 0.01 at 8 mag and degrades to 0.07 at 12 mag. These estimates were derived from the difference of successive measurements (the underlying assumption being that the variability is negligible for short time-scales; see Eyer & Genton 1999). The quoted individual photometric errors are not always fully consistent with our estimates.

The dates are given in Truncated Heliocentric Julian Dates (THJD) HJD −245 0000. The time sampling and number of measurements are quite diverse. There is a large gap between 585 and 1020 THJD, which was caused by a flood in the shelter containing the telescope and the control apparatus. The histogram of the number of measurements per star is given in Fig. 1 and the histogram of time differences between successive observations in Fig. 2. We also present the spectral window of a typical star in Fig. 3.

For our analysis, extreme values, which were more than four times the dispersion distant from the mean, were removed. We note that such a criterion may remove some points of an eclipsing binary during its eclipses. Different fields had different time coverage and time-span. This data heterogeneity complicates the analysis. Fig. 4 shows the time difference between the last and first epochs for each star. Because large gaps might cause problems in the analysis, only observations after 1020 THJD were considered. However, the data before 585 THJD were used to verify shorter periods.

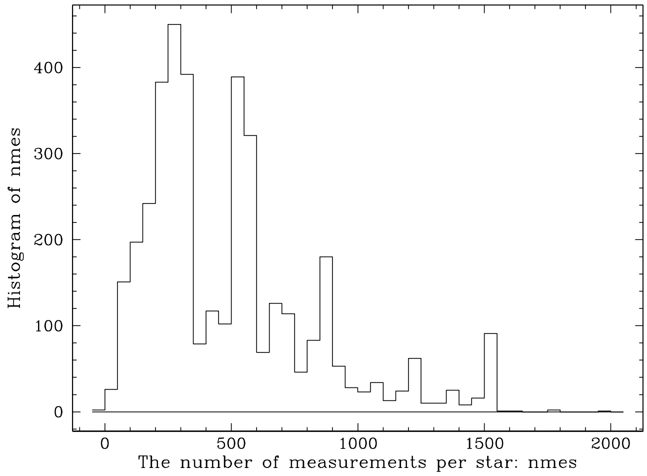

Figure 1. Histogram of the number of measurements per star.

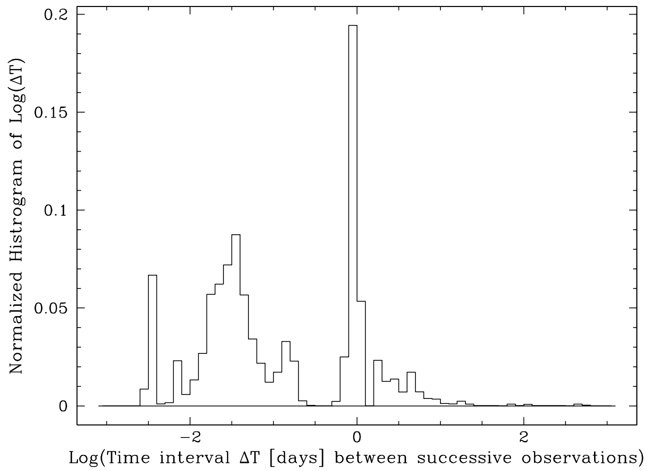

Figure 2. Normalized histogram of time intervals between successive measurements. We can see the regular nightly observing pattern. However, there are other characteristic time intervals, for instance ∼5 or ∼45 min.

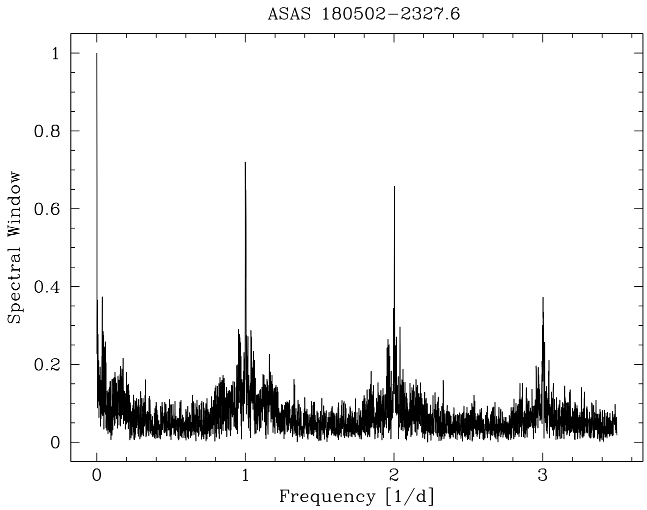

Figure 3. Spectral window of a typical star ASAS 180502–2327.6.

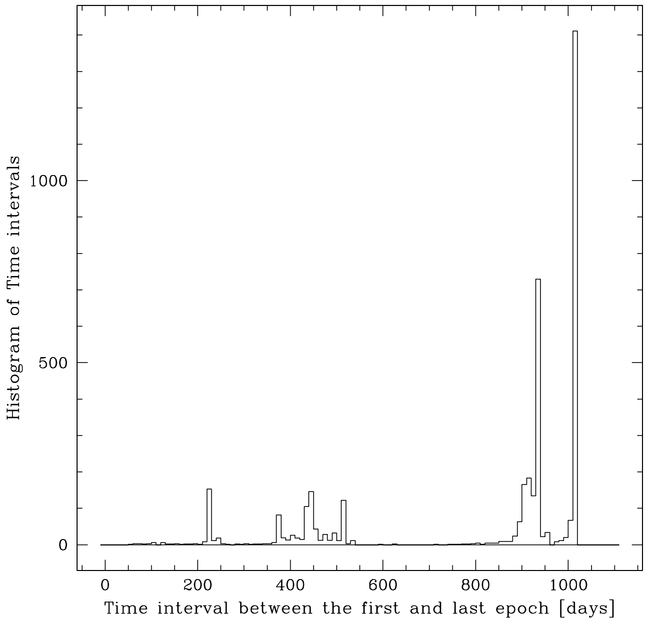

Figure 4. Histogram of the difference between the last and first epochs. The star sample is heterogeneous, but with many stars being covered over a period of nearly 3 yr. However, these stars may have large intervals without observations.

