THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Calibrating Long Period Variables as Standard Candles with Machine Learning

Citation for published version:
Rau, MM, Koposov, SE, Trac, H & Mandelbaum, R 2019, 'Calibrating Long Period Variables as Standard Candles with Machine Learning', *Monthly Notices of the Royal Astronomical Society* .
https://doi.org/10.1093/mnras/sty3495

Digital Object Identifier (DOI):
10.1093/mnras/sty3495

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Monthly Notices of the Royal Astronomical Society

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 23. Jul. 2021

MNRAS **000**, 1–14 (2015) Preprint 9 April 2019 Compiled using MNRAS LaTeX style file v3.0

arXiv:1806.02841v2 [astro-ph.IM] 5 Apr 2019

Calibrating Long Period Variables as Standard Candles with Machine Learning

Markus Michael Rau[1]*, Sergey E. Koposov[1,2], Hy Trac[1], Rachel Mandelbaum[1]

[1] *McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213*

[2] *Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Variable stars with well-calibrated period-luminosity relationships provide accurate distance measurements to nearby galaxies and are therefore a vital tool for cosmology and astrophysics. While these measurements typically rely on samples of Cepheid and RR-Lyrae stars, abundant populations of luminous variable stars with longer periods of $10-1000$ days remain largely unused. We apply machine learning to derive a mapping between lightcurve features of these variable stars and their magnitude to extend the traditional period-luminosity (PL) relation commonly used for Cepheid samples. Using photometric data for long period variable stars in the Large Magellanic cloud (LMC), we demonstrate that our predictions produce residual errors comparable to those obtained on the corresponding Cepheid population. We show that our model generalizes well to other samples by performing a blind test on photometric data from the Small Magellanic Cloud (SMC). Our predictions on the SMC again show small residual errors and biases, comparable to results that employ PL relations fitted on Cepheid samples. The residual biases are complementary between the long period variable and Cepheid fits, which provides exciting prospects to better control sources of systematic error in cosmological distance measurements. We finally show that the proposed methodology can be used to optimize samples of variable stars as standard candles independent of any prior variable star classification.

Key words: distance scale – cosmology: observations – stars: variables – Magellanic Clouds

1 INTRODUCTION

One of the most important aspects of modern cosmology and astrophysics is the measurement of accurate distances. Variable stars like Cepheids that exhibit tight relationships between their oscillation period and luminosity are among the primary tools to measure distances in the local universe. In the advent of precision cosmology, Cepheid distances are a vital rung in the cosmological distance ladder and calibrate local Type Ia supernovae (SNIa) samples. These SNIa distance measurements (e.g. Conley et al. 2011; Riess et al. 2016) provide an absolute distance scale in the low redshift universe, that complement constraints on the CMB sound horizon scale (Planck Collaboration et al. 2016a) at the high redshift border of the visible universe. The mild tension between both distance scales that currently persists in the literature (e.g. Riess et al. 2016; Zhang et al. 2017; Feeney et al. 2018) could require new theoretical interpretations of these complementary distance scales. Given sufficient observational evidence, this can motivate considering non-standard extensions to the cosmological model like the introduction of sterile neutrinos, dynamical dark energy or a nonzero curvature component (e.g. Wyman et al. 2014; Dvorkin et al. 2014; Leistedt et al. 2014; Planck Collaboration et al. 2016b; Di Valentino et al. 2016; Bernal et al. 2016; Zhao et al. 2017; Solà et al. 2017). A primary challenge in the field is therefore to clarify if the observed tensions are significant signs of new physics or caused by observational systematics.

Cepheid distances, which provide the primary calibration for local supernovae samples, are subject to a variety of potential observational and methodological systematics (e.g. Freedman et al. 2001; Conley et al. 2011; Efstathiou 2014; Kodric et al. 2015; Riess et al. 2016; Zhang et al. 2017; Feeney et al. 2018) such as inaccurate photometric calibration, metallicity differences between anchor samples or biases introduced by the treatment of outliers in fits of the period-luminosity (PL) relation. In addition, sufficiently large anchor samples for Cepheids are only available in ~ 9 nearby galaxies (Zhang et al. 2017). As a result, ≈ 30% of

* E-mail: markusr@andrew.cmu.edu

© 2015 The Authors

the total error budget on local H_0 measurements is related to the Cepheid distance calibration (see Riess et al. 2016, Fig. 1). Exploring additional sources of distance calibration for local supernovae measurements is therefore an interesting avenue to better control sources of systematic errors in H_0 measurements. For example Huang et al. (2018) recently proposed the tight period luminosity relation of oxygen-rich Mira variables as an additional rung in the cosmological distance ladder. Distance measurements that use variable stars exploit the tight relationship between period, metallicity and the luminosity of Cepheids, RR-Lyrae and Mira variables to calibrate them as cosmological standard candles. However, there exist a large variety of other luminous variable stars, like OGLE Small Amplitude Red Giant stars (OSARG), that also exhibit a variety of overlapping sequences in PL space.

The goal of this paper is to investigate how machine learning can be used to exploit these variable stars for cosmological distance measurements. This is facilitated by the rich feature set in variable star lightcurves, typically used in the context of variable star classification (e.g. Richards et al. 2011; Dubath et al. 2011; Palaversa et al. 2013; Kügler et al. 2015; Armstrong et al. 2016; Sesar et al. 2017; Naul et al. 2018), that contains much more information about the variable star luminosity than just the first dominant period.

This paper is structured as follows. In §2 we will describe the photometric data and the lightcurve features used in this analysis. §3 describes our methodology, the metrics used, and additional analysis steps such as outlier rejection. §4 details our analysis and §5 closes with a discussion and summary of our results.

