Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **414,** 2602–2617 (2011) doi:10.1111/j.1365-2966.2011.18575.x

Random forest automated supervised classification of *Hipparcos* periodic variable stars

P. Dubath,[1,2]★ L. Rimoldini,[1,2] M. Süveges,[1,2] J. Blomme,[3] M. López,[4] L. M. Sarro,[5] J. De Ridder,[3] J. Cuypers,[6] L. Guy,[1,2] I. Lecoeur,[1,2] K. Nienartowicz,[1,2] A. Jan,[1,2] M. Beck,[1,2] N. Mowlavi,[1,2] P. De Cat,[6] T. Lebzelter[7] and L. Eyer[1,2]

[1]*Observatoire astronomique de l'Université de Genève, ch. des Maillettes 51, 1290 Versoix, Switzerland*
[2]*ISDC Data Center For Astronphysics, ch. d'Ecogia 16, 1290 Versoix, Switzerland*
[3]*Instituut voor Sterrenkunde, K. U. Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium*
[4]*Centro de Astrobiología (INTA-CSIC), Departamento de Astrofísica, PO Box 78, E-28691 Villanueva de la Cañada, Spain*
[5]*Dpt. de Inteligencia Artificial, UNED, Juan del Rosal, 16, 28040 Madrid, Spain*
[6]*Royal Observatory of Belgium, Ringlaan 3, 1180 Brussels, Belgium*
[7]*University of Vienna, Department of Astronomy, Türkenschanzstrasse 17, A1180 Vienna, Austria*

Accepted 2011 February 21. Received 2011 February 18; in original form 2011 January 12

ABSTRACT

We present an evaluation of the performance of an automated classification of the *Hipparcos* periodic variable stars into 26 types. The sub-sample with the most reliable variability types available in the literature is used to train supervised algorithms to characterize the type dependencies on a number of attributes. The most useful attributes evaluated with the random forest methodology include, in decreasing order of importance, the period, the amplitude, the $V - I$ colour index, the absolute magnitude, the residual around the folded light-curve model, the magnitude distribution skewness and the amplitude of the second harmonic of the Fourier series model relative to that of the fundamental frequency. Random forests and a multi-stage scheme involving Bayesian network and Gaussian mixture methods lead to statistically equivalent results. In standard 10-fold cross-validation (CV) experiments, the rate of correct classification is between 90 and 100 per cent, depending on the variability type. The main mis-classification cases, up to a rate of about 10 per cent, arise due to confusion between *SPB* and *ACV* blue variables and between eclipsing binaries, ellipsoidal variables and other variability types. Our training set and the predicted types for the other *Hipparcos* periodic stars are available online.

Key words: methods: data analysis – methods: statistical – techniques: photometric – catalogues – stars: variables: general.

1 INTRODUCTION

The development of efficient automated classification schemes is becoming of prime importance in astronomy. Large surveys are monitoring millions, and soon billions, of targets. The resulting time series cannot possibly be scrutinized by eye. The identification and study of variable stars require the use of powerful statistical and data mining tools. Automated supervised classification methods provide object type predictions based on the values of a set of attributes characterizing the objects. In the first stage, a collection of prototype objects of known type and attribute values, referred to as the training set, is used to build a model of the dependencies of the types on the attribute values. In the second stage, this model is used to predict the types of other objects of unknown types but with available attribute values. As the naming schemes differ in different publications, we make the following definitions to be used throughout this paper: *objects* (i.e. *stars*) are classified into *types* making use of a number of *attributes*.

Tree-based classification methods are simple to use and popular in applications (see e.g. Hastie, Tibshirani & Friedman 2009). They can deal with complex structures in the attribute space and have low systematic classification errors if the trees are sufficiently deep. However trees are noisy and the resulting type estimator has large variance. With the random forest method (Breiman 2001) the variance is drastically reduced by averaging the results of many trees built from randomly selected subsamples of the training set (bootstrapping). In addition, this method uses the best one of a number of

★E-mail: pierre.dubath@unige.ch

randomly selected attributes at each branching, which has the effect of reducing the correlation between different trees and hence improving the averaging. The random method has been used in many scientific domains such as bioinformatics, biology, pharmacy and Earth sciences, confirming that it performs excellently on a broad range of classification problems. By construction it is relatively robust against overfitting; it is only weakly sensitive to choices of tuning parameters; it can handle a large number of attributes; it provides an unbiased estimate of the generalization error and the importance of each attribute can be estimated.

A number of variable star classification studies exploit recent or ongoing surveys, such as (1) ASAS (Pojmanski 2002, 2003; Eyer & Blake 2002, 2005), (2) OGLE (Sarro et al. 2009), (3) MACHO (Belokurov, Evans & Du 2003; Belokurov, Evans & Le Du 2004). (4) CoRoT (Debosscher et al. 2009), (5) Kepler (Blomme et al. 2010). A number of ambitious survey projects are also in an advanced stage of preparation, in particular (1) Pan-STARRS,[1] (2) LSST[2] and (3) *Gaia*.[3] Although these projects have different primary goals, the expected time series of measurements will provide data of unprecedented quality to study variable stars. These data are critical to both better characterize populations of variable stars in different environments and investigate in more depth typical or peculiar individual cases. Progress in this field not only impacts our understanding of variable star physics, but also leads to contributions to a wide range of astronomical topics, from stellar evolution and population-synthesis modelling to distance scale related issues.

The *Hipparcos* mission stands out as a rather original and ambitious astrometry space programme. Because of the whole sky repeated scanning, it provides accurate data for all the brightest stars in our close neighbourhood. The *Hipparcos* periodic star catalogueincludes most of the best studied stars and hence provides a unique set that can be used as a 'control sample'. Results obtained for these stars can be validated using the wealth of available published information. This is particularly useful for evaluating variable star classification methods before applying them to other large surveys. The variability types resulting from the classification can be compared with types available from the literature. The *Hipparcos* sample also includes almost all types of variable stars present in the solar vicinity. It is certainly a solid basis for building a training sample for supervised classification methods.

The comprehensive study of *Hipparcos* variable stars presented in volume 11 of the *Hipparcos* periodic star catalogue[4] (Eyer 1998) does not comprise a systematic automated classification. The variability types provided in this catalogue are extracted from the literature with two exceptions. First, the eclipsing binaries and the RV Tauri were identified and characterized on the basis of visual inspections of the folded light curves. Secondly, a systematic classification of variables with B spectral type was achieved by Waelkens et al. (1998) using a multivariate discriminant analysis. Later, Aerts, Eyer & Kestens (1998) used a similar technique to isolate variables with A2 to F8 spectral types. An attempt to obtain a systematic classification is also presented by Willemsen & Eyer (2007).

The subject of this paper is the development of the first systematic, fully automated classification of the complete sample of *Hipparcos* periodic variable stars. The performance of the random forest method is evaluated and the results are compared with those obtained from the multistage classifier developed recently (Blomme et al. 2011). In a companion paper (Rimoldini et al. in preparation), this work is extended to include non-periodic variables, both to study the classification of non-periodic stars and to evaluate the confusion between periodic and non-periodic types. The main goal is to fully validate our approach on a controlled sample of *Hipparcos* stars, before applying it in a more automated fashion on other surveys in future studies. An important outcome of our work is a homogeneous supervised classification training set, which can be adapted to other missions, in particular, the upcoming *Gaia* mission. Predicted types are also provided for almost all *Hipparcos* periodic variables, including some that have no types or only uncertain ones in the literature.

The procedure followed to build the training set from a subsample of the best-known *Hipparcos* stars is described in Section 2. The subject of Section 3 is the determination of the attribute values, including the period search. Section 4 describes the applied random forest methodology and shows the corresponding results. Section 5 presents an investigation of the influence of the period value errors on the classification process. Results from random forest results and a multi-stage classifier are compared in Section 6, while our best final predicted types are listed in Section 7. Finally, our conclusions are given in the last section (Section 8).

2 TRAINING SET COMPOSITION

Data from a set of objects of known types are needed to train the supervised classification algorithms. The quality of the classification directly rests on the reliability of the types of the stars included in this set, called the training set. For a given type, the selected objects should only include true representatives of the group with typical properties. In our case, we select a sub-set of our *Hipparcos* stars with most reliable types available from the literature, taking advantage of the fact that many of them are relatively bright, well-studied objects. Ideally, the relative frequencies of the different types in the training set should be representative of those in the population to-be-classified.

A search for periods in the *Hipparcos* data alone is inconclusive for 171 of the 2712 stars included in the *Hipparcos* periodic star cataloguedue to the incomplete phase coverage of the light curves. The period values published in the *Hipparcos* periodic star cataloguecome from the literature for these stars. Most of them are eclipsing binaries (152 EAs) with too few *Hipparcos* measurements during the eclipses. These stars are excluded from the training set as the scope of this paper is restricted to periodic stars with a light curve from which it is possible, at least in principle, to infer a period.

The variability types provided in the *Hipparcos* periodic star cataloguewere mainly extracted from the literature (to the notable exception of eclipsing binaries; see Section 3.2.2) available at the time of publication (1997). These types are revised for our study using more recent information. The main reference is the International Variable Star Index (Watson, Henden & Price 2010) catalogue from the American Association of Variable Star Observers (*AAVSO* catalogue hereafter). This index includes information from the General Catalogue of Variable Stars (GCVS) and the New Catalogue of Suspected Variables (NSV) and it is kept up-to-date with the literature with two releases per month. The release adopted herein is that from 2010 June 13. Information from private communications is preferentially used for some specific variability types (see below) as it is believed to be more reliable.

[1] http://pan-starrs.ifa.hawaii.edu/public
[2] http://www.lsst.org/lsst
[3] http://www.rssd.esa.int/Gaia
[4] http://www.rssd.esa.int/index.php?project=HIPPARCOS&page=Overview

Table 1. Training set composition.

Type		Num	Main reference
Eclipsing binary	*EA*	228	*Hipparcos*
	EB	255	*Hipparcos*
	EW	107	*Hipparcos*
Ellipsoidal	*ELL*	27	*Hipparcos*
Long period variable	*LPV*	285	Lebzelter (p. c.)
RV Tauri	*RV*	5	AAVSO
W Virginis	*CWA*	9	AAVSO
	CWB	6	AAVSO
Delta Cepheid	*DCEP*	189	AAVSO
(first overtone)	*DCEPS*	31	AAVSO
(multi-mode)	*CEP(B)*	11	AAVSO
RR Lyrae	*RRAB*	72	AAVSO
	RRC	20	AAVSO
Gamma Doradus	*GDOR*	27	De Cat (p. c.)
Delta Scuti	*DSCT*	43	AAVSO
(low amplitude)	*DSCTC*	81	AAVSO
SX Phoenicis	*SXPHE*	4	AAVSO
Beta Cephei	*BCEP*	30	De Cat (p. c.)
Slowly Pulsating B star	*SPB*	81	De Cat (p. c.)
B emission-line star	*BE*	9	AAVSO
Gamma Cassiopeiae	*GCAS*	4	AAVSO
Alpha Cygni	*ACYG*	18	AAVSO
Alpha-2 Canum Venaticorum	*ACV*	77	Romanyuk (p. c.)
SX Arietis	*SXARI*	7	Romanyuk (p. c.)
BY Draconis	*BY*	5	Eker et al. (2008)
RS Canum Venaticorum	*RS*	30	Eker et al. (2008)
	Total:	1661	

AAVSO : Watson et al. (2010).