3 PERIOD SEARCH

The data were processed through the Lomb (1976) period search algorithm, which is equivalent to adjusting the parameters of a sinusoidal curve. We used the algorithm given by Press et al. (1992), in its fast version called 'FASPER'. The algorithm has two parameters, which fix the resolution in frequency and the highest frequency searched. The first, OFAC, was set to 6 while the second, HIFAC, was changed according to the properties of the light curve and of the sampling. In FASPER, the highest frequency searched is computed given the time intervals, the number of data points and the ad hoc factor multiple (HIFAC) of Nyquist frequency, as defined by Press et al. (1992). This factor is needed because short periods can be detected even if the data have large time gaps between measurements (see Eyer & Bartholdi 1999). We did not use a constant HIFAC parameter, because in certain cases (large amplitude variables) it is unnecessary to search for high frequencies. On the other hand, for short time variability very high frequencies can be found.

We computed σ_N, the dispersion of the difference of measurements separated by a small time interval (less than 1.5 d), and divided it by the dispersion of the signal. If this ratio was small ($<$0.55), we used HIFAC such that the highest frequency was equal to 0.5 d^{-1}; if it was between 0.55 and 0.66 we used HIFAC $= 9$, and if above 0.66 we fixed HIFAC so that the highest frequency was equal to 17 d^{-1}. This was found to be a better compromise than using a single value for HIFAC. Long time-scale and large amplitude variable stars can show a variety of irregular behaviours and trends, and therefore are prone to aliasing. The determination of the highest frequency to be searched was limited to 17 d^{-1}. This limit was determined for the shortest pulsators which have high amplitudes ($>$0.05) using the δ Scuti star catalogue of Rodriguez, López-González & López de Coca (2000). The choice of δ Scuti stars is motivated by the fact that they have the shortest periods among the high-amplitude periodic variable stars. Furthermore, there is a large number of known candidates allowing statistical conclusions to be drawn. Only a small fraction of those stars would be overlooked if we take the limit at 17 d^{-1}.

The frequency of the highest peak in the Lomb periodogram is selected as being the main frequency of the signal. The amplitude A of a sinusoidal signal with a standard deviation of σ and the number of measurements, nmes, of it is related to the power P by the relation $A = \sqrt{4\sigma P/\mathrm{nmes}}$.

4 FOURIER SERIES

A Fourier series with n ($\geqslant$2) harmonics is fitted to the data,

$$S(t) = \sum_{i=1}^{n} A_i \sin(2\pi i \nu t) + B_i \cos(2\pi i \nu t) + C,$$

where ν is the frequency, t is the time and C is a constant. The parameters

$$R_{21} = \sqrt{\frac{A_2^2 + B_2^2}{A_1^2 + B_1^2}}$$

and

$$\phi_{21} = \arctan(-A_2/B_2) - 2\arctan(-A_1/B_1)$$

are computed. We linearize the equations with respect to the frequency, and search for the least-squares fit iteratively. The initial frequency is that obtained with the Lomb algorithm.

The determination of the number of harmonics is performed iteratively. Initially, the number of harmonics is fixed at 2, so that R_{21} and ϕ_{21} are always defined. The procedure is then to loop and stop at a maximum of six harmonics. We determine a first solution for a certain number of harmonics, then we increase by one the number of harmonics and recompute a second solution. We perform a Fisher test comparing the two models (Lupton 1993). If there is a significant reduction of the χ^2 by adding a harmonic, we repeat the procedure by adding one more, and if not we keep the first model. The majority of stars (58 per cent) have a solution with two harmonics, then 18, 11, 7 and 6 per cent have solutions with 3, 4, 5 and 6 harmonics, respectively.

4.1 Estimation of the error on the period

The general trend is that the error on the period is a function of the square of the period. However, peculiarities in the light curve, e.g. sharp features such as rising branches in Cepheids, may fix the period more precisely.

The method used for estimating the error on periods is the same as that used for the *Hipparcos* catalogue (ESA 1997) in the case of the Geneva solutions (Eyer 1998). We use the estimation of the error on the frequency given by the least-squares fit of a linearized Fourier series. Schwarzenberg-Czerny (1991) proposed an estimate for the error on the period by taking into account the correlation of the residuals. Such a correction is not implemented here.

5 CLEANING THE SAMPLE

To select a 'well-behaved' data set, we used the following criteria.

(i) Stars with less than 40 measurements were discarded.