2 DATA

We use star catalogs from the third public release of the Optical Gravitational Lensing Experiment (OGLE) survey (e.g. OGLE 2018; Soszyński et al. 2008, 2009, 2010, 2011), which provides photometric lightcurves of the Large Magellanic Cloud and Small Magellanic Cloud in the I and V bands. The catalog contains variable star classifications, mean lightcurve magnitudes in the I and V filters and the three primary periods P_{1-3} and amplitudes A_{1-3} extracted from the I band lightcurves by Fourier decomposition as described in Soszynski et al. (2004). We use the fundamental mode Cepheid and the Long Period variables (LPV) in the full catalog, where the LPV sample consist of semi-regular variables (SRV), Mira stars and OGLE Small Amplitude Red Giant (OSARG) variables. In the following we will further distinguish stars on the Red Giant Branch (RGB) and the Asymptotic Giant Branch (AGB). We correct the I band magnitude for reddening using the optical Wesenheit index (e.g. Madore 1982; Schlegel et al. 1998; Soszynski et al. 2007)

$$W_I = I - 1.55 \cdot (V - I), \tag{1}$$

where I and V denote the mean apparent magnitudes in the I and V band respectively. We restrict the Wesenheit range of the LPV sample to $3 < W_I < 16$ and use $|W_I| < 17.5$ for the fundamental mode Cepheid sample, which is slightly fainter than the LPV population. Beyond these Wesenheit magnitude limits we only find LPV and fundamental mode Cepheid stars that lack magnitude measurements in the V or I band (values are set to -99.99) or a total of 14 (7) very faint LPV in the LMC (SMC) with up to $W_I \lesssim 20$.

Fig. 1 plots the first period against the optical Wesenheit index (or ‘P-W relation’), for the different types of variable stars in the selected catalog. The left panel shows the P-W relations for the population in the Large Magellanic Cloud (LMC). Besides the population of fundamental mode Cepheids (F-Cepheids), we see a large sample of bright variable stars with longer periods covering two orders of magnitude from 10 to 1000 days. The sample size of each stellar population is shown in the legend. The largest population, ≈ 80.000 variable stars, are OGLE Small Amplitude Red Giant stars (OSARGs). Notably, the OSARG sample is significantly brighter than the corresponding Cepheid sample for $P > 20$ days. The right panel of Fig. 1 shows the P-W relation of OSARGs in the LMC (blue) and SMC (red). Besides the shift towards fainter magnitudes due to the distance modulus between the two galaxies, we also note that the stellar population is significantly different. While both LMC and SMC OSARGs follow the same characteristic P-W sequences, the SMC OSARGs do not cover the full range of magnitudes and periods that would be expected from the LMC population. Furthermore the P-W relations in the SMC appear slightly tilted compared with the LMC sequences. A likely explanation for this discrepancy could be differences in the metallicity between both stellar populations.

2.1 Features: Periods

The overlapping P-W sequences that are in the current literature referred to as OSARG, Mira and SRV stars were initially described by Wood et al. (1999) and will be referred to as ‘Wood sequences’. They correspond to modes of stellar oscillation and can be related to clusters of period ratios. These can be efficiently analysed using so-called Petersen diagrams, shown in Fig. 2. This diagram, applied earlier by e.g. Soszynski et al. (2004), shows the distribution of pulsators in the space of longest period vs. shorter period to long period ratio. The left (right) panel shows the Petersen diagram for the OSARG population in the LMC (SMC). Soszynski et al. (2004) studied the OSARG population in the Magellanic Clouds and showed that the clusters in the Petersen diagram can be associated with the Wood sequences and thus to the type of stellar pulsation mode. Period ratios around 0.97 to 1.0 can be associated with nonradial oscillations that can also be found in Cepheids[1] and RR-Lyr (Soszynski et al. 2004, and references therein).

The lower right clump around $P_L \approx 200$ days can be associated with Long Secondary Period (LSP) oscillations, whose origin is still an area of active research (see Trabucchi et al. 2017, and references therein). The near horizontal sequences visible in the Petersen diagram are related to the aforementioned Wood sequences in P-W space and thus correspond to characteristic period ratios of stellar oscillations. Comparing the left and right panels of Fig. 2, we see that the basic structure of the Petersen diagram is the same in both

[1] We refer to Smolec et al. (2017) for a more detailed discussion of Petersen diagrams for Cepheids and RR-Lyr.