The type-assignment process for *Hipparcos* variables is as follows (see Table 1 for type acronym definitions).

(1) For eclipsing binaries and ellipsoidal variables, the *Hipparcos* periodic star catalogueis taken as the reference as (1) the classification done at the time included reliable visual checking of the *Hipparcos* light curves and (2) these light curves are known to have a good enough eclipse coverage to allow a successful period determination (see above and Section 3.2.2).

(2) Lists of *Hipparcos* stars of the types *GDOR*, *SPB* and *BCEP* are provided by P. De Cat and of *LPV* by T. Lebzelter; both maintain up-to-date compilations of literature information for these types.

(3) The type determination for *ACV* and *SXARI* stars that have a measured magnetic field is considered as particularly reliable. As a consequence, for these types, only the *Hipparcos* stars included in a list of magnetic stars provided by I.I. Romanyuk (private communication) are retained.

(4) Only the subset of *Hipparcos RS* and *BY* stars listed in the third edition of the 'catalogue of chromospherically active binary stars' (Eker et al. 2008) is included.

(5) All stars from the *AAVSO* catalogue with a type matching any of the above mentioned types are excluded. For example, a star identified as Mira, SR, LB or SARV in *AAVSO* catalogue that is not in the Lebzelter list of *LPV* is discarded from the training set. The *AAVSO* catalogue is then used to assign a type to the remaining stars from the *Hipparcos* periodic star catalogue. This procedure leads to a subset of 1963 stars.

(6) A visual inspection of the folded light curves of all 1963 stars leads to the elimination of 64 stars due to either poor sampling or excessive noise.

(7) Types with less than three representatives are discarded. This concerns seven stars of six different types: FKCOM, HADS, INSA, INSB, nra, CW-FU in Watson et al. (2010).

(8) Finally, 92 stars with uncertain type (denoted with a colon in the original sources) are also excluded.

The above type-assignment process leads to a training set with 1800 stars of 26 different types. A further 32 stars are excluded due to diverse difficulties in the light-curve processing (see Section 3) and another 107 are discarded because of missing colour indices in the *Hipparcos* periodic star catalogue. This leaves a training set of 1661 stars. Table 1 summaries the final composition of the training set as a function of the variability type.

It is important to note that combined types, such as an intrinsic variable included in an eclipsing binary, are excluded from our training set as a result of the above-type assignment process.

3 CLASSIFICATION ATTRIBUTES

The classification experiments presented in this paper rely on a number of attributes. To achieve the most accurate classification, attributes should be chosen so as to characterize the stars as thoroughly as possible. Some attributes reflect stellar global properties, such as the mean colour or the absolute brightness whereas others describe features of the light curve. The shape of the folded light curve is one of the key indicators of variability type.

The strategy used in this paper is to compute a large number of attributes and to use some algorithms to estimate their merits. In this section, we describe the principle of the attribute derivation. The attribute ranking and selection is described in Section 4, which is devoted to classification. The exact attribute definition is also deferred to avoid describing some that may finally not be retained.

3.1 Statistical parameters

A number of statistical parameters are derived from the distribution of photometric measurements. The list includes the distribution moments (mean, standard deviation, skewness and kurtosis), the range and percentiles. Weighted and un-weighted formulations are used as well as robust estimators.

3.2 Period search

The period values provided in the *Hipparcos* periodic star catalogue are particularly reliable as the corresponding folded light curves were all visually checked prior to publication. Using directly these values in our analysis would lead to optimum results. However, current and upcoming surveys can include millions of variables which cannot all be visually checked. One of the main goals of this paper is to investigate what can be achieved through an *automated* classification process, including the period search.[5] An investigation of the increase of the classification errors resulting from the use of an incorrect period is presented in Section 5.

A number of well-known period search methods such as Deeming (1975), Lomb–Scargle (Lomb 1976; Scargle 1982), harmonic least-squares analysis of generalized Lomb–Scargle methods (Zechmeister & Kürster 2009), String Length methods (Lafler &

[5] The problem of spurious and aliased periods (or frequencies) typically showing up at fractions and multiples of 1 day and 1 yr in ground-based surveys does not affect the period search in *Hipparcos* data.

Kinman 1965; Burke, Rolland & Boy 1970; Renson 1978; Dworetsky 1983) and Jurkevich–Stellingwerf (Jurkevich 1971; Stellingwerf 1978) are employed to search for periodicity in the light curves. The resulting periods are compared with the *Hipparcos* periodic star catalogue values to derive the fraction of correct results. Extensive testing shows that a single method can lead to a recovery fraction of around 80 per cent, while an ideal combination of all methods could potentially raise that value to close to 100 per cent (Cuypers in preparation). Unfortunately, no automated strategy is found to predict which method leads to the correct period for a specific light curve. The best overall recovery fraction is obtained with a combination of the classic Lomb–Scargle method (Lomb 1976; Scargle 1982) and of a generalized Lomb–Scargle variant (Zechmeister & Kürster 2009). The former is used for all stars with a large magnitude distribution skewness while the latter is used for all remaining stars.

The period search results obtained for eclipsing binaries and for the other types of variables are different. They are presented separately in the next two sections.

3.2.1 *Non-eclipsing variable periods*

Fig. 1 displays the results of the period search for all variables excluding the eclipsing binaries and the ellipsoidals using the classic or generalized Lomb–Scargle methods (see Section 3.2) depending on the skewness value. Out of the 1044 non-eclipsing variables, a good period estimate is derived for 951 stars, i.e. a correct period recovery rate of 91 per cent. The period is considered as good if the difference between the extracted period and the *Hipparcos* catalogue value does not lead to a cumulative shift in phase of more than 20 per cent over the full time-span of the light curve.

3.2.2 *Eclipsing variable periods*

The most common type of photometric variability is due simply to binary stars eclipsing each other. This represents a real challenge for classification because almost all kinds of stars can form a binary system. For example, the colour index, usually a powerful indicator of stellar type, is no longer a useful discriminant as it is a mean from two stars that can have almost any kind of stellar colours. In addition, one of the stars, or even both, can exhibit other types of variability leading to a wide range of combined behaviours. Close interaction can also trigger other types of variability such as the RS Canum Venaticorum phenomenon.

There is an important complication in deriving the periods of eclipsing binaries. The folded light curves (or the pulse profile) of non-eclipsing periodic variable stars exhibit usually a single excursion per cycle going through a unique minimum and maximum. As two eclipses are often observed over one binary system revolution, the resulting light curves exhibit two minima over one cycle. In significant number of cases, the two minima have almost the same depth and width. The light curves exhibit two almost identical excursions and, consequently, the period search usually returns half of the true period. Fig. 2 shows examples of actual *Hipparcos* light curves to illustrate the difficulty in extracting the correct period for eclipsing binary systems.

In the preparation of the *Hipparcos* periodic star catalogue, the light curves of all the 2712 periodic variables were visually inspected. Making sometimes use of additional information from the literature, the eclipsing binaries were identified and when necessary the period doubled. Introducing these period values into a general

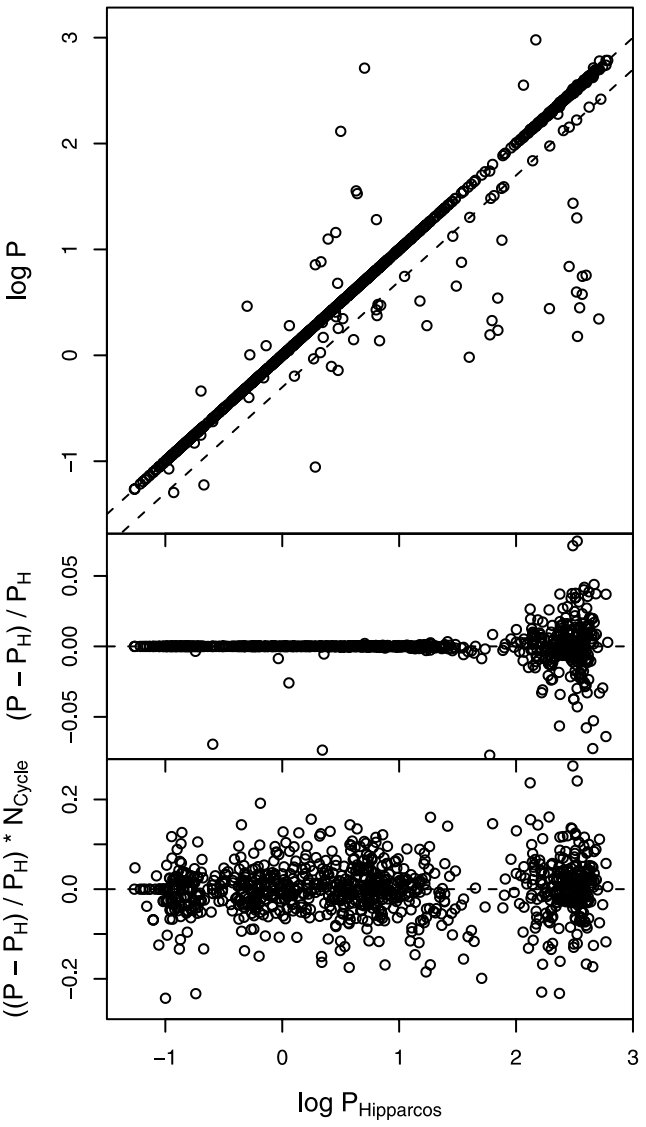

Figure 1. Periods extracted using the Lomb–Scargle method for non-eclipsing variables as a function of the *Hipparcos* periodic star catalogue periods $P_{\rm H}$ expressed in days on a decadic log scale. The upper plot shows the Lomb–Scargle period (P), the middle one the relative difference, and the lower one the relative difference multiplied by $N_{\rm Cycle}$. $N_{\rm Cycle}$ is the light-curve span divided by the period value, i.e. the number of cycles between the beginning and the end of the light curve. The lower diagram shows the cumulative shift, expressed in units of phase, after $N_{\rm Cycle}$ resulting from the inaccuracy of the period value. The dashed diagonal lines in the upper plot display the relationships $P = P_{\rm H}$, and $P = 0.5\,P_{\rm H}$. The middle and lower plots have much enlarged y-scales so that almost all outliers visibly scattered in the upper plot fall outside of the displayed ranges.

classification algorithm is not appropriate. Because of the double-excursion behaviour of their light curves they could be very easily separated from the other variables. But this success would be an illusion as it simply reflects the fact that these eclipsing binaries were carefully identified and their period confirmed through a thorough visual check in the first place.