(ii) We defined a reduced time-span: the total observational time-span where the three largest gaps are subtracted (we define a gap as the time interval between two consecutive measurements). We retained periods which were smaller than the reduced time-span divided by 1.2. It was found empirically, by visual inspection, that

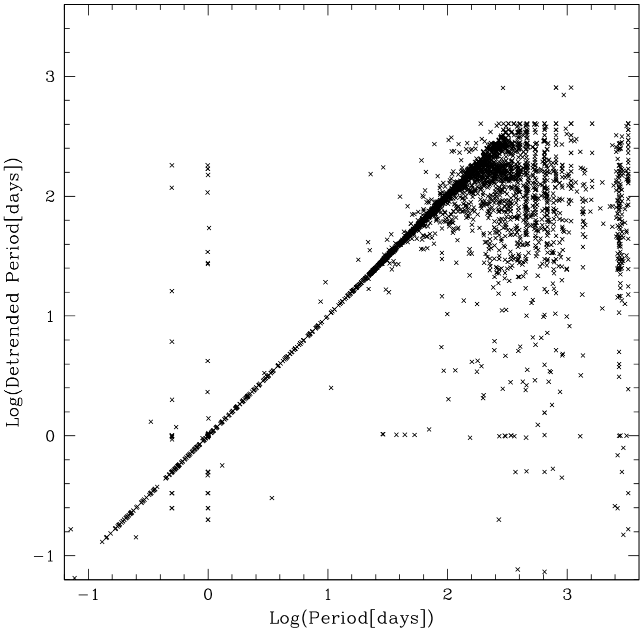

Figure 5. Comparison of the periods, obtained by the Lomb algorithm on the data and when a parabolic fit is subtracted from the data, allows some aliases as well as solutions with less robust period determinations to be identified.

many short frequencies are spurious. Such a criterion may exclude true LPV stars with a poorly covered light curve.

(iii) We rejected objects with a skewness on the I magnitude larger than -1 (rejecting time series containing bright outliers such as flares, cosmic rays).

(iv) We selected objects which had a mean I magnitude smaller than 12.65.

(v) A parabolic curve was fitted to the data, then subtracted from it, and the Lomb algorithm was recomputed on the resulting data under the same conditions as the initial trial described in Section 3. We present the diagram of the initial period and the period found after having subtracted that parabolic fit in Fig. 5. We notice that 1-d and 0.5 d spurious periods are present. We also notice that there are periods which are significantly different after reanalysis. We selected the objects with periods which are not significantly different.

(vi) Periods near the aliases of 1 d and 0.5 d were removed.

These criteria were established mostly empirically after the samples rejected were studied in a detailed manner to avoid rejecting valuable objects.

6 AUTOCLASS

Humans by nature have trouble visualizing multidimensional data sets, especially those with more than three dimensions. More than three attributes are needed to classify the light curves by the proposed Fourier decomposition. Furthermore, the variable star population is diverse. Some classes have very well-defined characteristics, and others have overlapping properties. Some classes are divided almost arbitrarily into subclasses, although they represent a continuum. For example, as defined in the General Catalogue of Variable Stars (Kholopov et al. 1985), Mira stars have a peak-to-peak amplitude in V larger than 2.5 mag. An unsupervised program might give some indications for better divisions. Another advantage of such an unsupervised algorithm is that it can point out new classes of objects.

AUTOCLASS (Cheeseman & Stutz 1996) is a Bayesian classifier. The algorithm looks for the number of classes and the classification which is most probable, given the observed data. The method was successfully applied to several astronomical sets: *IRAS* sources (Goebel et al. 1989), asteroids (Ivezic et al. 2001) and *Hipparcos* data (Eyer, private communication).

The method is not fully automated because there is an interactive part (probability distributions of the parameters have to be specified), but the method takes the data in totality and proposes a broad classification. As pointed out by the authors of AUTOCLASS, this interactivity is necessary.

Another useful aspect of the Bayesian classifier is that it computes a class membership probability. Therefore, this probability can be used to sort according to reliability levels within the classification.

The attributes (i.e. the parameters) chosen for the classification are the period, amplitude, phase difference ϕ_{21} and amplitude ratio R_{21}. With only these four parameters we show that we reach a rather reliable classification. For the irregular variables we used the period, second, third and fourth moments of the light curves as the classification parameters.

As period and amplitude are positive quantities, we choose the logarithm of those values for the classifier.

The phase difference ϕ_{21} is defined as modulo 2π. Circular or angular real valued attributes are not yet available in AUTOCLASS. The eclipsing binaries that constitute a major part of the sample we classify have a period which is generally wrong by a factor of 2 (typical for Fourier-type methods). This means ϕ_{21} is often around zero. Therefore, the eclipsing binaries will be split into two groups (those with a ϕ_{21} above 0 and those below 2π), giving rise to many more groups. For this reason, we redefined the phase difference ϕ_{21} as being between $3\pi/4$ and $(3\pi/4) + 2\pi$. Few were found to have a ϕ_{21} around $3\pi/4$.

There is a relation found between period and amplitude (see the next section). The stars forming this relation have irregular light curves and so the parameters have a large enough dispersion to weaken the abilities of AUTOCLASS to perform the classification. We previously (Eyer & Blake 2002) selected a subsample of these objects manually, to retain only fairly well-behaved objects. Here, we want a more automated process. Instead of finding a method for selecting objects with high residuals, we just divided the sample into two with the relation amplitude $= 10^{-3.2}$ period$^{1.6}$. We then applied AUTOCLASS to these two samples with different attributes.

In our experience, adding parameters often does not improve the classification. So the general method is to start with very few parameters, apply AUTOCLASS and analyse the result of the classification (on a subsample, for example). If well-known classes are not separated, we can add a parameter and iterate the process.