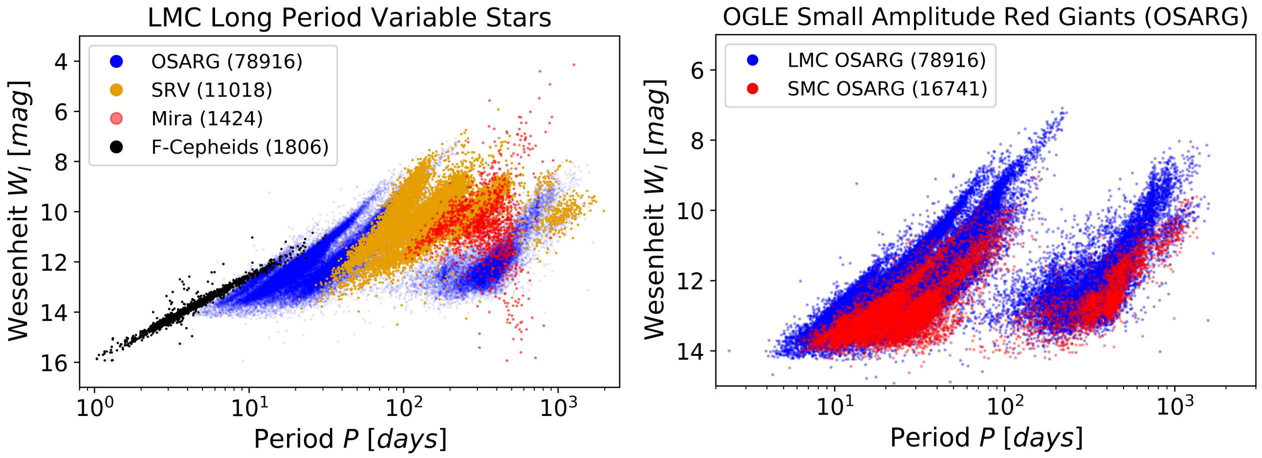

Figure 1. Overview of the Period-Wesenheit (P-W) relations for the different subsamples used in this analysis. The legends list the number of stars in the respective subsamples in brackets. *Left*: Fundamental Mode Cepheids (F-Ceph), Mira stars (Mira), semi-regular variables (SRV) and OGLE Small Amplitude Red Giant stars (OSARGs) in the Large Magellanic Cloud (LMC). *Right*: P-W relation of OSARG variables in the LMC and the Small Magellanic cloud (SMC).

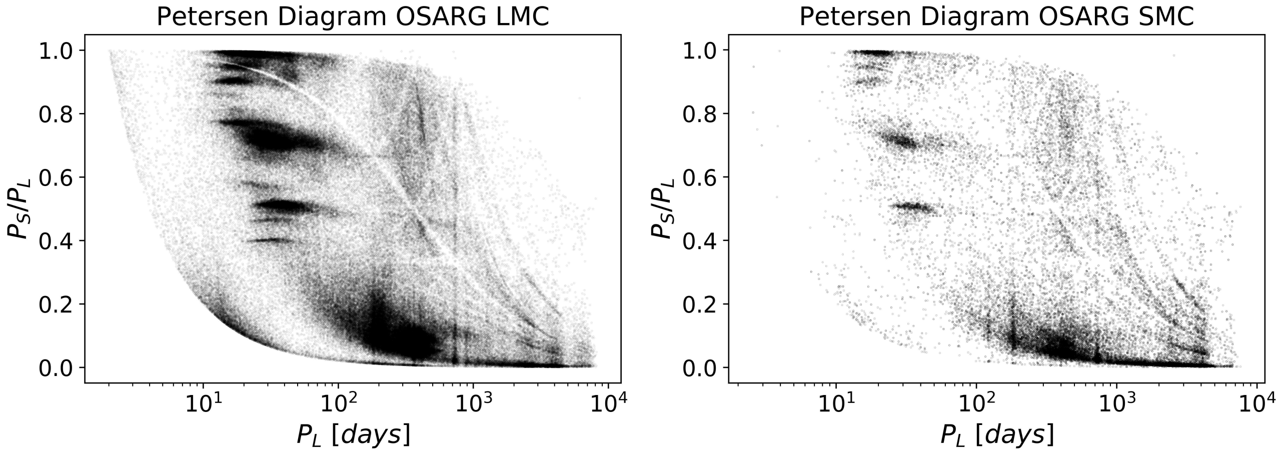

Figure 2. Petersen diagram of the full OSARG AGB and RGB star samples in the LMC (left panel) and SMC (right panel). Forming all combinations from the three primary periods, we plot the longer period P_L against the ratio between shorter and longer period P_S/P_L. A star can therefore appear up to 3 times in the Petersen diagram, depending on the number of measured periods.

the LMC and SMC samples. While the number of OSARG stars is lower in the SMC than in the LMC, the horizontal clusters of stars with similar period ratios at $P_L < 10^2$ are populated in both diagrams. We refer to Soszynski et al. (2004) for a more detailed analysis of Petersen diagrams of OSARG variables in the Magellanic Clouds.

The left panel of Fig. 3 shows the P-W relation for the RGB and AGB populations of LMC OSARGs. As expected, the AGB stars extend towards brighter magnitudes, but largely cover the same Wood sequences as the RGB population, in agreement with Kiss & Bedding (2003); Soszynski et al. (2004). We note a small period shift between the RGB and AGB samples, which can be explained by the fact that the characteristic oscillation period P for solar oscillations scales with effective temperature $T_{\rm eff}$, stellar mass M and luminosity L as $P \sim L/(M\,T_{\rm eff}^{3.5})$ as given in Stello et al. (2007). At constant luminosity the AGB stars will have a higher effective temperature compared to RGB stars. This temperature difference induces a small period shift between samples of AGB and RGB stars (Kiss & Bedding 2003; Soszynski et al. 2004).

We conclude that the first three dominant periods contain information about the type of stellar oscillation, the position on the respective Wood sequences as well as the evolutionary state of the giant star. The luminosity information contained in the multiple oscillation periods of RGB stars was already exploited for distance measurements by e.g. Tabur et al. (2010). Accordingly the multiple oscillation periods can be expected to be important features in our Machine Learning approach.

2.2 Features: Amplitudes

The right panel of Fig. 3 plots the first amplitude A_1 against the Wesenheit magnitude W_I for the AGB and RGB OSARG variables. We plot the range between the 0.05 and