Fig. 3 displays the results of the period search for the 617 eclipsing binaries (EA, EB and EW) and ellipsoidal variables using the classic or generalized Lomb–Scargle methods (see Section 3.2) depending on the skewness value. Out of the 617 variables, approximately half

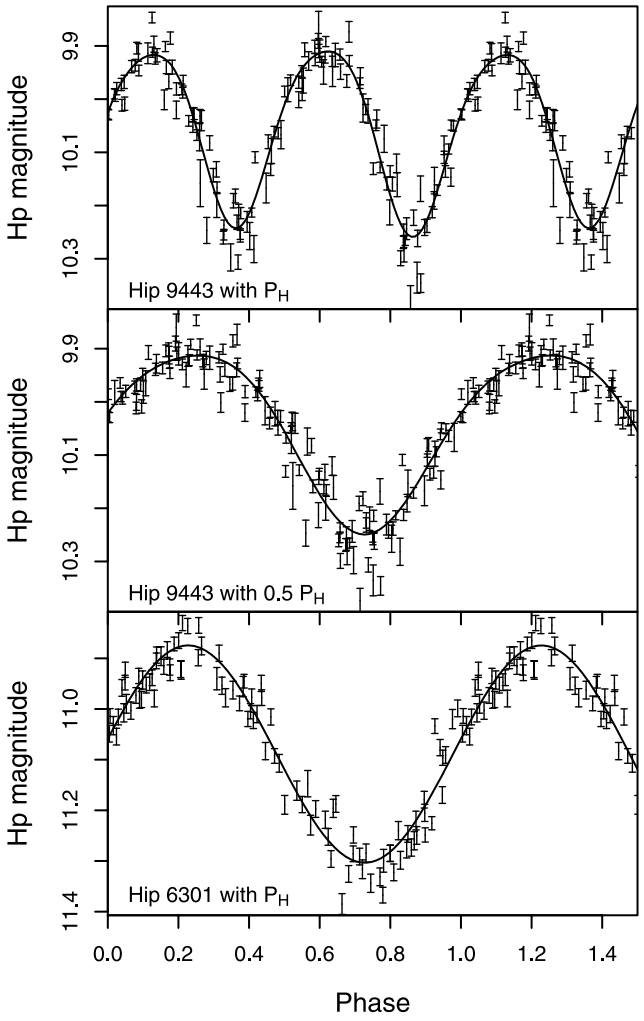

Figure 2. The top and bottom panels show that the folded light curves of the *EB* Hip 9443 and of the *RRC* Hip 6301 can be easily differentiated. However, an automatic period search is most likely to result in half of the correct period for Hip 9443 leading to the folded light curve displayed in the middle panel. This curve is similar to that of the bottom panel illustrating the difficulties of distinguishing these two stars in an automated process.

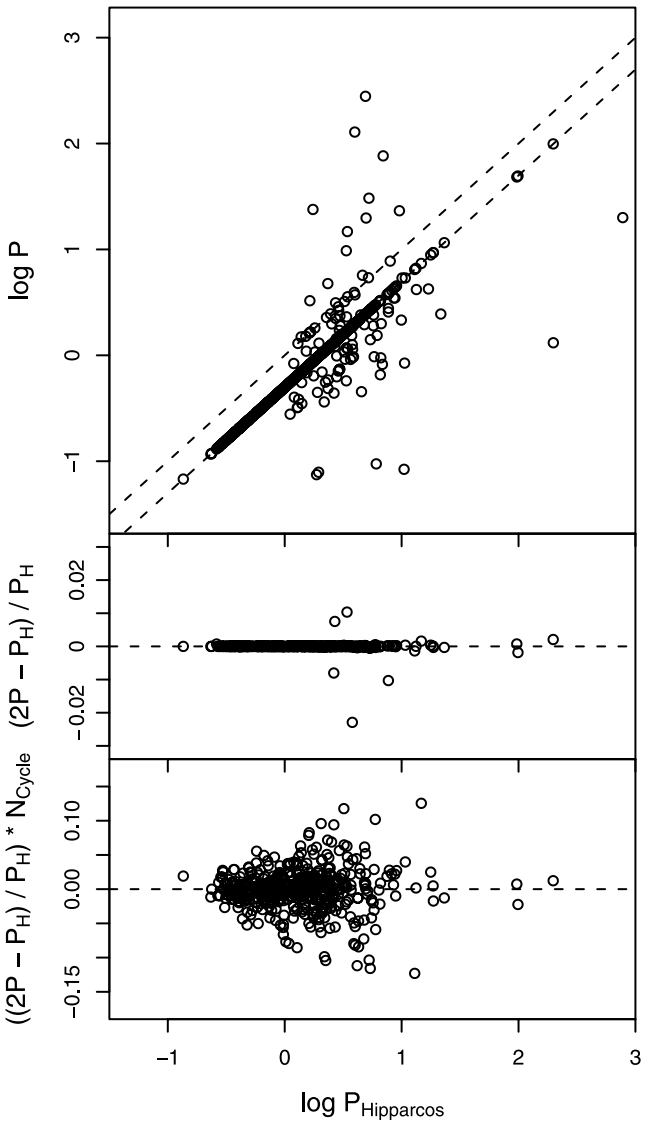

Figure 3. Periods extracted using the Lomb–Scargle method for eclipsing variables as a function of the *Hipparcos* periodic star catalogue periods P_H expressed in days on a decadic log scale. The upper plot shows the Lomb–Scargle period (P), the middle one the relative difference and the lower one the relative difference multiplied by N_{Cycle}. N_{Cycle} is the light-curve span divided by the period value, i.e. the number of cycles between the beginning and the end of the light curve. The lower diagram shows the cumulative shift, expressed in unit of phase, after N_{Cycle} resulting from the inaccuracy of the period value. The dashed diagonal lines in the upper plot display the relationships $P = P_H$ and $P = 0.5\,P_H$. The middle and lower plots have much enlarged y-scales so that almost all outliers visibly scattered in the upper plot fall outside of the displayed ranges.

of the *Hipparcos* period is obtained for 508 stars, i.e. for 82 per cent of the sample. For these 508 cases, the difference between the double of the extracted period and the *Hipparcos* value does not lead to a cumulative phase shift of more than 15 per cent over the full time-span of the light curve. A period close to the *Hipparcos* full period value is obtained only for 10 cases. Some period search algorithms are better than Lomb–Scargle to find the correct period. With Jurkevich–Stellingwerf, for example, the correct value and half of it are obtained in similar numbers of cases (38 per cent). The trouble is that when the true period is unknown, it is impossible to know in which cases, the double or the full value of the extracted period is the correct result. For this paper, we have found that the best strategy is to use the Lomb–Scargle period to model the light curves and to double the period values after classification for all objects that have been identified as eclipsing binaries and ellipsoidal variables. In this way, almost all these stars are modelled with half of the correct periods, as shown in the middle panel of Fig. 2. The use of half the true period in the case of eclipsing binaries does not much confuse the classifier and leads to the best overall results.

3.3 Light-curve modelling

The *Hipparcos* periodic star catalogue provides a unique period for each source. Although a number of these sources are truly multi-periodic, looking at the folded light curves displayed in the *Hipparcos* periodic star catalogue shows that in the vast majority of cases the curves obtained with the single, dominant period look good. Indications of additional significant periods, such as an apparent superposition of two curves or a strong scatter excess with respect to the nominal photometric uncertainties are only evident in a few cases. In this paper, we show that a light-curve modelling

carried out with the dominant period is sufficient to achieve a reliable classification.

The time-series model is given by

$$y = a_0 + \sum_{k=1}^{N_\mathrm{h}} b_k \cos(2\pi k \nu t) + c_k \sin(2\pi k \nu t), \quad (1)$$

where N_h is the number of significant harmonics. Here, $k = 1$ is the fundamental frequency ν (or first harmonic) and $k = 2$ is the first overtone (or second harmonic). This model is linear in its coefficients $\beta = \{a_0, b_k, c_k\}$. An alternative notation involves the amplitudes and phases of the oscillations and their overtones

$$y = a_0 + \sum_{k=1}^{N_\mathrm{h}} A_k \sin(2\pi k \nu t + \varphi_k). \quad (2)$$

The coefficient correspondence is

$$A_k = \sqrt{b_k^2 + c_k^2} \quad (3)$$

$$\varphi_k = \arctan(b_k, c_k). \quad (4)$$

The number of harmonics (N_h) to be used to best fit a given light curve is unknown and must be determined through the modelling process. A model with more parameters, i.e. a higher number of harmonics, will always fit the data at least as well as the model with fewer harmonics. The question is whether the model with additional harmonics gives a *significantly* better fit to the data. A *forward selection* regression process is followed, starting with only the fundamental coefficients, i.e. with $N_\mathrm{h} = 1$ and adding one more harmonic at a time. The different models are *nested*, i.e. a simpler model can be obtained by zeroing one or more coefficients of a more complex model. Different models are compared pairwise using an F-test. For example, in order to compare model 1 and model 2, having different harmonic numbers ($N_{\mathrm{h}2} > N_{\mathrm{h}1}$), the following F statistic is computed

$$f = \frac{RSS_1 - RSS_2}{RSS_2} \frac{n_\mathrm{obs} - P_2}{P_2 - P_1}, \quad (5)$$

where *RSS* are the residual sum of squares, n_obs is the number of data points (i.e. of observations), P is the number of free parameters ($P = 2N_\mathrm{h} + 1$), where subscripts 1 and 2 refer to models 1 and 2, respectively. Under the null hypothesis model 2 does not provide a significantly better fit than model 1 and f follows an F-distribution, with $(P_2 - P_1)/(n_\mathrm{obs} - P_2)$ degrees of freedom. The null hypothesis is rejected if the f calculated from the data is greater than the critical value of the F-distribution for some specified false-rejection probability. A succession of models with a number of harmonics ranging from 1 to N_max are computed and compared in turn. The N_h-th harmonic is only retained if the corresponding model is significantly better than the model with the lower harmonic number. The threshold to keep a given model is α/N_max (Bonferroni correction) where α, set to 5 per cent in this study, is the overall type-I error rate (i.e. accepting unduly one of the harmonics). There may be gaps in the harmonic sequence when some lower harmonics are rejected while higher ones are retained (e.g. a model with $N_\mathrm{h}+2$ may be better than the model with N_h while model with $N_\mathrm{h}+1$ was rejected).

The sampling of the *Hipparcos* time series is irregular with occasional large time gaps. In some cases, models with high harmonic number which fit the data very well exhibit large, unphysical excursions within the gaps. In order to prevent such cases, a stop criterion based on the width of the maximum phase gap of the time series is introduced. The addition of new harmonics is stopped when

$$N_\mathrm{h} \geq \frac{C}{2P_\mathrm{gap}}, \quad (6)$$

i.e. N_max, the maximum number of harmonics, is taken as the first harmonic number N_h in the increasing sequence that satisfies the above inequality. P_gap is the width of the maximum phase gap (between 0 and 1), and C is a constant whose appropriate value is derived empirically from extensive testing. A optimum of 1.4 determined through visual inspections is obtained, but changing this value in a wide range, i.e. from 1.2 to 1.6, only leads to modifications in a handful of the *Hipparcos* light-curve models.

Extensive visual inspection of the resulting models shows that this approach is robust and that it provides good models for almost all cases. The only notable exception is for some of the eclipsing binaries of type *EA* and *EB*, with sharp eclipses, where the number of harmonics is not sufficiently high to fully model the eclipses. In some cases, the number of points in the eclipse is too few to provide sufficient weight in the fitting process. There may also be phase gaps larger than the eclipse duration. In this second case, application of the phase gap criterium stops the iterative process and the optimum harmonic number cannot be reached. However, this is deemed less harmful than the large model artefacts that can occur within the gaps if the stop criterion is not applied.

4 RANDOM FOREST

Random forest (Breiman 2001) is a tree-based classification method. Extensive documentation and Fortran programs by Breiman and Cutler are available at http://www.stat.berkeley.edu/~breiman/RandomForests/. Both the R RANDOMFOREST package (Liaw & Wiener 2002) and the weka (Hall et al. 2009) implementations are used in this work.

4.1 Algorithm

The random forest algorithm aggregates the results of a number (n_tree) of classification trees. Each tree is built as follows.