7 RED GIANTS AND PERIOD–AMPLITUDE RELATION

Fig. 6 shows that there is a period–amplitude relation for a very large fraction of stars, as already remarked in Eyer & Blake (2002). We find that about 83 per cent of the variables fall in this broad region (not including Cepheids). If we compare the population of stars observed by *Hipparcos* in that same region of the period–amplitude diagram (see ESA 1997; Koen & Eyer 2002), we find that most stars have spectral types from K giants for the lower-left part of the relation to M giants for the higher-right part. At the small amplitude and short period side of this relation, we find the small amplitude

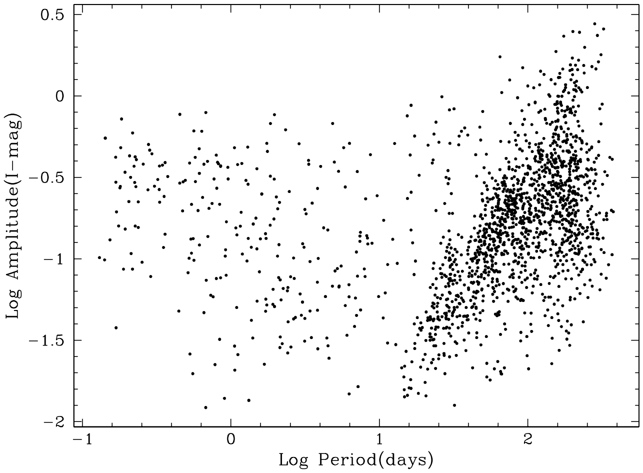

Figure 6. The raw diagram log(period), log(amplitude). The diagram is dominated by red stars, which seem to fall on a relation (or several neighbouring sequences).

red giants studied, for example, by Percy, Wilson & Henry (2001). At the large amplitude and long period side, we find the well-known Mira stars. With the continuation of the ASAS, more stars forming this relation will be found, more data will be available per star, and the morphology of this relation will be described with better precision. For the moment we notice that this relation is most likely composed of two or three parallel sequences.

This period–amplitude relation is also observed in K-band infrared photometry (see van Loon 2002) and r-MACHO band (see Glass & Schultheis 2003). Substructures and parallel relations have also been observed by Minniti et al. (1998) and Wray, Eyer & Paczyński (2004).

8 RESULTS OF THE CLASSIFICATION

A prior work included 458 stars (Eyer & Blake 2002), and now the sample is extended to 1731 stars divided into two groups of 302 stars and of 1429 stars. Thus, 45 per cent of the stars in the sample have a sufficiently regular behaviour to have been selected by our criteria.

For the subsample of 302 stars, the stars are classified into nine groups. The tabled results of the classifcation can be accessed online at http://www.blackwellpublishing.com/products/journals/suppmat/MNR/MNR8651/MNR8651sm.htm. An excerpt of that table is shown in Table 1. Certain groups appear to be very clean and others seem to contain more difficult cases. See Figs 7 and 8 for the result of the classification in a log(period)–log(amplitude) diagram and log(period)–ϕ_{21} diagram or R_{21}diagram, respectively. We have the following groups.

(i) Eclipsing binaries (∼192): one group with (63 stars) eclipsing binaries of EA and EB type. This group has no ambiguity of classification. Another group (36 stars) of EW-type eclipsing binaries includes very few potential pulsating stars (such as δ Scuti stars) or Ap stars with very sinusoidal curves. The third group (38 stars) contains more difficult cases: a large majority are eclipsing binaries, some seem marginal, and others clearly have a wrong period.

(ii) Cepheids ($\sim 19 + 13 = 32$): we can find this type of variable in two different classes. One is very well defined (19 stars). Only one star seems to be peculiar in changing its amplitude. There are about 13 other cases which could be Cepheids – some undoubtedly recognizable to human eyes – while others are difficult to recognize.

(iii) RR Lyrae stars (∼4): probably one δ Scuti and three RR Lyrae of ab type; RR Lyrae of c types will be mixed with eclipsing binaries of EW type.

(iv) LPVs: eight stars are classified in a group with poorly defined light curves, and with periods above 60 d. The time sampling is often sparse and does not cover many cycles of the light curve. The phase coverage often presents gaps.

(v) Small amplitude variables (∼44): many light curves seem to be marginal cases. Very few unambiguous cases could be identified. However, it makes sense to have such a group. There are some αCVn, RS CVn, ellipsoidal variables which could be present in this group. The mean amplitude of this group is of 0.04 I mag.

(vi) The last group (∼58) is also composed of difficult cases but of larger amplitude (mean amplitude is 0.12 I mag) than the previous one. Here the variability is strongly detected. It is worth noting that the formation of this group is a remarkable aspect of the classifier. Instead of spreading those objects among well-defined classes, they are put in a separate group.

In total, from this classification, there are five groups out of nine, which contain clear cases with an error of classification below 7 per cent, two groups, with some mixed classification, one group of small amplitude variables, which can be caused by many different effects, and one group with very difficult cases.

Because of the limiting magnitude of the ASAS, the RR Lyrae stars are too faint to be numerously detected. Indeed, the Sloan Digital Sky Survey (SDSS; Ivezic et al. 2000) shows that the halo RR Lyrae stars are very rare below I mag 13. Therefore, only three to four RR Lyrae stars are found in the sample. The classification algorithm sometimes identifies the RRab type, depending on the precision that we estimated for the period; however, these stars are not forming a stable class. However, it is remarkable that the program can form a new group with such a small number of stars (about 1 per cent of the sample). The classification was found to be sensitive to error parameters, and sometimes group RR Lyrae are lost.