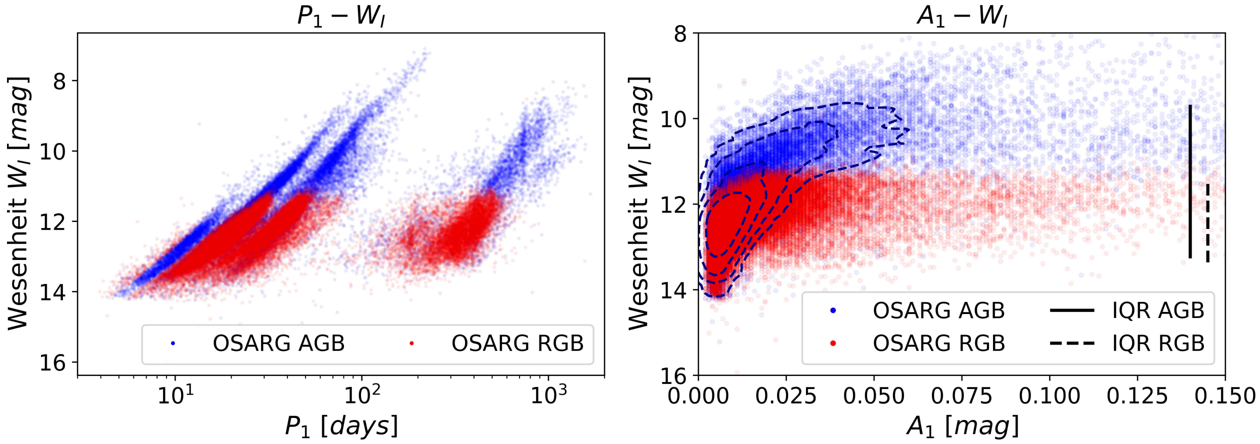

Figure 3. Differences in the primary period-amplitude distribution between the OSARG AGB and OSARG RGB subsamples. *Left*: Primary period P_1 against Wesenheit W_I distribution for the AGB (blue) and RGB (red) subsamples. *Right*: Primary amplitude A_1 against Wesenheit W_I distribution for the AGB (blue) and RGB (red) subsample. The vertical lines show the inter quantile range of the W_I distribution between the 0.05 and 0.95 quantiles of the respective subsamples. The blue dashed density contours highlight the A_1 against W_I distribution for OSARG AGB variables, that is partly overplotted by the OSARG RGB points.

0.95 quantile, i.e. the inter-quantile range **IQR**, for the AGB (RGB) population as solid (dashed) vertical lines and highlight the amplitude-Wesenheit distribution of the OSARG AGB population with density contours. The amplitude and the Wesenheit are anti-correlated, where AGB and RGB OSARG samples with smaller amplitudes extend towards fainter magnitudes as to be expected from the peak amplitude scaling $A \sim L/(M T_{\rm eff}^2)$ predicted by Stello et al. (2007). As discussed in Trabucchi et al. (2017), the observed amplitudes are related to the growth rate of the stellar oscillation modes and can therefore be expected to help in distinguishing between different modes of stellar oscillation.

From these considerations we conclude that the inclusion of the first three amplitudes into the feature set is well motivated by both their correlation with the luminosity of the star and by their connection to the growth rates of the respective stellar oscillation modes.

3 METHODOLOGY

As noted previously, our goal is to select samples of long period variables for which we can obtain accurate Wesenheit predictions. In this way we are able to jointly optimize the sample selection, i.e. the identification of ‘standard candle-like’ stars, and make accurate predictions of the Wesenheit magnitudes based on the high dimensional feature set. The following analysis uses the lightcurve features discussed in §2.1 and 2.2:

- the primary P_1, secondary P_2 and tertiary P_3 oscillation periods;
- the corresponding amplitudes A_1, A_2 and A_3 of the oscillations.

The full information of the mapping between the lightcurve features $\mathbf{f}$ and the apparent Wesenheit index W_I is contained in the conditional probability density function (pdf) $p(W_I|\mathbf{f})$. Once this distribution is estimated for each variable star in the catalog, we define a statistic of this distribution that will be used to obtain Wesenheit predictions for a given variable star. In the classical regression setting this is the conditional mean, which is also our choice, but alternatives such as the conditional median can be justified if $p(W_I|\mathbf{f})$ is expected to exhibit wide wings, due to a significant fraction of outliers in the data. The estimated distribution $p(W_I|\mathbf{f})$ is a convolution of a pdf that describes the photometric error in W_I, the intrinsic error of the data[2] and other effects like attenuation bias from inconsistencies in the errors on input features across different datasets or inaccuracies in the machine learning algorithm. Since the OGLE photometry is of exceptional quality ($S/N > 1500$) and very similar in both the LMC and SMC, we do not incorporate the photometric error or attenuation bias into the modeling and assume that $p(W_I|\mathbf{f})$ is dominated by the intrinsic error in the data.

As a selection criterion of high-quality standard candles we use the standard deviation $\sigma(W_I|\mathbf{f})$ of the conditional pdf $p(W_I|\mathbf{f})$ and select only those objects for which this quantity is small. This is motivated by our previous decision to use the mean of $p(W_I|\mathbf{f})$ as the regression statistic. The procedure therefore essentially approximates the conditional pdf $p(W_I|\mathbf{f})$ as a normal distribution. The mean is used to predict the Wesenheit magnitude of the variable star and the standard deviation allows us to select those stars that are expected to occupy regions in feature space where the most accurate predictions are possible. While the analysis presented in this paper only relies on estimates for two statistics of the conditional pdf, the general method that we demonstrate constructs the full shape of this distribution. This enables a possible extension to alternative point statistics like the conditional median.

[2] Here, the intrinsic error refers to the standard deviation of the conditional pdf obtained by a perfect estimator on noiseless data. This error depends only on the intrinsic information in the lightcurve features to predict the Wesenheit, but not on inaccuracies in the estimator or data.

3.1 Conditional Density Estimation

In this section we describe our machine learning methodology to construct an estimate for the conditional pdf $p(W_I|\mathbf{f})$ from which the mean and standard deviation are derived. To avoid overfitting, we split the available data randomly into two disjunct subsamples; the training and test set. The model is then fitted, or trained, on the training set and subsequently applied to the disjunct test set. This work will use 90% of the data to train the model and 10% as test data, in a so-called '10-fold cross validation' procedure described in §3.3.