(1) A *bootstrap* star sample is obtained by drawing a sample with replacements from the training set. The bootstrap sample has the same size as the original set, but some stars are represented multiple times, while others are left out. The omitted stars, called *Out-Of-Bag (OOB)*, can be used to estimate the prediction error (see below).

(2) The tree is grown by recursively partitioning the bootstrap sample into subgroups with more and more homogeneous type content. At each node, m_try divisions into two groups are considered, each using one attribute from a randomly selected set of m_try attributes. The best split is selected and the process is repeated for the child nodes with a new set of m_try attributes at each node.

(3) A so-called *maximum* tree is constructed, i.e. a tree with terminal nodes containing only a single type of stars (or a single star in extreme cases).

Typically, large numbers (500–10 000) of trees are built. Each tree provides a predicted type for a star. The most probable type is simply the most frequent type in the sample of predictions of the different trees.

An estimate of the error rate can be obtained from the training set. Any training set star is OOB in some fraction (about one-third) of the trees. The most frequent type obtained from all the trees where a star is OOB provides a predicted type for each star. Note that the sample of trees in which a given star is OOB is different for each star. The error rate and confusion matrix can be built by comparing the predicted with the actual types. This is similar to a CV performance estimate, but at a much lower computational cost.

It is known that random forest is insensitive to the precise value of $m_{\rm try}$. In this paper, $m_{\rm try}$ is taken by default as the recommended value, i.e. the square root of the number of considered attributes, unless specified explicitly.

4.2 Attribute importance

Random forest can produce an attribute importance score based on the following idea. The classification accuracy computed by passing the OOB sample down a specific tree is recorded. The values of a given attribute are permuted in the OOB sample, i.e. the value for a star is randomly taken out of the sample of all other star values. The classification accuracy is computed again with the OOB sample with permuted values for one attribute. If this attribute is important, the permutation should noticeably degrade the classification accuracy. Conversely, it should not change significantly the predictions if this attribute is ineffective in the classification process in the first place. The attribute importance is given by the difference in classification error averaged over all trees and normalized by the standard deviation (of these differences). These importance values are extensively used in our attribute selection scheme.

4.3 Attribute correlation and selection

Many of the derived star attributes are highly correlated. As explained in Section 3, there may be several alternative ways to characterize a given physical property. For example, there are different ways to measure the amplitude of a light curve, and different colour indices are all, to the first order, a measure of the star effective temperature. The idea is now to investigate which of these alternatives leads to the best results in classification.

The attribute importance estimates provided by random forest can be used to rank attributes. The limitation is that it is not sensitive to correlations. Two highly correlated attributes will score equally highly in this process. Experience shows that random forest classification results are not much affected by the use of some almost redundant, highly correlated attributes, but it is interesting to investigate what is the minimum set of attributes to be used for an optimum classification of our stars.

The recursive procedure to build a list of the most important, not too correlated attributes is as follows.

(1) A ranked list of attributes, from the most to the least important, is built using a 2000-tree random forest with the full attribute set.

(2) The most important attribute is selected and all other attributes with a Spearman correlation coefficient above 80 per cent with this one are discarded.

(3) A new ranked attribute list is built re-running a random forest with the selected and the remaining attributes.

(4) The second most important attribute is selected and all other attributes highly correlated with any of the first two are discarded.

(5) This process is iterated until a full ranked list of not-too-correlated attributes is obtained.

This procedure is somewhat unstable if the number of attributes is too large. The most important attributes are always highly ranked, but the order of the moderately important ones may change drastically from one run to the next. Clearly, the importance measurement of a given attribute depends to some level on the background of the other attributes. This is even amplified if the attributes are correlated. If the attribute under evaluation is highly correlated with other ones, replacing that attribute with random noise does not affect much the results as the other attributes have similar classification power. An effective way around this difficulty is to remove a large fraction of the least important attributes before starting the above recursive procedure.

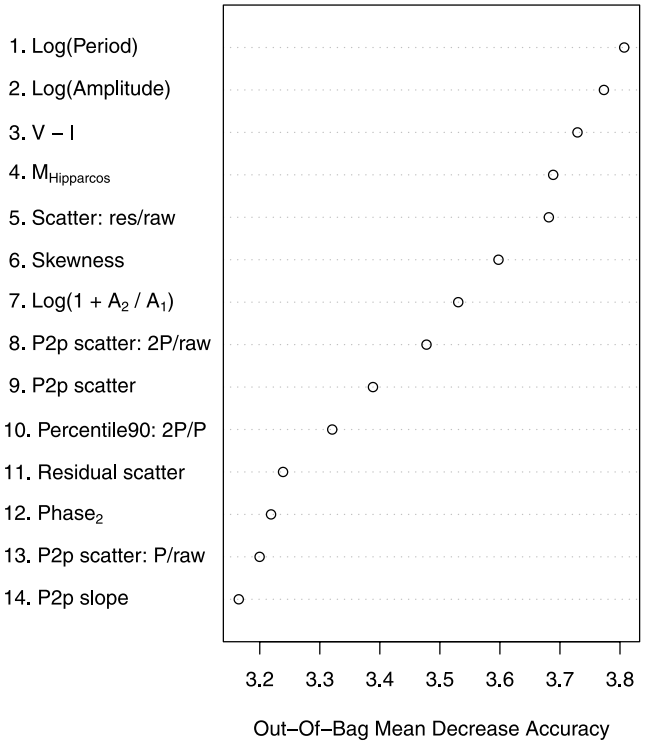

Figure 4. The ranked list of the 14 most important, not-too-correlated attributes (defined in Section 4.5). The Spearman correlation coefficient of any of the above attribute pairs is smaller than 80 per cent. The attribute importance is measured with the random forest OOB mean decrease accuracy.

Some astronomical insight is also injected into this selection process. When two, or several, attributes have similar importance, the one with a simpler and/or more widely used definition is preferred. Fig. 4 displays the results of the above attribute ranking procedure for the 14 most important attributes. A detailed attribute description is provided in Section 4.5.

4.4 Towards a minimum attribute list

The procedure described in the last section is used to derive a ranked list of not-too-correlated attributes. The importance value decreases in the list but it never reaches zero. A key question is where to cut the list. Are all attributes really useful? Or, are the low-importance attribute contributions already included in those of more important attributes? This second possibility is more likely as many of the low-importance attributes are correlated at some level with some of the more important attributes in the list.

In order to reduce the number of attributes, a variant of the method proposed by Svetnik et al. (2004) is adopted. Only the list of not-too-correlated attributes, derived as described in the previous section, is used through the following algorithm.

(1) The data are partitioned for a 10-fold CV.

(2) On each CV training set, a ranked list of attributes is established using the random forest importance measures as described in the previous section.

(3) On each CV training set, a model is trained on all attributes and used to predict types for the CV test set. The CV error rate

is recorded and the process is repeated after removing the least important attribute. Iterating by removing one attribute at a time and stopping when only two attributes are left, a vector of CV error rates is obtained for an attribute number ranging from two to the total.

(4) At the end of the 10-fold process, a mean error vector is computed by taking the mean of the 10 values obtained for each attribute sub-set.

(5) Steps 1 to 4 are repeated 20 times. The mean value and the standard deviation of the 20 CV mean errors are computed for each attribute number, combining the results of the classification experiments achieved with a specific attribute number.

Fig. 5 shows the error rates as a function of the number of attributes resulting from the above procedure. The optimum number of attributes can then be inferred from this figure. As the attribute number increases, Fig. 5 shows that the error rates first decrease and then level-off at some value. A CV error rate under 18 per cent is obtained with the first seven most important attributes, and a minimum of 16.4 per cent is reached with seven more attributes. The large number of additional attributes tested in this study are not mentioned as they do not lead to any further improvement of the classification results.

It is interesting to note that Fig. 5 is much more contrasted than Fig. 4. While the latter indicates a steady, almost linear decrease in attribute importance, the drop in error classification, and hence in attribute merit, seen in Fig. 5 is much more abrupt. This is probably due to the fact that the importance displayed in Fig. 4

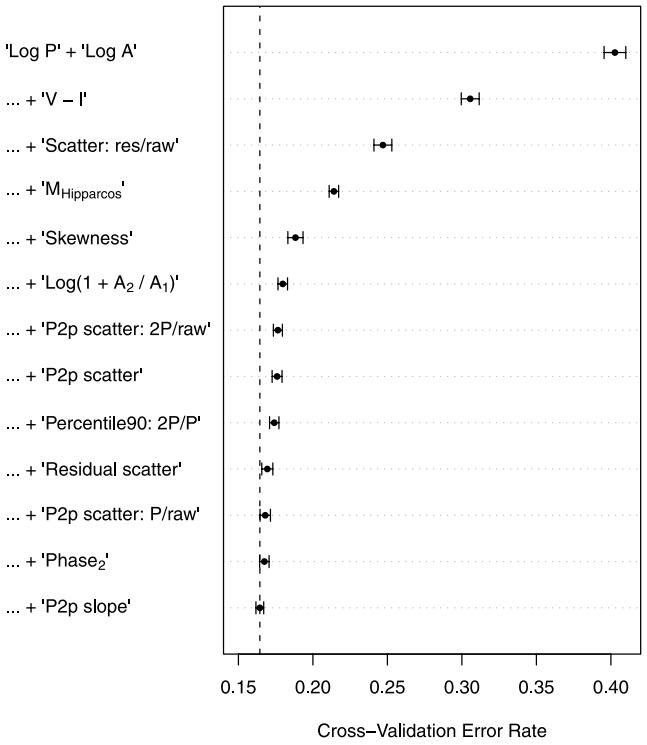

Figure 5. Evolution of the CV error rate as more and more attributes are added into the classification process. The seven most important attributes already drive the error rate under 18 per cent, while a minimum of 16.4 per cent is reached with an additional seven attributes. As explained in the main text, some randomness is included in our classification process. As a consequence, the attribute order can vary slightly in the different CV experiments. The attribute name provided at a given line in this figure is the name of the attribute appearing most frequently at that position in the different experiments.

is measured against the background of the other 13 attributes. As there is some remaining correlation between attributes, some of the other attributes can compensate for the loss of the specific, evaluated attribute, whose values have been permuted (see Sections 4.2 and 4.3).

4.5 The most important attributes

The 14 most important attributes listed in Figs 4 and 5 are defined below.

(1) **Log(Period)** : decadic log of the period extracted with the Lomb–Scargle method (see Section 3.2).

(2) **Log(Amplitude)** : decadic log of the amplitude of the light-curve model.

(3) $\boldsymbol{V-I}$: the mean $V-I$ colour.

(4) $\mathbf{M}_{\mathrm{Hipparcos}}$: a *Hipparcos* absolute magnitude derived from the parallaxes neglecting interstellar absorption. Because of measurement uncertainties, some stars have negative parallax values. Each of these values is replaced by a positive value taken randomly from a Gaussian distribution with zero mean and a standard deviation equal to the measurement uncertainty. In many cases, the derived absolute magnitudes represent lower limits as the parallax measurements are not significant.

(5) **Scatter: res/raw** : Median absolute of the residuals (obtained by subtracting model values from the raw light curve) divided by the Median Absolute Deviation (MAD) of the raw light-curve values around the median.

(6) **Skewness** : unbiased weighted skewness of the magnitude distribution.

(7) **Log(1 + $\mathbf{A}_2$ / $\mathbf{A}_1$)** : decadic log of the amplitude ratio between the second harmonic and the fundamental (plus one, to avoid negative values).