Beltrame & Poretti (2002) found that the star ASAS 112843–5925.7 (HD 304373) is a double-mode Cepheid, the second one detected in our Galaxy, pulsating in the first and second radial overtones. In our study, unfortunately the period search was limited for that star to a frequency interval up to 0.5 d^{-1}, missing the main peak of this short period pulsator (period = 0.92 d). It is classified in the group of stars that is a mix of pulsating and eclipsing.

The second group of 1429 stars, where we use the moments of the distribution, is divided by AUTOCLASS into five groups (see Fig. 9). The classification divides these stars into SARV (∼230), SR ($\sim 1158 = 102 + 484 + 572$) and Mira (∼41).

With these groups it is difficult to determine error levels because the classification is extremely difficult to establish. There are some erroneous cases, such as the star ASAS 180057–2333.8, which is in the group of Mira stars but is clearly a long-period eclipsing binary.

The catalogue, the light curves and folded curves (see Fig. 10 for a sample), on an individual basis and on a class basis, can be seen at the website of LE (http://www.astro.princeton.edu/∼leyer/ASAS/).

The eclipsing binaries have recomputed solutions where the initial period is doubled because the Lomb–Scargle algorithm usually gives half of the true period. The star 144245–0039.9 even has a factor of 4 between the true period and the period found by the Lomb algorithm.

Table 1. Results of the classification (extract – the full catalogue is available at http://www.blackwellpublishing.com/products/journals/suppmat/MNR/MNR8651/MNR8651sm.htm). The columns are the ASAS ID (equatorial coordinates in equinox 2000), mean I magnitude $\bar{I}$, standard error σ_I, the number of measurements N, the period, the amplitude, the amplitude ratio R21, the phase difference ϕ_{21}, the number of harmonics nh, the class cl, and the probability of membership Prob.

ASAS ID	$\bar{I}$	σ	N	Period	Ampl.	$R21$	ϕ_{21}	nh	cl	Prob
005759+0034.7	10.444	0.028	1204	0.7980	0.076	0.396	6.242	6	3	0.42
015647−0021.2	11.041	0.047	417	0.5427	0.108	0.102	7.996	2	1	0.43
030201−0027.2	10.311	0.049	1133	3.1062	0.118	0.052	5.505	2	1	0.47
034803−0023.5	10.793	0.032	1314	7.0156	0.045	0.187	6.816	2	2	0.47
044830+0017.9	12.205	0.124	1457	0.2250	0.294	0.307	6.374	3	4	0.49
044944+0056.0	11.447	0.086	862	0.7116	0.204	0.028	6.557	2	1	0.50
045017+0100.7	11.379	0.091	503	0.2056	0.224	0.288	6.294	2	4	0.50
045024+0013.2	9.913	0.015	1530	6.2548	0.015	0.040	3.738	2	2	0.51
045128−0032.7	7.596	0.013	1534	0.7818	0.022	0.037	3.143	2	2	0.51
045206−7043.9	10.542	0.086	312	1.1724	0.193	0.758	6.266	4	0	0.52
045423−7054.1	11.773	0.175	423	34.4540	0.566	0.385	5.068	5	6	0.53
045506−6728.5	12.630	0.194	267	29.8318	0.511	0.366	4.665	3	6	0.53
045511−0101.7	10.247	0.023	792	3.1366	0.031	0.128	7.689	2	2	0.54
045702−6759.7	12.065	0.123	517	45.1273	0.322	0.279	5.124	2	6	0.55
045712−6723.2	12.162	0.147	273	22.7070	0.387	0.293	4.809	4	5	0.55
045720−8023.0	11.615	0.228	3448	0.1835	0.721	0.368	6.321	6	4	0.56
045728−7033.1	12.053	0.113	676	0.8249	0.383	0.866	6.301	5	0	0.56
045750−6957.4	12.354	0.147	309	23.3178	0.583	0.331	5.039	5	6	0.57
045810−6957.0	11.839	0.197	637	39.3889	0.575	0.385	5.162	5	6	0.57
045817−0013.9	10.942	0.047	1296	0.2522	0.095	0.083	6.223	2	4	0.58
045832−7020.8	11.970	0.177	621	35.6997	0.607	0.435	5.332	5	6	0.59
045836−7006.6	12.591	0.183	295	17.2697	0.567	0.313	5.349	2	5	0.59
045914−6935.7	11.061	0.100	700	0.3289	0.266	0.109	5.087	3	1	0.59
045941−6927.4	12.541	0.238	596	31.8223	0.833	0.463	5.099	4	6	0.60

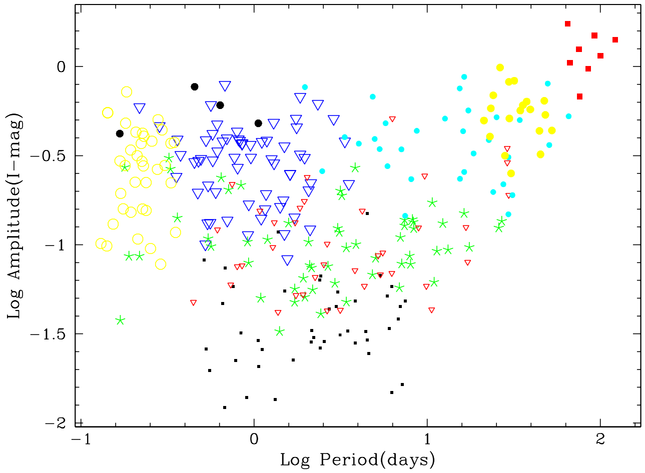

Figure 7. The classification obtained in the diagram log(period), log(amplitude). Red giant stars are excluded from the classification. The Fourier decomposition is used. The colour description is for the on-line version. The symbols are as follows: eclipsing binaries, large blue open triangle, open yellow circles and red small open triangle; RR Lyrae, black large full circles; Cepheids, yellow small full circles and blue small circles; LPVs, red full large squares; small amplitude variables, small black squares; uncertain cases, green five-branch star.