To estimate the conditional pdf $p(W_I|\mathbf{f})$, we discretize the Wesenheit index of the training set into 300 equally spaced bins.[3] In this way we reformulate the regression problem of predicting the continuous Wesenheit index W_I as a classification problem (e.g. Schapire et al. 2002; Frank & Bouckaert 2009; Rau et al. 2015). The model will then estimate probabilities of Wesenheit-bin membership. These probabilities can then be combined into a histogram that is an estimate of the conditional pdf $p(W_I|\mathbf{f}_i)$ for each variable star i in the sample.

It is convenient to express this estimate as a weighted sum over the Wesenheit magnitudes W_I^i of the stars i in the training set. The Wesenheit interval $j \in [1, n_{\rm bins} = 300]$ into which the star i falls is denoted as $\mathcal{I}_j$. Denoting the bin probability of bin j as $\mathcal{P}_j$, we can define a weight $w_i(\mathbf{f})$ for each variable star in the training set as

$$w_i(\mathbf{f}) = \sum_{j=1}^{n_{\rm bins}} \left(\frac{\mathcal{P}_j}{n_j}\right) \Theta(W_I^i \in \mathcal{I}_j), \qquad (2)$$

where n_j is the number of all variable stars in the training set that fall into bin $\mathcal{I}_j$ and W_I^i denotes the Wesenheit magnitude of the training set star. For a variable star in the test set with Wesenheit magnitude W_I and feature vector $\mathbf{f}$, we can write the conditional distribution $p(W_I|\mathbf{f})$ as

$$p(W_I|\mathbf{f}) = \sum_{i=1}^{N} w_i(\mathbf{f}) \sum_{j=1}^{n_{\rm bins}} \frac{\Theta(W_I \in \mathcal{I}_j)\,\Theta(W_I^i \in \mathcal{I}_j)}{r_j}, \qquad (3)$$

where $N = \sum_{j=1}^{j=n_{\rm bins}} n_j$ denotes the number of variable stars in the training set and r_j is the width of bin $\mathcal{I}_j$. The boolean function $\Theta(x)$ is 0 if its argument is false and unity if it is true. Accordingly $\Theta(W_I \in \mathcal{I}_j)\,\Theta(W_I^i \in \mathcal{I}_j)$ is unity if both W_I^i and W_I are in bin $\mathcal{I}_j$ and 0 otherwise. Note that these weights are a function of the features $\mathbf{f}$ of the respective variable star in the test set, as the bin probabilities will depend on the position in feature space.

The conditional mean $\langle W_I|\mathbf{f}\rangle$ can be estimated on the weighted training set as

$$\langle W_I|\mathbf{f}\rangle = \sum_{i=1}^{N} w_i(\mathbf{f})\, W_I^i\,, \qquad (4)$$

and the conditional standard deviation $\sigma(W_I|\mathbf{f})$ as

$$\sigma(W_I|\mathbf{f}) = \sqrt{\sum_{i=1}^{N} w_i(\mathbf{f})\, \left(W_I^i - \langle W_I|\mathbf{f}\rangle\right)^2} \qquad (5)$$

In §4, we rank the variable stars in order of increasing conditional standard deviation and select a subsample that contains only stars that are expected to yield very accurate predictions of their Wesenheit magnitude given their feature vectors $\mathbf{f}$.

3.2 The Random Forest

In the following we describe the Random Forest[4] classifier (Breiman 2001) that we use to estimate the bin probabilities $\mathcal{P}_j$.

Given a training set of variable stars with known combinations of features $\mathbf{f}_i$ and Wesenheit indices W_I^i, the algorithm starts by forming N_T number of bootstrap realizations of this training set. In the process of bootstrapping, we randomly select N elements from the original dataset with replacement, where N denotes the sample size of the original dataset. On each of these bootstrap realizations a single decision tree is fitted. The index T identifies a particular tree in the Random Forest. The Random Forest predicts bin probabilities by averaging the bin probabilities estimated by all N_T decision trees in the Random Forest.

A single decision tree is a binary partitioning tree that is recursively grown on the bootstrapped dataset by selecting splits such that the newly formed partitions are optimized to contain only training set stars of high similarity. Each grown partition with index $\tau \in [1, N_\tau]$ corresponds to a region in input space $\mathcal{R}_\tau$. We denote N_τ as the total number of partitions, or 'leaf nodes' of the tree, and n_τ as the number of training set elements in partition τ. The probability that a variable star whose features fall into a region $\mathbf{f} \in \mathcal{R}_\tau$ has a Wesenheit index in bin j, $W_I \in \mathcal{I}_j$, is given as the fraction of training set elements in $\mathcal{R}_\tau$ that fall into $\mathcal{I}_j$

$$\mathcal{P}_{\tau,j} = \sum_{i=1}^{n_\tau} \frac{\Theta(W_I^i \in \mathcal{I}_j)}{n_\tau}\,, \qquad (6)$$

where the sum runs over all n_τ training set elements in region $\mathcal{R}_\tau$. Similarity can be optimized by minimizing the Gini criterion $\mathcal{G}_\tau$ in region $\mathcal{R}_\tau$:

$$\mathcal{G}_\tau = \sum_{j=1}^{n_{\rm bins}} \mathcal{P}_{\tau,j}(1 - \mathcal{P}_{\tau,j})\,. \qquad (7)$$

Note that $\mathcal{G}_\tau$ vanishes if all training set elements in region $\mathcal{R}_\tau$ reside in the same Wesenheit bin, and is maximal if they are equally distributed across the Wesenheit bins. In each recursion step we select a binary split in feature space such, that the summed $\mathcal{G}_\tau$ over all regions $\mathcal{R}_\tau$ is minimized compared with the previous state. The splitting stops if a minimum number of training set objects are located in the respective region, which is a tuning parameter of the model.