(8) **P2p scatter: 2P/raw** : sum of the squares of the magnitude differences between pairs of successive data points in the light curve folded around twice the period divided by the same quantity derived from the raw light curve.

(9) **P2p scatter** : median of the absolute values of the differences between successive magnitudes in the raw light curve normalized by the MAD around the median.

(10) **Percentile90: 2P/P** : the 90th percentile of the absolute residual values around the 2P model divided by the same quantity for the residuals around the P model. The 2P model is a model recomputed using twice the period value.

(11) **Residual scatter** : mean of the squared residuals around the model.

(12) **Phase**$_2$: phase of the second harmonic after setting the phase of the fundamental to zero by an appropriate transformation $\mathrm{Phase}_2 = \arctan(\sin(\varphi_2 - 2\varphi_1), \cos(\varphi_2 - 2\varphi_1))$ (Debosscher et al. 2007).

(13) **P2p scatter: P/raw** : median of the absolute values of the differences between successive magnitudes in the folded light curve normalized by the MAD around the median of the raw light curve.

(14) **P2p slope** : sum of the square of the slopes of lines joining the data points before and after a number of selected outliers towards faint magnitude (e.g., data points during eclipses). This is set to zero if there are no such outliers in the light curve.

$$\mathrm{P2p\ slope} = \left\{ \sum_i \left[\left(\frac{d_i - d_{i-1}}{t_i - t_{i-1}} \right)^2 + \left(\frac{d_{i+1} - d_i}{t_{i+1} - t_i} \right)^2 \right] \right\}^{0.1} \quad \text{for } d_i > 3,$$

with $d_i = (y_i - P_{25})/\delta_i$, $\delta_i = (\sigma_i^2 + \delta^2)^{1/2}$ and $\delta = P_{25} - P_5$, where y_i and σ_i are the observed magnitude and its error, respectively, at time t_i, and P_n is the nth percentile of the magnitude distribution.

4.6 Attribute display

Figs 6 and 7 display the distributions of the eight most important attributes for each of the variability types.

4.7 Classification error analysis

Fig. 8 displays the confusion matrix resulting from a 10 000-tree random forest classification of our training-set stars. The 14 attributes described in Section 4.5 are used and the number $m_{\rm try}$ (see Section 4.1) of attributes tried at each node is three. The matrix rows indicate the reference types resulting from the literature survey presented in Section 2, while the columns represent the classifier *predicted* types. The classification process fails to separate some of the types, namely (1) *DSCT* from *SXPHE*, (2) *BE* from *GCAS* and (3) *BY* from *RS*. These types are pairwise merged in Fig. 8 to improve the matrix readability. In principle, the *SXPHE* could be separated if a metallicity estimator was used in the classification and *BY* and *RS* could be distinguished if better absolute magnitudes were available. The case of the *BE* and *GCAS* is discussed below.

The overall classification error rate derived from the OOB samples is 15.7 per cent. It is slightly lower than the 16.4 per cent presented in Fig. 5 because of the above-mentioned merging of six types. However, this overall rate does not bear much meaning as confusions within groups of similar stars are less problematic than others. The most important confusion cases are detailed in the

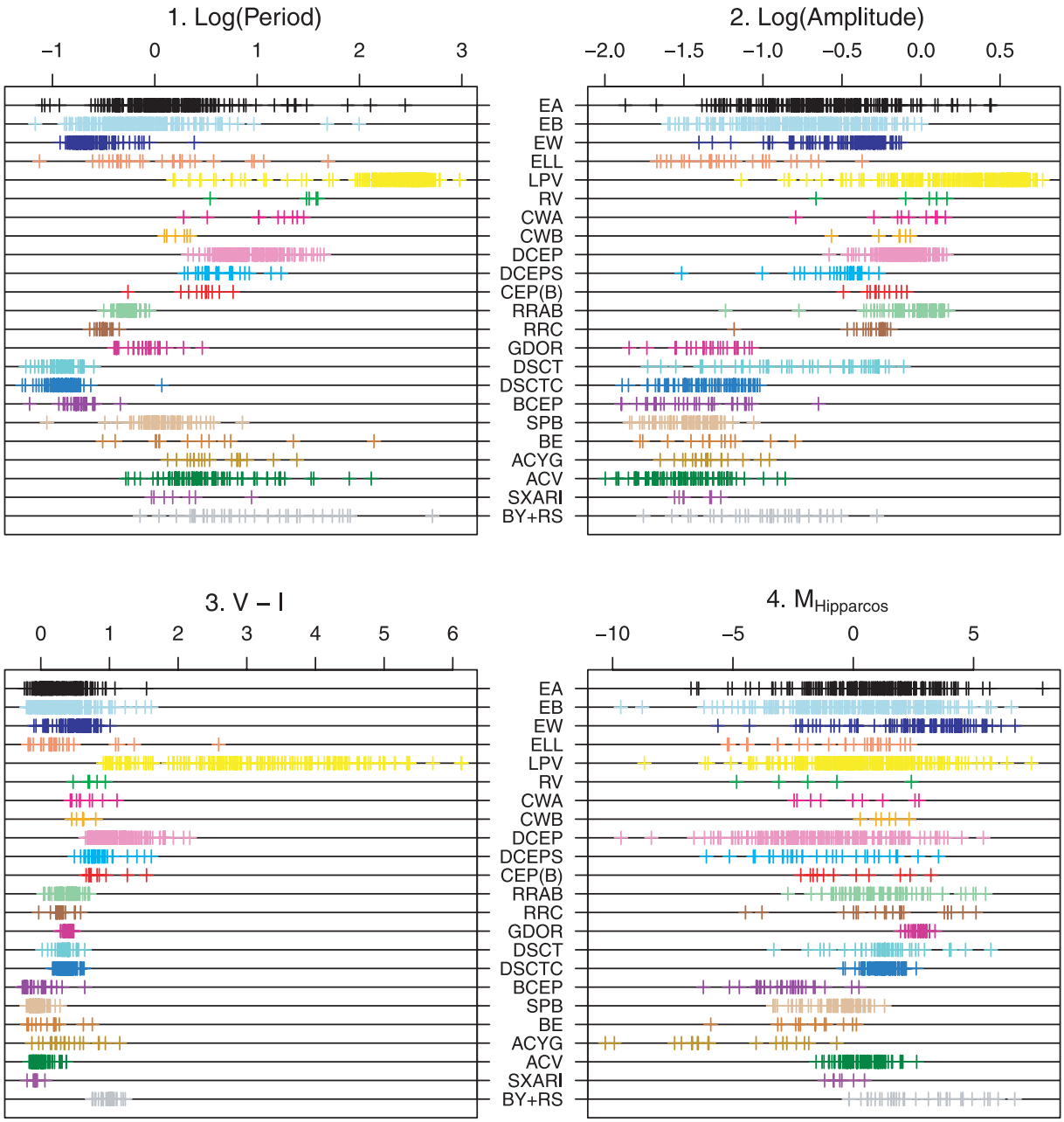

Figure 6. Distributions of the four most important attributes obtained for the training-set members.

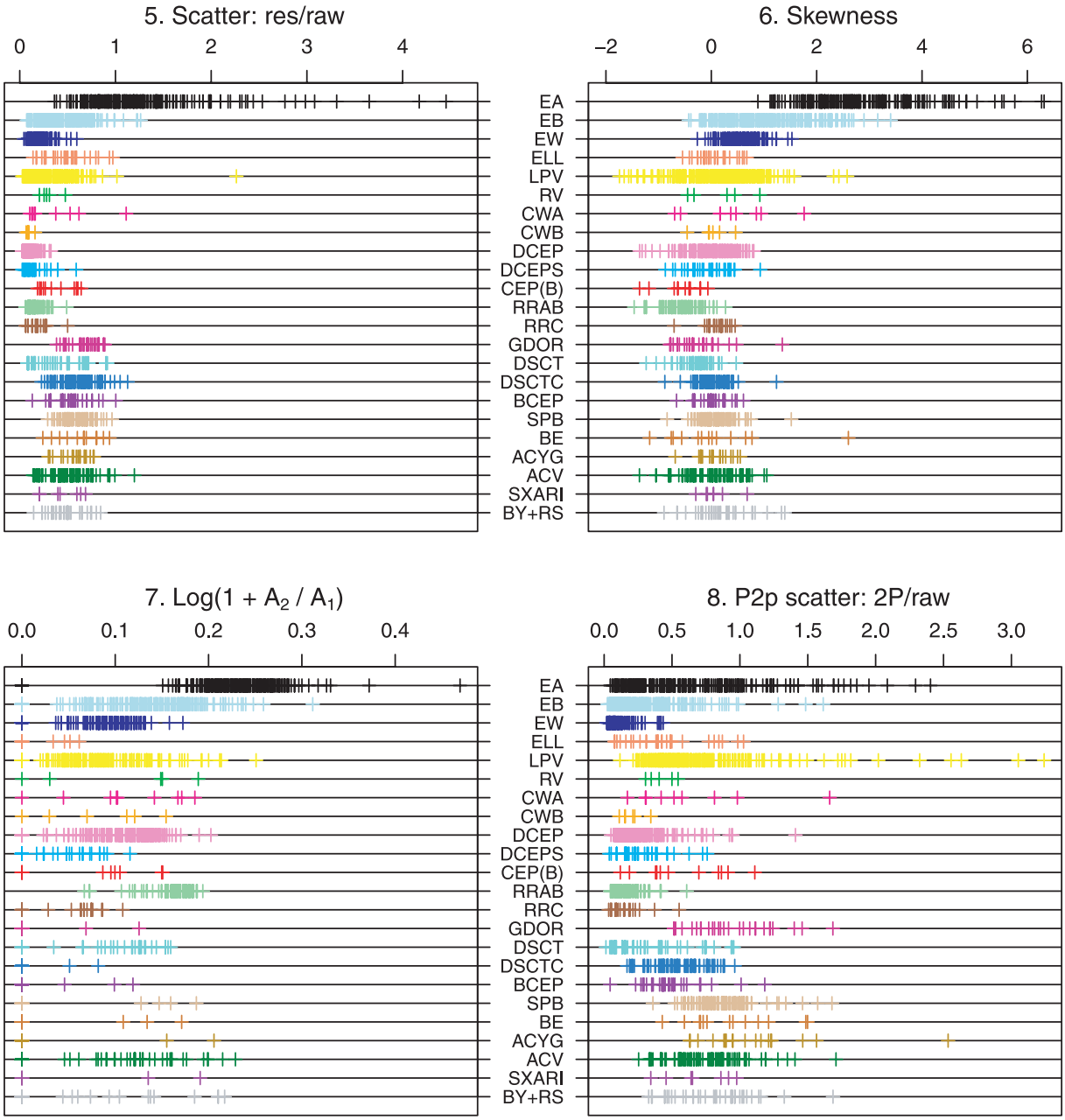

Figure 7. Distributions of the attributes ranking 5 to 8 in random forest importance obtained for the training-set members.

following sections. It is important to remember that random forest involves randomness in the sample bootstrapping and node attribute selection (see Section 4.1). As a consequence, the confusion matrices obtained in successive identical runs differ slightly at the level of a few cases.

4.7.1 Eclipsing binaries and ellipsoidal variables

As already alluded to in Section 3.2.2, eclipsing binaries are expected to be a difficult case. It is therefore not surprising to see from Fig. 8 that they are involved in the most important confusion cases. The classification disperses 17 *EB* into other types of non-eclipsing variables. There are 14 *ELL* variables mis-classified as *EB* while 13 of them are scatter into six other types. In addition, 19 and six cases of diverse other non-eclipsing and non-ellipsoidal variables are unduly classify as *EB* and *EW*, respectively.