9 DISCUSSION

Pojmański (2000) produced a list of 3900 stars and extracted 400 periodic variables. From our procedures, 1700 stars were extracted in an automated way and then classified. This subsample was needed in order to have a set of sufficiently well-behaved variable stars.

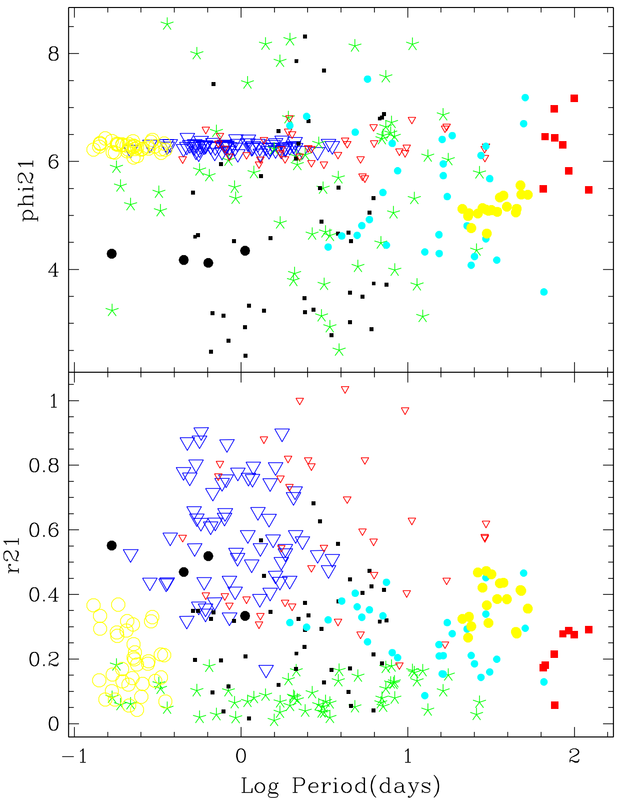

Figure 8. The classification obtained in the diagram log(period), ϕ_{21} and log(period), R_{21}. Symbols as in Fig. 7.

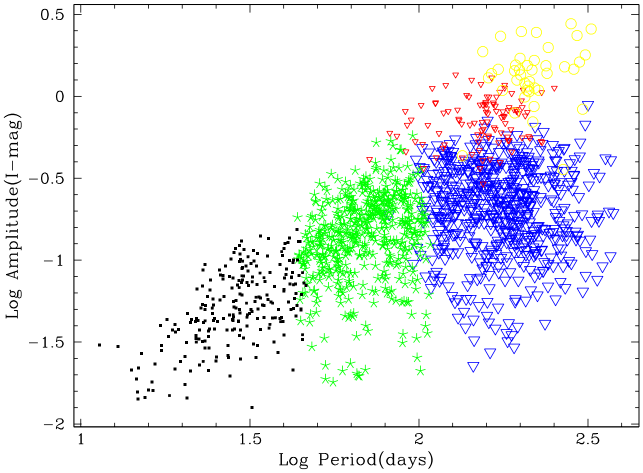

Figure 9. The classification obtained in the diagram log(period), log(amplitude) for the red giants. The moments of the distribution are used. The classifier forms five groups which can be identified to three general known variability types: small-amplitude red variables (small black squares), semiregular variable stars (five-branch star, open triangles) and Mira stars (open circles).

Very irregular objects, objects with too few data points, too small variability and artefacts were removed with automated procedures. So 45 per cent of the stars are classified. Obviously there are open questions. What do we do with the remaining stars? What other parameters can be used? What are other possible procedures? We point out that, up to now, there have been very few attempts of published global automated classification. This situation will probably change in the near future because of the growing importance of surveys.

We note that in our procedure the amount of human interaction is fairly small. It is mostly related to checking the procedures, tuning some parameters, or controlling the classification. The amount of work is increasing only slowly as the number of stars to classify increases.

Other studies using different methods are awaited, for example machine learning algorithms from Woźniak et al. (2001); see also a published result with self-organizing maps by Brett, West & Wheatley (2004). It would be important to compare the efficiency of the methods, their capability to detect new classes, their error levels, and their CPU consumption. It might be interesting to apply such a Bayesian classifier to the data of ASAS 3 and compare it with the results of Pojmański (2002, 2004) and Pojmański & Maciejewksi