If a new variable star is queried down the tree into region $\mathcal{R}_\tau$, the tree returns the bin probabilities $\mathcal{P}_{\tau,j}$, as defined in Eq. 6, for each Wesenheit bin $\mathcal{I}_j$. For a more in-depth description of the Random Forest algorithm, we refer the interested reader to the literature (e.g. Hastie et al. 2001; Bishop 2006).

[3] The results are not very sensitive to this choice, but choosing an overly-coarse binning can lead to biased estimates.

[4] We use the implementation provided by the scikit-learn package (Pedregosa et al. 2011), using the default parameters from the 'Random Forest Classifier'.

3.3 Evaluation and Metrics

The analysis presented in §4, applies the Random Forest classifier to each variable star in the considered sample. To ensure that the trained model generalizes well to unseen data, we iterate the splits into training and test set using the k-fold cross validation technique. The complete dataset is randomly split into k non-overlapping, equal-sized parts, where we use the common choice $k = 10$. The model is then trained on all but the first of these partitions and subsequently applied to the held-out partition. The procedure then continues in the same manner with the remaining partitions, where each partition is held out once. In this way we generate an estimate of the conditional density $p(W_I|\mathbf{f})$ for each star in the sample in $k = 10$ chunks. We can then evaluate the performance metrics, such as accuracy of the mean and variance.

We measure the prediction quality using the root mean squared error between the true Wesenheit magnitude $W_{\rm true}$ and the predicted Wesenheit magnitude $W_{\rm pred}$ as

$$\mathrm{RMSE} = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(W_{\rm true,i} - W_{\rm pred,i}\right)^2}, \qquad (8)$$

and the mean error (ME) in the prediction as

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}\left(W_{\rm true,i} - W_{\rm pred,i}\right). \qquad (9)$$

To accurately use variable stars for distance measurements, an important quality requirement is that the variation in the slope of the regression function is small across the several distance anchors (see e.g. Zhang et al. 2017). To compare the accuracy of the prediction across samples, we consider the linear regression function

$$W_{\rm true} = b\, W_{\rm pred} + \mu, \qquad (10)$$

where μ represents the offset of the linear regression fit and $W_{\rm pred}$ ($W_{\rm true}$) the predicted (true) Wesenheit indices. Since the offset in these relations is connected with the distance modulus between both samples, we compare the similarity between both relations in terms of the relative bias in the slope b between a reference sample $\mathcal{R}$ and a comparison sample C:

$$b_\Delta = \left(b_{\mathcal{R}} - b_C\right)/b_{\mathcal{R}}\,. \qquad (11)$$

3.4 Outlier Removal

Outlier removal is an essential part of accurately using period-luminosity relations of variable stars for distance measurements. A common technique (Zhang et al. 2017) removes outliers that deviate more than $\alpha\,\sigma$ from the linear $W_{\rm pred}$-$W_{\rm true}$ regression line. In this work we use $\alpha = 2.25$, which is a common choice used in the literature (Zhang et al. 2017). Specifically, we iteratively repeat $\alpha\,\sigma$ outlier rejection and linear regression of non-rejected datapoints until we reach convergence.

4 ANALYSIS AND RESULTS

In the following section we apply the methodology described in §3 to the OGLE photometric variable star catalogs of the Magellanic Clouds. We rank the long period variable stars in these samples based on our estimates of their conditional standard deviation (Eq. 5). We select variable stars where these values are small, and for which therefore the lightcurve parameters are highly predictive of their Wesenheit. We then evaluate the metrics defined in §3.3 on these selections on the selected subsamples in the LMC and the SMC catalogs.

4.1 Large Magellanic Cloud

In this section we investigate how the accuracy of the Random Forest predictions improve when we optimize the sample selection using the methodology described in §3. To this end, we use the 10-fold cross validation procedure to obtain predictions of the conditional mean (Eq. 4) and conditional standard deviation (Eq. 5) for the variable star samples in the LMC. For each of the 10 equal-sized folds, we select the variable stars based on the smallest estimated conditional standard deviation and evaluate the performance metrics on this selection. Thus for each selected sample size we obtain 10 cross validated performance estimates. This procedure is separately applied to the full sample of long period variables, and to four subclasses of variable stars: AGB/RGB OGLE Small Amplitude Red Giants, Mira variables and semi-regular variables (SRV). The left (right) panel of Fig. 4 shows the root mean squared RMS error (mean error ME)[5] as a function of the selected sample size, i.e. the number of variable stars that remain in the sample after the selection by conditional standard deviation $\sigma(W|f)$ as described above. The solid mean line shows the average performance across the $k = 10$ equal sized folds and the error contours the standard deviation across them. We quote the combined sample size of the 10 folds on the horizontal axis and divide the width of the 1σ errorbars by $\sqrt{k}$ to correct for the increased sample size.

To compare with the performance of a well-known standard candle, we show these metrics, obtained in the same manner, for the fitted Period-Wesenheit (P-W) relation of the LMC Cepheids as a grey band. The black bullet corresponds to the sample size of this Cepheid sample. We note that the AGB and RGB OSARGs show small RMSE values that are consistent with the performance of the Cepheid P-W relation within the statistical errors, while having a factor of three times larger sample size.

Fig. 4 also shows the results of our method when applied to the full LPV sample ('All LPV') with all its subclasses, i.e. Mira, SRV and OSARG variables. The performance in this case is quite similar to AGB OSARG and RGB OSARG. Comparing with the performance of e.g. the SRV population at constant sample size, we see that our methodology robustly identifies subpopulations of LPVs with especially tight conditional pdfs purely based on light curve features, without a prior step of variable star classification.