Could this confusion be diminished if the full training set is first separated into eclipsing and non-eclipsing variables? To investigate this issue the attribute ranking and selection procedure are repeated considering two type groups, *EA*, *EB*, *EW* and *ELL* on the one side, and all other types on the other. The resulting attribute ranking is quite different but the classification results do not improve. It is possible to lower the number of variables falsely classified as eclipsing binaries to about 20 cases, but then, the number of mis-classified eclipsing binaries increases to about 50, so that the total is slightly worse than the result of a direct classification into all types.

The trouble is that the light curves of some *EB*, *EW* and *ELL* are quite symmetrical and resemble those of other variability types.

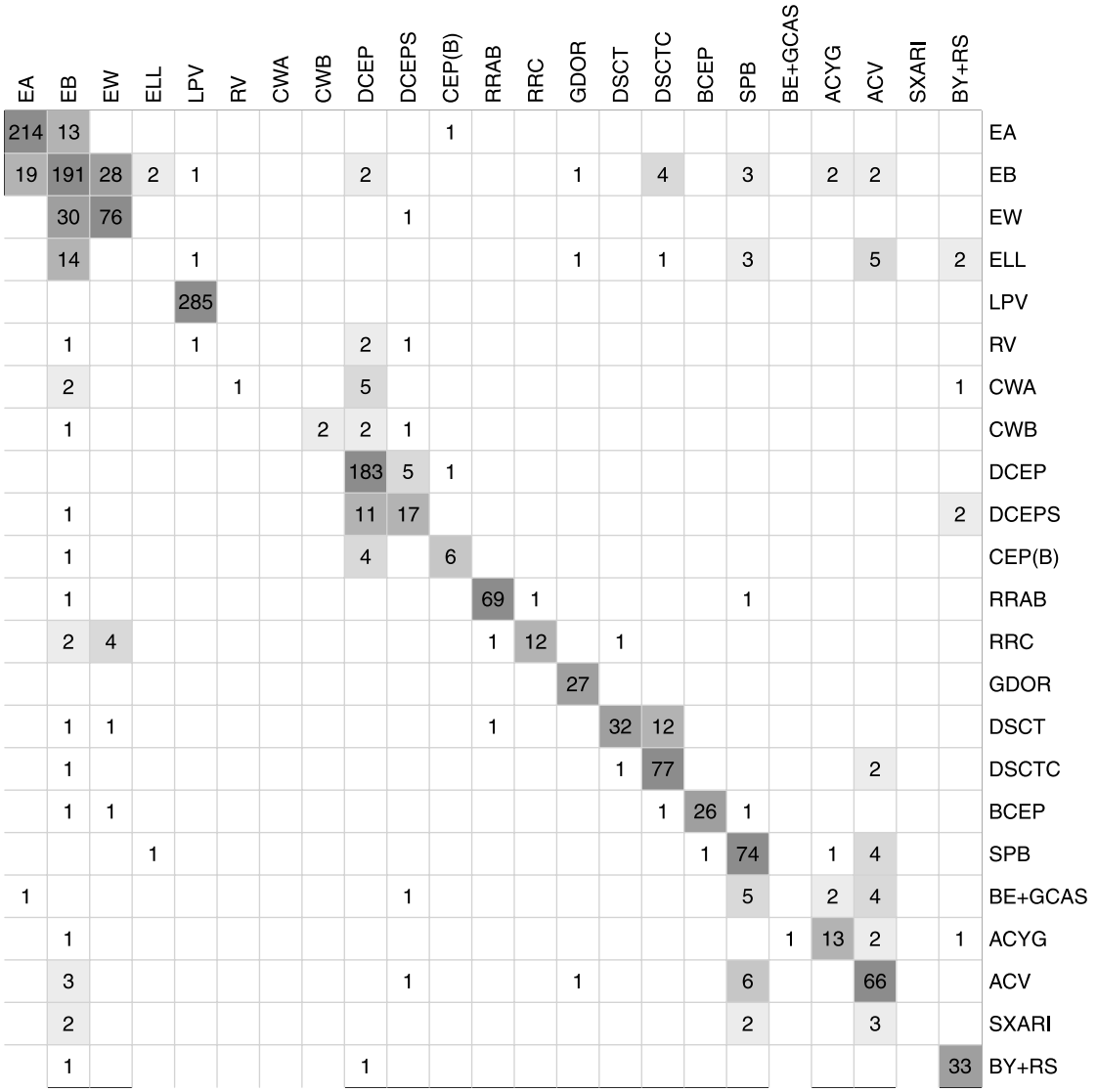

Figure 8. Confusion matrix obtained for our training set with a 10 000-tree random forest classification using a group of 14 attributes. The rows indicate the reference types resulting from the literature survey (see Section 2), while the columns represent the classifier *predicted* types. Type labels are as described in Table 1.

In addition, stellar properties such as colour and absolute magnitude can take almost any possible value as they are the combination of the properties of the two stars of the binary system. As a consequence, it is likely that these confusion cases represent a true physical difficulty that cannot be fully solved by any classification method.

4.7.2 Cepheid- and Cepheid-like variables

As can be clearly seen in Fig. 8, all types of Cepheid-like variables are confused with the Delta Cepheid type. More precisely the following cases can be listed.

(1) The *CW* (*CWA* and *CWB*) variables are Population II Cepheid stars of lower absolute magnitude (and smaller mass). The *Hipparcos* parallax measurements are not significant for these relatively bright and remote stars. As a consequence, the derived absolute magnitudes are dominated by noise and this explains the confusion as these stars are otherwise similar to *DCEP*. The *CW* and *DCEP* variables could be separated using a metallicity indicator or more reliable luminosity estimates.

(2) The *DCEPS* stars are Cepheids with smaller amplitude and period values, which probably pulsate in the first overtone. There is, however, an overlap with the *DCEP* stars in the log(Period)–log(Amplitude) diagram and this probably explains the confusion cases. Out of the sample of 31 *DCEPS* stars, 17 are correctly classified while 11 fall wrongly in the *DCEP* category.

(3) The *CEP(B)* stars are Cepheids which exhibit two or more pulsation modes. They could almost certainly be better singled out by searching for additional significant periods and using them in the characterization and classification processes. This concerns, however, a small number of stars [only four out of 10 *CEP(B)* stars are wrongly classified as *DCEP*] and it is outside of the scope of this paper, which is restricted to single-period analyses.

In addition, there seems to exist a not well-understood confusion between *RV* and *DCEP* although small number statistics here is a limitation.

4.7.3 *Blue variables*

The third case of confusion concerns the blue variables. First *BE* and *GCAS* have been put together in Fig. 8. These types can only be separated on the basis of long-term behaviour. *GCAS* show eruptive, non-periodic events (Samus et al. 2009)[6] and our sample includes only those stars where the observed signal is periodic. The short-term periodic behaviour of some *BE* + *GCAS* is also similar to the one observed in the *SPB* stars (e.g. Diago et al. 2009). As a consequence, it is not surprising to observe a confusion between *SPB* and *BE* + *GCAS*.

The confusion between *BE* + *GCAS* and *ACV* cannot be so easily understood. It most probably comes from a true confusion between *ACV* and *SPB*, seen at the 10 per cent level in Fig. 8, which also concerns the few *SXARI* which are physically relatively close in attribute space to the *ACV*.

4.8 Random forest and linear discriminant analysis

The linear discriminant analysis (LDA) (Mardia, Kent & Bibby 1979; Hastie, Tibshirani & Friedman 2009) approach is applied to derive an optimum set of independent linear discriminants (LDs). The goal is to run a random forest classification using these LDs as an alternative set of attributes.

In the attribute multi-dimensional space, objects of a particular type can be visualized as a 'cloud' of data points. The complete training set is then viewed as a set of generally overlapping clouds that are to be separated by the process of classification. The idea of LDA is to derive linear transformations of the attributes which maximize the ratio of the cloud centre variance divided by the variance of the data points within the clouds. In other words, the transformation seeks to rotate the axes so that when the objects are projected on the new axes, the differences between the different clouds (i.e. types) are maximized.

The LDA-based classification scheme goes through the following steps.

(1) For each attribute, the distribution of values is standardized by subtracting the mean and dividing by the standard deviation.

(2) An LDA is carried out. The resulting LDs are ranked as a function of the singular values, i.e. the most important LD is the one with the largest ratio of the variance of the group centres over the within-group variance.

(3) The attributes are ranked according to their maximum contribution to any of the most important LDs.

(4) Different attribute selection schemes are used, iterating and removing one, or a few of the least important attributes and of the highly correlated ones in each of the successive steps. Although, this process is completely independent from the one used previously for random forest, the final list of selected attributes is very similar, with period, amplitude and colour attributes always standing out as the best three.

(5) The resulting list of attributes is used in a final LDA. The LDs are computed for all stars and used for a random forest classification.

(6) The classification errors are estimated using the OOB sample and a 10-fold CV method.

Although many attempts have been performed varying the attribute selection scheme, the resulting classification errors are always significantly worse (by at least 3 per cent) than those obtained when applying random forest to the original attributes. The derived LDs are less correlated than the original attributes (even when relaxing the selection criteria and accepting more highly correlated attributes), but surprisingly, it did not lead to better results in our case. Thus, LDA is not used to produce any of the results presented in this paper.

[6] http://www.sai.msu.su/gcvs/gcvs/iii/vartype.txt

Table 2. Confusion induced by incorrect periods for non-eclipsing variables.

	Total	Misclassified stars	Error rate
Stars with correct period	951	102	10.7 per cent
Stars with incorrect period	93	20	21.5 per cent
All stars	1044	122	11.7 per cent

5 DEGRADATIONS DUE TO ERRORS IN THE PERIOD DETERMINATION

As shown in Fig. 1, in some cases, the period values resulting from the search done in this work are completely different from the *Hipparcos* periodic star catalogue values. Although, the latter values are probably more reliable as they were visually checked, our classification is based on our own values as the idea is to evaluate the performance of an automated classification process (see Section 3.2). It is, however, interesting to investigate the classification degradation induced by wrong period values. The stars with incorrect periods can be traced to evaluate how well they are classified. The level of confusion observed for these stars can be compared with that seen for stars with correct periods.

We exclude from this comparison the eclipsing binaries and the ellipsoidal variables as it is known (see Section 3.2.2) that the period found is systematically half of that of the true period for these stars. Since the eclipsing binary periods span a wide range of values, the confusion due to an incorrect period is likely to be less severe than that observed for other stars.

The comparison presented in Section 3.2.1 shows that out of the total of 1044 non-eclipsing variables, an incorrect period is found for 93 stars. Table 2 displays the classification errors obtained in the different cases. The eclipsing binaries are excluded, but the cases where a non-eclipsing is classified as an eclipsing (or an ellipsoidal) variable are accounted for in the numbers of misclassified stars.

Although statistical uncertainties due to the small numbers is a limitation, this table shows that 73 out of the total of 93 stars with an incorrect period are successfully classified into their proper types. This is surprising as the period always stands as the most important attribute. Somehow, other attributes, such as the amplitude, the colour, or the pseudo-absolute magnitude, compensate and safeguard against an incorrect classification in an important number of cases.