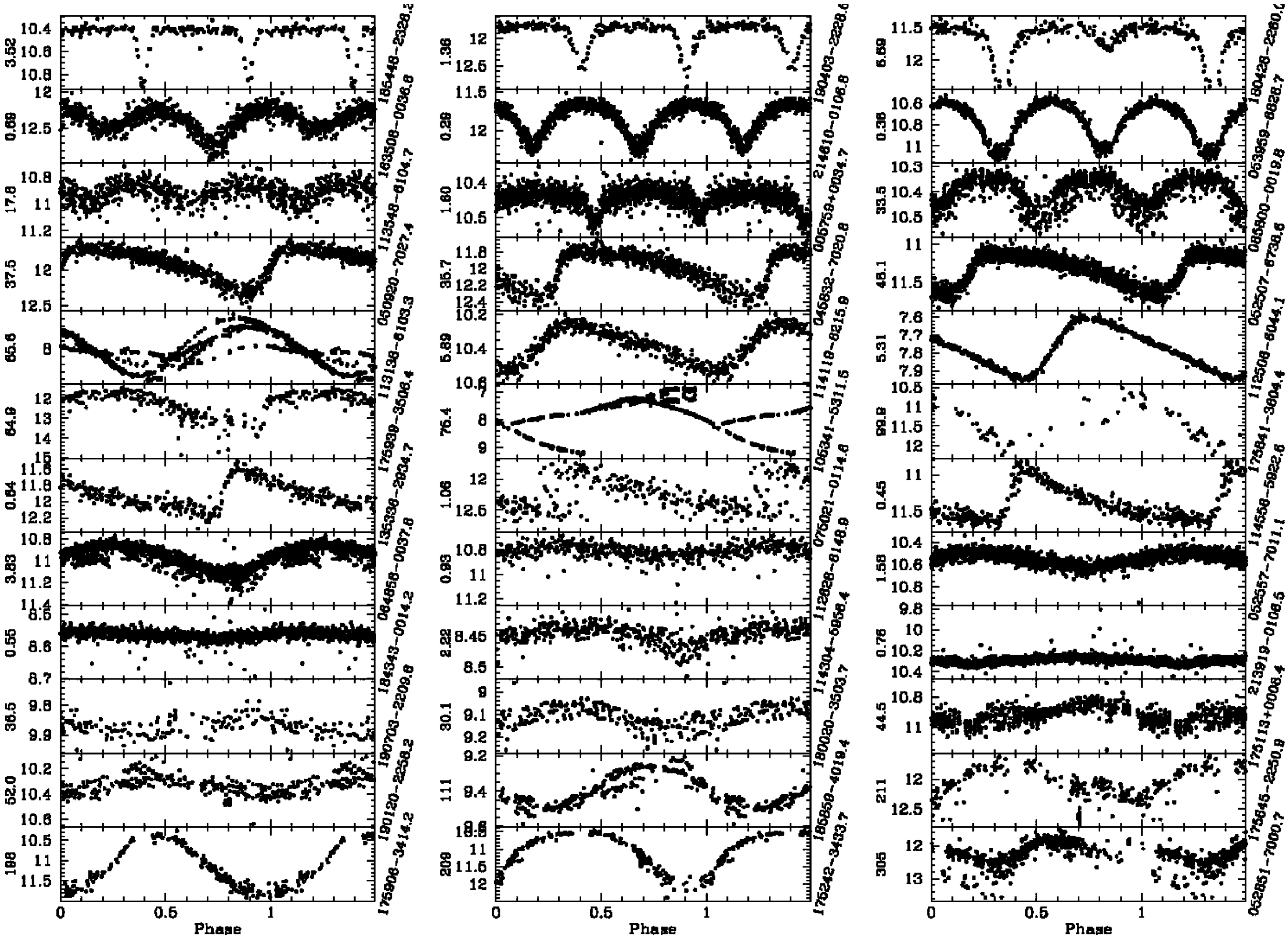

Figure 10. Three examples (on one line) of each of the 12 major classes are represented. We have, from top bottom: eclipsing binaries (∼EA, EB types); eclipsing binaries (∼EW types); eclipsing binaries (more marginal cases); Cepheids; Cepheids (with more marginal cases); LPVs; RR Lyrae candidates; various case class; small amplitude variables; SARVs; SRs; Miras. On the right of the folded curve is the ASAS coordinate (equinox 2000), and on the left is the period in d.

(2004). The domain of global automated classification is still in an exploratory era.

Real time detection and classification of phenomena such as supernovae will require additional software development to be scientifically valuable. For example, OGLE has possessed an early warning system (EWS) since 1994, and received further development in 2003 to include the detection of the effect of planets in a microlensing event. It is envisioned that the OGLE data will be put into the public domain within 24 h from data acquisition. If so, OGLE-III opens the possibility for anyone to make quasi-real time detection of variable phenomenon.

The ASAS also has an alert service. The photometric reduction pipeline is available in real time within 5 min. The current service is focused on the monitoring of cataclysmic variable stars.

10 CONCLUSION

We have developed a scheme for general and automated classification for the periodic variable stars of the ASAS 1–2 data set. Of course, every survey has its own properties, so even if the approach is transferable in its principles, it probably requires modifications for every data set. However, the general method can easily be applied to larger data bases.

The work was broken into three parts: (i) selection of periodic objects; (ii) modelling that subsample with Fourier series or with simpler parameters such as moments of the distribution; (iii) application of the AUTOCLASS Bayesian classifier. However, at every step it is critical to check the quality of the analysis, to interact with the data and to visualize data or the defined parameters. Work is also needed finally to analyse the output.

At present, and in the very near future, there are several good opportunities to apply such a classification method to other data sets, as follows.

(i) The ASAS continues to survey the sky in the *V* and *I* bands.

(ii) The Hungarian Automated Telescope (HAT) project (Bakos et al. 2002) surveys some specific regions of the sky in the *I* band. The HAT has released data about 1700 suspected variable stars (Hartman et al. 2004)

(iii) The Magellanic Clouds data of OGLE-II (Żebruń et al. 2001) are available (58 000 variable stars), as well the 49 bulge fields from OGLE-II (Woźniak et al. 2002). The classification of bulge field 1 has been carried out by Mizerski & Bejger (2002); specific extractions of eclipsing binaries, RR Lyrae and Cepheids stars have been accomplished for the Magellanic clouds.

(iv) The third phase of OGLE, OGLE-III, is functional and is taking data. The data rate is multiplied by a factor of 10 with respect to OGLE-II.