We highlight this point in the following analysis: we use the 'All LPV' sample, i.e. the sample of all long period variables in the LMC, and investigate how the fraction of different types of LPV variables changes, when we select

[5] The root mean squared error (RMSE) and the mean error (ME) are defined in Eq. 8 and Eq. 9 respectively.

subsamples using our proposed method. We measure this change as the fractional population difference:

$$\Delta_{\mathrm{frac,type}} = \frac{N_{\mathrm{select}}/N_{\mathrm{select,total}}}{N_{\mathrm{orig}}/N_{\mathrm{orig,total}}} \,. \tag{12}$$

Here N_{select} denotes the number of stars of a given type after the selection and $N_{\mathrm{select,total}}$ denotes the total number of stars, including all types, after the selection. N_{orig} denotes the number of stars of a given type before the selection, i.e. for the full sample, and $N_{\mathrm{orig,total}}$ denotes the total number of stars for the full sample, including all types. This quantity therefore measures how the fraction of a given type of variable stars changes in the sample, when we apply a more restrictive selection.

Fig. 5 plots $\Delta_{\mathrm{frac,type}}$ as a function of the selected sample size. We see that the fraction of MIRA and SRV variable stars decreases significantly, while the fraction of OSARG AGB and OSARG RGB stars remains constant and even increases. We note that the horizontal line shows the expectation of a random selection, where the fraction of LPV types remains constant. The 1σ errorbars show the deviation between the 10 cross validation folds.

Compared with the performance of the OSARG AGB/RGB subsamples, we note that Mira or SRV variables show a substantially larger RMSE, due to a larger luminosity range at fixed period compared with OSARG variables. We expect however that the inclusion of deep infrared photometry in the K band (Whitelock et al. 2008; Yuan et al. 2017) will improve the residuals obtained for the Mira sample, as Miras follow a well defined P-W relation in this wavelength range (see e.g. Huang et al. 2018). Similarly the mean errors obtained for those samples that show a small RMSE, i.e. the AGB OSARG, RGB OSARG and the full LPV sample, are consistent with the results obtained using Cepheids, albeit having a small bias towards underpredicting the true Wesenheit. We note that this result is stable across all considered sample sizes, where the width of the error decreases with the inclusion of more stars to the sample. For large sample sizes of $N > 10.000$ stars, the error on the mean error (ME), shown as 1σ contours, obtained on the LPV samples is significantly smaller than the one for Cepheids. At the sample size of the Cepheid sample, these errorbars are comparable between Cepheids and LPVs.

The left panel of Fig. 6 color codes the selected long period variable stars in the P-W relation. We see that the best 2000 long period variables[6], cluster around the faint part of the respective Wood sequences, in a tight correlation between log period and Wesenheit. This result again highlights that the algorithm is able to effectively select samples of variable stars solely based on the provided input features without the need of a prior classification step. In addition we note that the selected sample is on average significantly brighter than the Cepheid sample. While the Cepheid sample has Wesenheit of $11 \lesssim W_I \lesssim 16$, the selected LPV sample span a range of $11 \lesssim W_I \lesssim 13$ (see Fig. 6). Their brightness makes these LPV stars attractive as potential standard candles for cosmological distance measurements. We note that imposing a less stringent sample selection on $\sigma(W|\mathbf{f})$ will allow more and even brighter objects into the sample. We discuss this in more detail in §4.3. The right panel of Fig. 6 plots the true Wesenheit against the difference between predicted and true Wesenheit for both the full sample and our selection. We see that the performance of the selected, or optimized, sample is much better compared with the full sample. We also note that the multimodality in Wesenheit magnitude at a given period as shown in Fig. 6, is removed and the Wesenheit distribution as estimated by the Random Forest algorithm given the full feature set is now more peaked and unimodal.

4.2 Small Magellanic Cloud

In the previous analysis we trained our model exclusively on LMC data using the 10-fold cross validation procedure. In order to test the robustness of our model on other datasets, we now train our model on LMC data and apply it to the corresponding SMC datasets. Since we demonstrated in the previous section that the OSARG stars are the best-performing subpopulation of LPV in the LMC, we concentrate in this section on this subtype for simplicity. In the following section we will train separately on the AGB and RGB subpopulation.

We use the Random Forest models trained on the 10 cross validation folds of AGB/RGB OSARG variables in the LMC and query the corresponding SMC datasets. We therefore obtain 10 sets of predictions for the SMC AGB and SMC RGB population respectively, each corresponding to a cross validation fold in the LMC. The variance between these 10 predictions therefore quantifies the variance in the training of the Random Forest. Since the Random Forest is trained on the LMC data, the predicted Wesenheit magnitudes obtained on SMC data will be biased low due to the distance modulus between LMC and SMC. We can measure this offset by fitting linear regression functions to the selected relations between predicted W_{pred} and true Wesenheit W_{true} for both the LMC and SMC samples. To mimic a typical distance measurement procedure using variable stars, we apply the outlier removal algorithm described in §3.4 to each set of SMC predictions and to each of the 10 cross-validated LMC prediction sets. In analogy to the previous section, we compare the performance of our methodology with the performance of the P-W fits of the fundamental mode Cepheid sample in the LMC and SMC. To make a comparison easier we do not consider the P-W relation directly but instead consider the predicted Wesenheit indices from the fitted P-W relation.

In analogy to the Machine Learning (ML) approach, we fit the P-W relation on the LMC and apply the model to the corresponding SMC data. Note that the outlier rejection and 10-fold cross validation[7] are analogously applied to the Cepheid sample to obtain error contours in the quoted metrics. The outlier rejection algorithm will remove a certain fraction of the data after the cut on the conditional standard deviation is performed. We will refer to the sample

[6] For this sample size, the 'All LPV' sample in Fig. 4, is roughly of comparable size to the Cepheid sample and has similar prediction accuracy.

[7] Our fundamental mode Cepheid sample in the LMC (SMC) contains 1806 (2603) stars from which 89 ± 3 % (89 ± 0%) remain after the outlier rejection method is applied.