6 COMPARISON WITH A MULTI-STAGE CLASSIFIER

For comparison with the results shown in Section 4, a methodology based on a divide-and-conquer approach is applied whereby the overall classification problem with 26 variability types included in Table 1 is sequenced into several stages. The variability zoo in that table is grouped into categories and subcategories, and the classification of a star proceeds by assigning a probability vector for each category and subcategory until one of the variability types defined in Table 1 is reached (the leaves of the tree defined in Fig. 9).

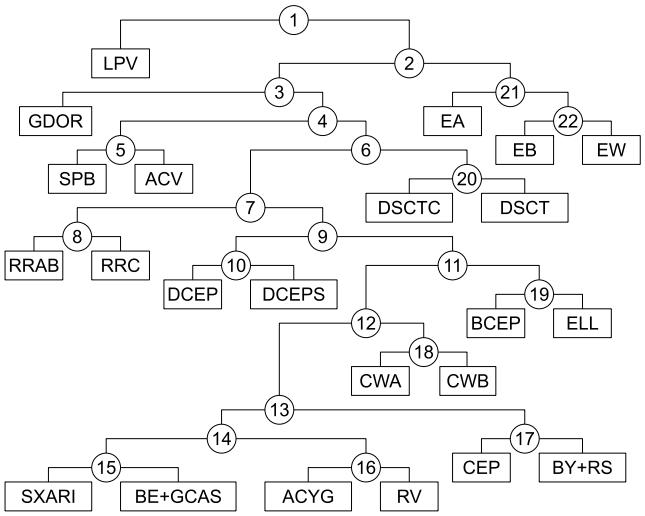

Figure 9. Sequential scheme for the separation of all variability types with dichotomic classifiers.

This multi-stage scheme is defined in more detail in Blomme et al. (2011). Here a different grouping of variability types based on an automatic definition of categories is tested.

The algorithm used to define the multi-stage scheme shown in Fig. 9 is based on the confusion matrix obtained by using a monolithic, single-stage classifier. From this, the similarity between types is determined according to the metrics

$$\text{Similarity}(\mathcal{T}_i, \mathcal{T}_j) = \begin{cases} 1, & i = j \\ \frac{X_{ij}+X_{ji}}{X_{ij}+X_{ji}+X_{ii}+X_{jj}}, & i \neq j \end{cases} \quad (7)$$

where $\mathcal{T}_i$ represents type i and X_{ij} represents the element in the ith row and jth column of the confusion matrix.

The idea behind this metric is that types that are easy to separate should be in the topmost levels of the scheme since they do not affect other types too much, and the most problematic types are positioned at the bottom of the sequence. The algorithm starts from the complete set of types and, in each step, the two most similar types are merged into a new type, and the similarities are calculated again over the new set of types. Thus, the final multi-stage tree is composed of a series of dichotomic classifiers.

The advantage of the divide-and-conquer approach is that the optimal set of attributes used for classification and the optimal classification algorithm can be selected in each node of the tree. In our case, the best algorithm in each node is selected from a set composed of Bayesian Networks (Pearl 1988) and Gaussian Mixtures (Debosscher et al. 2007). These two methods are chosen because both Bayesian Networks and Gaussian Mixtures allow for a simple procedure in order to account for missing attributes. In both cases this is accomplished via marginalization of the posterior-type probabilities given the attributes. In the case of the Gaussian Mixture classifier this can be achieved analytically as

$$p(\mathcal{T}|\boldsymbol{x}_{\text{avail}}) = \int p(\mathcal{T}|\boldsymbol{x}_{\text{avail}}, \boldsymbol{x}_{\text{missing}})\, p(\boldsymbol{x}_{\text{missing}})\, \mathrm{d}\boldsymbol{x}_{\text{missing}}, \quad (8)$$

where $\boldsymbol{x}_{\text{avail}}$ and $\boldsymbol{x}_{\text{missing}}$ are the subsets of available and missing attributes, respectively, which make up the complete attribute vector $\boldsymbol{x}$, and the probability density functions are always normal. $p(\mathcal{T}|\boldsymbol{x}_{\text{avail}}, \boldsymbol{x}_{\text{missing}})$ is an outcome of the training-set classification, and the distribution of the missing attribute $p(\boldsymbol{x}_{\text{missing}})$ can be established from the sub-sample of other stars for which $\boldsymbol{x}_{\text{missing}}$ is available or from astrophysical knowledge of the distribution.

In each node, the classifier that shows the smallest misclassification rate is chosen. The misclassification rate is obtained by averaging the misclassification rates obtained in 10 experiments of 10-fold CV [the so-called multiple runs k-fold CV (Bouckaert 2003); see below for details]. This type of experiment allows for the comparison between two classifiers by statistically testing the null hypothesis that the two classifiers perform equally well. Here for simplicity Bayesian Networks is selected in those nodes where the null hypothesis could not be rejected (i.e. where there was not sufficient evidence that one of the classifiers outperformed the other).

As stated above, one of the advantages of the multi-stage classification is that it allows for context-dependent feature selection. That is, the optimal attribute set for classification can be selected in each node of the tree. This is particularly useful for variability classification where the variables that discriminate between types depend on the types themselves. The procedure adopted here for variable selection starts with an empty set of attributes in each node. Then, the attribute that conveys the largest mutual information with the type is added. Attributes are added to the set following this greedy strategy until the addition of a new attribute produces an increase in the mutual information of less than 0.1. This threshold is found to avoid in most of the cases the inclusion of attributes which are deemed irrelevant on the basis of expert astronomical knowledge. These irrelevant attributes are sometimes picked by the algorithm due to spurious correlations caused by the small training set sample sizes.

The comparison between the classification strategy described in Section 4 and the multi-stage classifier is done following the same procedure (Bouckaert 2003) used to select the best classifier in the nodes of the multi-stage tree. 10 experiments of 10-fold CV are carried out. For each 10-fold CV experiment, the misclassification rate of the two alternative classifiers is subtracted. Let a_{ij} denote the misclassification rate of one of the classifiers in the ith run and jth fold, and b_{ij} that of the alternative. Then, the difference $x_{ij} = a_{ij} - b_{ij}$ is calculated and the values of x_{ij} within the same run are sorted in increasing order. Finally, the values of x_{ij} in each fold are averaged over 10 different runs. Thus, this ends up with 10 sorted average misclassification rates and the corresponding variance estimates, one for each fold. Then, the Z-statistic is computed as follows:

$$Z = \frac{m}{\sqrt{\hat{\sigma}^2}\sqrt{df + 1}}, \quad (9)$$

where m is the mean of the 100 misclassification rates, $\hat{\sigma}^2$ is the variance averaged over the 10 fold-wise variance estimates and df is the number of degrees of freedom. In our case, the calibration by Bouckaert (2003) (i.e. $df = 10$) is used.

It can be shown that the Z-statistic follows a t distribution for two classifiers that perform equally well (the null hypothesis). In our case, a value of $Z = 0.516$ is obtained. It corresponds to a p-value of 0.31 which is clearly above any reasonable confidence threshold. Therefore, there is no evidence to reject the null hypothesis that the two classification strategies (the random forest and the multi-stage tree) perform equally well.

7 AUTOMATED CLASSIFICATION

A complete set of attributes is available for 2543 stars out of the total of 2712 *Hipparcos* periodic variables. A sub-set of 1661 of these stars, selected following the procedure described in Section 2, forms the training set used in previous sections. There are 882 stars left, for which either only an uncertain type is available from the literature (832), or no type at all can be found (50). These stars

Table 3. A sample of the *Hipparcos* training set star list with literature types and attribute values. The full table is available in the online version of the article (see Supporting Information).

Hip	Type	Log(Period)	Log(Amplitude)	$V-I$	$M_{\rm Hipparcos}$	Scatter: res/raw	Skewness	$\log(1+A_2/A_1)$	P2p scatter: 2P/raw	P2p scatter	Percentile90: 2P/P	Residual scatter (per cent)	$\rm Phase_2$	P2p scatter: P/raw	P2p slope
8	LPV	2.5229	0.61	3.92	2.12	2.60	0.61	0.00	0.48	0.07	0.68	11.77	1.00	1.57	1.63
63	ACV	0.5726	−1.40	−0.03	−0.24	1.62	−0.62	0.13	0.85	0.75	1.10	0.01	1.20	1.70	0.00
109	DSCTC	−0.7819	−1.41	0.45	0.95	0.95	−0.18	0.00	0.84	1.36	1.01	0.02	1.13	1.57	0.00
226	RRAB	−0.3068	0.11	0.29	−0.47	11.16	−0.75	0.17	0.11	0.51	0.95	0.11	0.30	2.32	0.00
270	EA	−0.1576	−0.71	0.16	0.77	0.68	2.00	0.17	1.09	1.31	1.01	0.55	1.09	−1.51	4.15
316	DSCTC	−0.7693	−1.19	0.42	0.99	1.73	0.01	0.00	0.49	0.99	0.96	0.04	0.69	1.57	2.37
344	LPV	2.5100	0.71	3.91	4.35	7.38	−0.20	0.00	1.62	0.06	0.60	6.25	0.99	1.57	2.59
623	GDOR	−0.0375	−1.38	0.44	3.40	1.12	−0.36	0.00	1.40	0.94	1.01	0.08	1.42	1.57	0.00
703	LPV	2.5591	0.34	1.53	1.00	5.19	0.47	0.04	0.71	0.15	1.15	1.75	1.02	−0.55	2.78
746	DSCTC	−0.9955	−1.49	0.40	1.24	3.05	0.32	0.00	0.21	1.59	0.95	0.00	0.32	1.57	3.46

Table 4. Results obtained for the *Hipparcos* stars excluded from the training set. This table shows the *Hipparcos* numbers, the literature types, the predicted types and the attribute values for a subset of the sample. The full table is available in the online version of the article (see Supporting Information).

Hip	Type	Predicted type	Log(Period)	Log(Amplitude)	$V-I$	$M_{\rm Hipparcos}$	Scatter: res/raw	Skewness	$\log(1+A_2/A_1)$	P2p scatter: 2P/raw	P2p scatter	Percentile90: 2P/P	Residual scatter (per cent)	$\rm Phase_2$	P2p scatter: P/raw	P2p slope
262	EA:	EA	0.3518	−0.04	0.49	2.31	0.69	5.03	0.23	1.06	0.82	1.86	2.79	1.73	−1.56	3.75
516	LPV:	LPV	2.1605	0.14	2.43	−3.25	2.68	−0.39	0.10	0.86	0.07	0.82	3.38	1.24	2.97	0.00
664	RS+BY:	RS+BY	1.6836	−0.76	1.33	−1.12	4.96	0.26	0.00	0.34	0.16	0.72	0.02	1.03	1.57	1.93
723	LPV:	RS+BY	2.5575	−1.31	0.89	6.19	1.26	0.02	0.00	0.90	1.10	0.95	0.05	1.05	1.57	1.71
864	LPV:	RS+BY	1.5759	−0.71	1.72	−1.17	2.75	−0.34	0.08	0.77	0.29	0.61	0.12	1.10	1.54	0.00
871	EB:	EA	1.2616	−0.05	0.73	0.32	1.61	2.70	0.20	0.27	0.85	1.23	0.29	0.83	−1.81	2.73
988	LPV:	LPV	1.6542	−0.52	2.67	−4.32	2.43	−0.70	0.00	0.67	0.31	1.02	0.10	0.91	−1.37	0.00
1110	LPV:	LPV	1.1795	−0.56	2.06	0.28	3.18	−0.41	0.00	0.68	0.25	1.01	0.37	1.23	1.57	0.00
1263	EB:	EA	0.9776	−0.62	0.46	2.26	0.91	3.19	0.21	0.94	1.50	0.91	0.32	0.79	−1.60	3.41
1378	ACV:	SPB	−0.0238	−1.22	−0.06	−2.30	1.30	0.25	0.00	0.71	1.14	1.01	0.04	0.75	1.57	1.88

are classified applying the best model obtained through the random forest processing of the training set.