Similar software has to be developed for large surveys from the ground and from space, such as the GAIA mission (ESA 2000), so it is important to gain knowledge of different classification methods.

ACKNOWLEDGMENTS

We are grateful for the help, comments and encouragement from B. Paczyński, C. Alard, A. Gautschy, M. Grenon, Z. Ivezic and G. Pojmański. Our thanks go to M. C. S. Peterson for English corrections, and to W. O'Mullan and A. Dartois for their help in JAVA programming. The software SM (Lupton & Monger 1997) was widely used for this research. Partial support for this project was provided by the Swiss National Science Foundation, the National Science Foundation grant AST-0204908 and the Carnegie Institution of Washington, Department of Terrestrial Magnetism.

REFERENCES

Bakos G. Á., Lázár J., Papp I., Sári P., Green E. M., 2002, PASP, 114, 974

Belokurov V., Evans N. W., Du Y. L., 2003, MNRAS, 341, 1373

Beltrame M., Poretti E., 2002, A&A, 386, L9

Brett D. R., West R. G., Wheatley P. J., 2004, MNRAS, 353, 369

Cheeseman P., Stutz J., 1996, in Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., eds, Advances in Knowledge Discovery and Data Mining, Bayesian classification (AUTOCLASS): theory and results. AAAI Press/MIT Press, Cambridge, MA, p. 61

ESA, 1997, ESA SP-1200, The *Hipparcos* and Tycho Catalogues. ESA Publications Divison, Noordwijk

ESA, 2000, ESA-SCI(2000)4, GAIA (Concept and Technology Study Report). ESA Publications Division, Noordwijk

Eyer L., 1998, PhD thesis, Univ. Geneva

Eyer L., Bartholdi P., 1999, A&ASS, 135, 1

Eyer L., Blake C., 2002, in Aerts C., Bedding T., Christensen-Dalsgaard J., eds, ASP Conf. Ser. Vol. 259, Radial and Nonradial Pulsations as Probes of Stellar Physics. Astron. Soc. Pac., San Francisco, p. 160

Eyer L., Grenon M., 1997, ESA SP-402, 467

Eyer L., Genton M., 1999, A&ASS, 136, 421

Glass I. S., Schultheis M., 2003, MNRAS, 345, 39

Goebel J., Stutz J., Volk K., Walker H., Gerbault F., Self M., Taylor W., Cheeseman P., 1989, A&A, 222, L5

Hartman J. D., Bakos G., Stanek K., Noyes R. W., 2004, AJ, 128, 1761

Ivezic Z. et al., 2000, AJ, 120, 963

Ivezic Z. et al., 2001, AJ, 122, 2749

Kholopov P. N. et al., 1985, General Catalogue of Variable Stars. Nauka, Moscow

Koen C., Eyer L., 2002, MNRAS, 331, 45

Lomb N. R., 1976, Ap&SS, 39, 447

Lupton R. H., 1993, Statistics in Theory and Practice. Princeton Univ. Press Princeton

Lupton R. H., Monger P., 1997, The SM Reference Manual (http://www.astro.princeton.edu/~rhl/sm/sm.html)

Minniti D. et al., 1998, in Takeuti M., Sasselov D. D., eds, Pulsating Stars: Recent Developments in Theory and Observation. Universal Academy Press, Tokyo, p. 5

Mizerski T., Bejger M., 2002, Acta Astron., 52, 61

Paczyński B., 1997, in Ferlet R., Maillard J.-P., Raban B., eds, Variables Stars and the Astrophysical Returns of the Microlensing Surveys. Editions Frontières, Gif-sur-Yvette, p. 35

Paczyński B., 2000, PASP, 112, 1281

Paczyński B., 2001, in Banday A. J., Zaroubi S., Bartelmann M., eds, Mining the Sky. Springer-Verlag, Berlin, p. 481

Percy J. R., Wilson J. B., Henry G. W., 2001, PASP, 113, 983

Piquard S., Halbwachs J.-L., Fabricius C., Geckeler R., Soubiran C., Wicenec A., 2001, A&A, 373, 576

Pojmański G., 1997, Acta Astron., 47, 467

Pojmański G., 2000, Acta Astron., 50, 177

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

Pojmański G., 2003, Acta Astron., 53, 341

Pojmański G., 2004, Astron. Nachr., 325, 553

Pojmański G., Maciejewksi G., 2004, Acta Astron., 54, 153

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in Fortran. Cambridge Univ. Press, Cambridge

Rodriguez E., López-González M. J., López de Coca P., 2000, A&ASS, 144, 469

Ruciński S. M., 1993, PASP, 105, 1433

Ruciński S. M., 1997, AJ, 113, 1112

Schwarzenberg-Czerny A., 1991, MNRAS, 253, 198

van Loon J. Th., 2002, in Aerts C., Bedding T., Christensen-Dalsgaard J., eds, ASP Conf. Ser. Vol. 259, Radial and Nonradial Pulsations as Probes of Stellar Physics. Astron. Soc. Pac., San Francisco, p. 458
Woźniak P. et al., 2001, BAAS, 33, 1495
Woźniak P., Udalski A., Szymański M., Kubiak M., Pietrzyński G., Soszyński I., Żebruń K., 2002, Acta Astron., 52, 129
Woźniak P. et al., 2004, AJ, 127, 2436
Wray J. J., Eyer L., Paczyński B., 2004, MNRAS, 349, 1059
Żebruń K. et al., 2001, Acta Astron., 51, 317

This paper has been typeset from a TEX/LATEX file prepared by the author.