Table 3 shows the literature types and the attribute values for the training-set sample, while Table 4 shows the literature types (when available), the predicted types and the attribute values for the 882 uncertain-type star sample. The predicted types are also compared to the literature types using a confusion-matrix type of display in Figs 10 and 11. Quite expectedly, the confusion is larger in these matrices, but the main confusion cases are the same as those observed with the training set (see Fig. 8 and the discussion of Section 4.7). Note that only the sub-set of stars with available uncertain types from the literature is incorporated in these figures.

8 CONCLUSIONS

The results presented in this paper show that it is possible to classify remarkably well the *Hipparcos* periodic variable stars into types that reflect their stellar physical properties. As detailed in Section 4.7, the main confusion cases are quite well understood. They originate from any of: (1) problems in extracting the correct period in the case of eclipsing binaries and *ELL* variables, (2) real similarities between different types of Cepheid stars, (3) a known, true difficulty for disentangling different types of blue variables, in particular the *SPB* and *ACV* stars. In any case, as seen in Fig. 8 the classification errors related to these confusion cases are generally below the 10 per cent level. This figure also shows that they are only a handful of additional confusion cases.

Similarly good results are obtained with the random forest methodology as with the multi-stage approach. Important advantages of random forest include a very useful attribute ranking method and a simple set-up and tuning. Is it also surprisingly robust to the presence of irrelevant or highly correlated attributes. The multi-stage approach allows a controlled selection of a particular classification algorithm and of a different attribute set at each node.

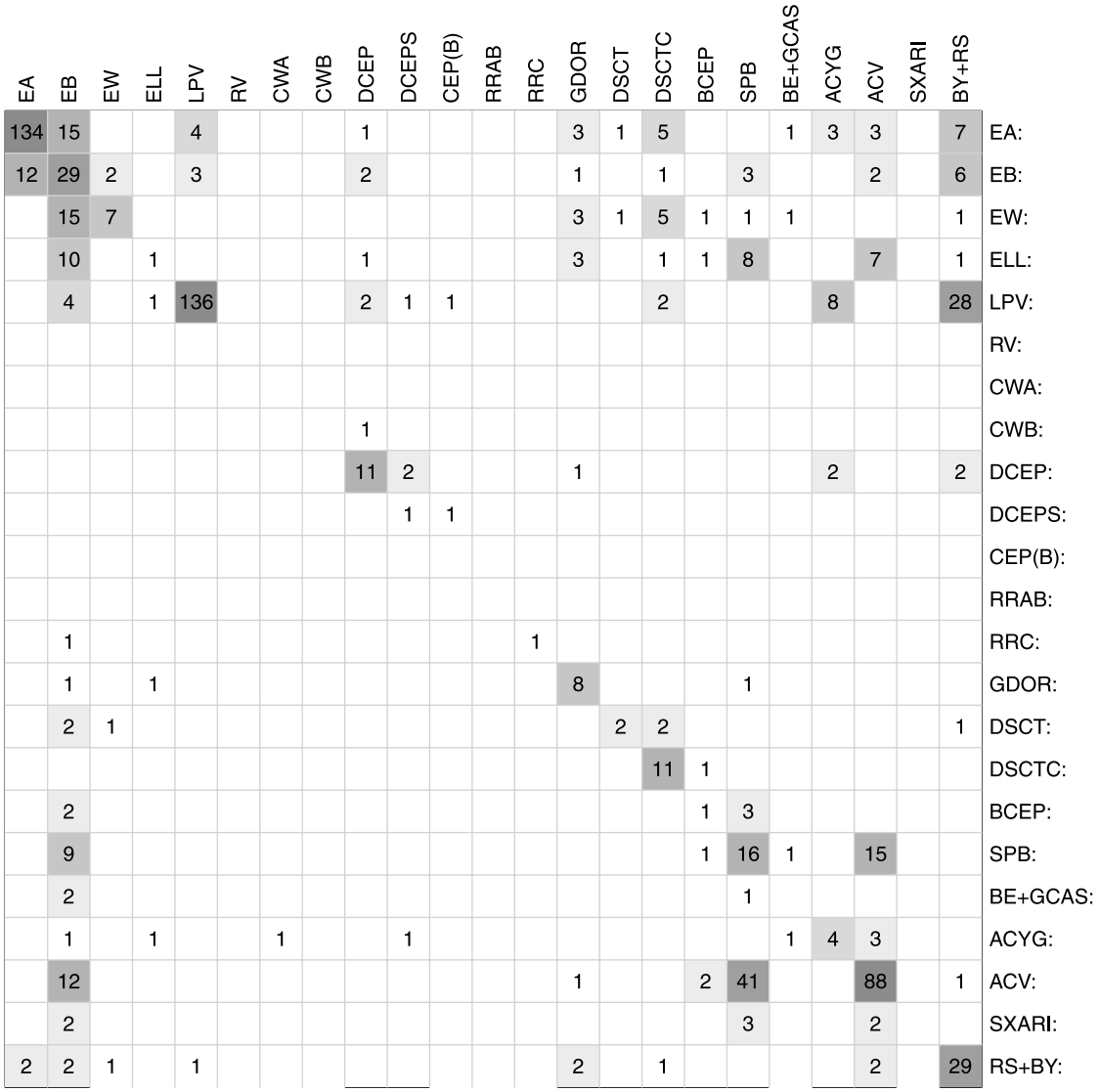

Figure 10. Confusion matrix obtained for a subset of *Hipparcos* stars with uncertain types from the literature. These types are shown in rows while columns indicate the types predicted by the random forest model derived from the training set analysis. Type labels are described in Table 1.

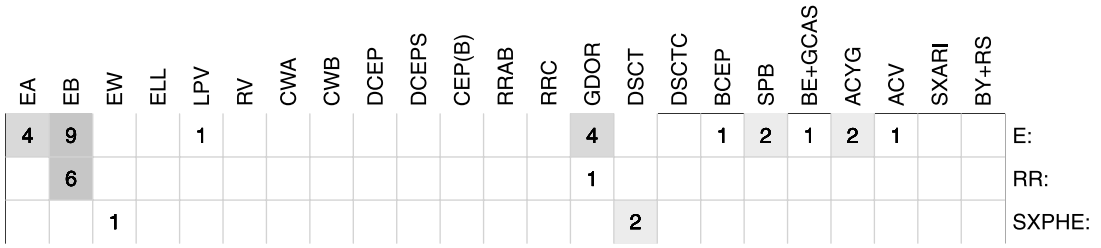

Figure 11. Same as Fig. 10 but for stars with types that do not directly match any of the types from training set stars.

Such choices can be useful in specific cases, but they also require more extensive and time-consuming optimization work.

Experience with various classification methods, random forest, multi-stage and other alternative methods, suggests that significant improvements are unlikely to come from better classification algorithms. Important progress rather can be expected through the introduction of new attributes which better reflect features of the physical processes responsible for the variability. They may even be specifically designed to disentangle some known cases of confusion. This is possible, for example, when additional independent data such as colour light curves or radial-velocity time series are available.

In addition to presenting the first systematic automated classification of the *Hipparcos* periodic variable stars, this paper describes the construction of a homogeneous training set of periodic stars. In a companion paper (Rimoldini et al. in preparation), this training set is completed with non-periodic variable stars. The complete training set can then be adapted to other surveys as a starting point for further

classification studies. Some challenging topics, such as variability detection, period search reliability and possible confusion between periodic and non-periodic types are deferred to subsequent investigations.

ACKNOWLEDGMENTS

We warmly thank Andy Liaw, the author of the random forest R implementation, for suggesting the optimization procedure used in Section 4.4. TL acknowledges support by the Austrian Science Fund FWF under project P20046-N16. We thank an anonymous referee for a careful review of this paper.

NOTE ADDED IN PROOF

While this paper was being submitted to MNRAS for publication, we became aware of a similar and interesting work by Richards et al. (2011). In their study, a number of classification methods are tested on a training set built from *Hipparcos* and OGLE data. They find that the random forest technique leads to the most accurate overall classification of variable stars, in good agreement with the results presented in this paper.

REFERENCES

Aerts C., Eyer L., Kestens E., 1998, A&A, 337, 790
Belokurov V., Evans N. W., Le Du Y., 2003, MNRAS, 341, 1373
Belokurov V., Evans N. W., Le Du Y., 2004, MNRAS, 352, 233
Blomme J. et al., 2010, ApJ, 713, L204
Blomme J. et al., 2011, MNRAS, in press (arXiv:1101.5038)
Bouckaert R. R., 2003, Choosing between Two Learning Algorithms Based on Calibrated Tests. International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, p. 51
Breiman L., 2001, Machine Learning, 45, 5
Burke E. W., Jr, Rolland W. W., Boy W. R., 1970, JRASC, 64, 353
Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159
Debosscher J. et al., 2009, A&A, 506, 519
Deeming T. J., 1975, Ap&SS, 36, 137
Diago P. D. et al., 2009, A&A, 506, 125
Dworetsky M. M., 1983, MNRAS, 203, 917
Eker Z. et al., 2008, VizieR Online Data Catalogue, 5128, 0
Eyer L., 1998, PhD thesis, Geneva Univ.
Eyer L., Blake C., 2002, in Aerts C., Bedding T. R., 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., Blake C., 2005, MNRAS, 358, 30
Hall M. et al., 2009, SIGKDD Explorations, 11, 10
Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York
Jurkevich I., 1971, Ap&SS, 13, 154
Lafler J., Kinman T. D., 1965, ApJS, 11, 216
Liaw A., Wiener M., 2002, R News, 2/3, 18
Lomb N. R., 1976, Ap&SS, 39, 447
Mardia K., Kent J., Bibby J., 1979, Multivariate Analysis. Academic Press, New York
Pearl J., 1988, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, Inc., San Francisco, CA
Pojmanski G., 2002, Acta Astron., 52, 397
Pojmanski G., 2003, Acta Astron., 53, 341
Renson P., 1978, A&A, 63, 125
Richards J. W. et al., 2011, preprint (arXiv:1101.1959)
Samus N. N. et al., 2009, VizieR Online Data Catalogue, 1, 2025
Sarro L. M., Debosscher J., Aerts C., López M., 2009, A&A, 506, 535
Scargle J. D., 1982, ApJ, 263, 835
Stellingwerf R. F., 1978, ApJ, 224, 953
Svetnik V. et al., 2004, in Roli F., Kittler J., Windeatt T., eds, Lecture Notes in Computer Science, Vol. 3077. Springer, Berlin, p. 334
Waelkens C., Aerts C., Kestens E., Grenon M., Eyer L., 1998, A&A, 330, 215
Watson C., Henden A. A., Price A., 2010, VizieR Online Data Catalogue, 1, 2027
Willemsen P. G., Eyer L., 2007, preprint (arXiv:0712.2898)
Zechmeister M., Kürster M., 2009, A&A, 496, 577

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table 3. The *Hipparcos* training set star list with literature types and attribute values.
Table 4. Results obtained for the *Hipparcos* stars excluded from the training set.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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