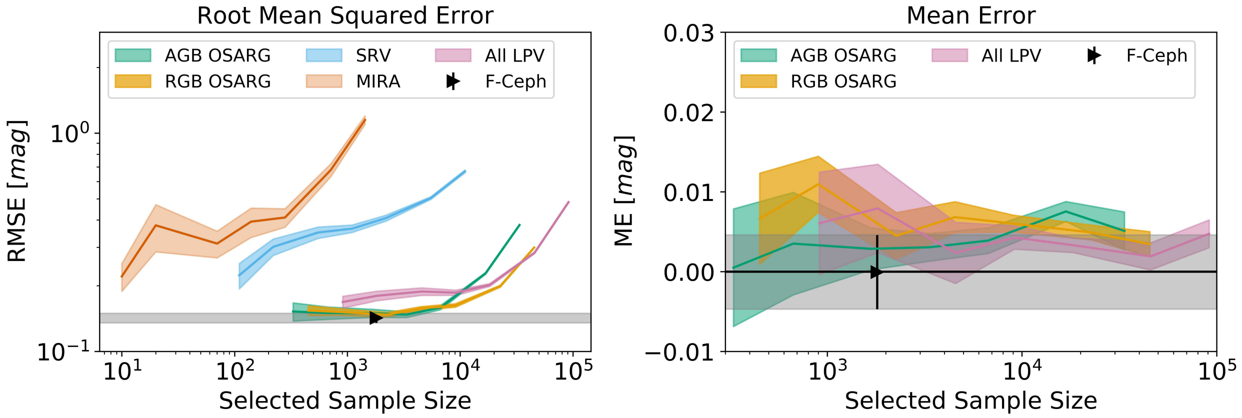

Figure 4. Root mean squared error and mean error of the Machine Learning predictions evaluated on different samples of long period variable stars (LPV) in the LMC as a function of the selected sample size. The horizontal axis shows how many stars are kept in the sample after applying our sample selection methodology (§3). In the left panel, the y-axis shows the root mean squared error (Eq. 8) for asymptotic giant branch (AGB) OSARG, red giant branch (RGB) OSARG, SRV and MIRA stars as well as for all LPV. The right panel shows the mean error (Eq. 9) for asymptotic giant branch (AGB) OSARG, red giant branch (RGB) OSARG and for all LPV. The horizontal grey contour shows the respective results obtained using fundamental mode Cepheids (F-Ceph), where the black right triangle indicates the Cepheid sample size. The mean and errorbars are obtained by 10-fold cross validation as explained in the text.

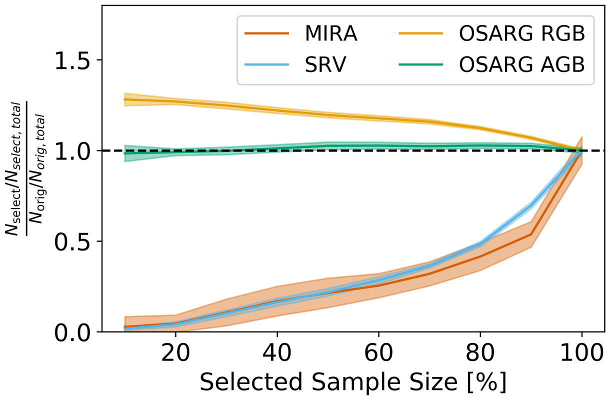

Figure 5. Excess probability over a random selection for the different types of long period variables as a function of the selected sample size in the full LMC LPV sample. The black horizontal dashed line shows the selection probability that would be expected in a completely random selection. The brown, blue, yellow and green lines show the population fractions for the different types of LPV, if the selection is done using our method.

Orig	LMC AGB	SMC AGB	LMC RGB	SMC RGB
10 %	9.3 ± 0.1 %	**8.5** ± 0.2 %	9.0 ± 0.2 %	**9.0** ± 0.1 %
20 %	18.6 ± 0.3 %	**16.8** ± 0.4 %	18.1 ± 0.3 %	**17.7** ± 0.1 %
30 %	27.4 ± 0.3 %	**25.0** ± 0.3 %	27.0 ± 0.3 %	**26.3** ± 0.2 %
40 %	35.6 ± 0.6 %	**33.1** ± 0.6 %	35.8 ± 0.4 %	**35.0** ± 0.2 %
50 %	43.5 ± 0.6 %	**41.0** ± 0.5 %	44.4 ± 0.5 %	**43.5** ± 0.2 %
60 %	51.0 ± 0.7 %	**49.4** ± 0.7 %	52.8 ± 0.7 %	**52.1** ± 0.2 %
70 %	58.8 ± 0.9 %	**58.3** ± 0.9 %	60.9 ± 0.9 %	**60.3** ± 0.2 %
80 %	66.4 ± 1.5 %	**68.0** ± 1.1 %	68.4 ± 0.9 %	**68.8** ± 0.3 %
90 %	73.1 ± 1.4 %	**77.0** ± 1.1 %	75.7 ± 1.1 %	**77.5** ± 0.3 %
100 %	78.9 ± 1.6 %	**84.9** ± 1.0 %	82.6 ± 1.3 %	**86.0** ± 0.3 %

Table 1. Reduction in sample size due to outlier removal in percent. The first column shows the fractional sample size after imposing the cut on the conditional standard deviation, 'original selection' in the following. The other columns report the corresponding fractional sample sizes after the outlier removal for the different subsamples. The SMC results are shown in boldface and will be used as reference in this analysis. The 1σ errors quantify the variation across the 10 cross validation folds.

size that includes this cut but not the outlier rejection step as the 'original selection' in the following. The exact number of stars being culled depends on the subsample and, to a much lesser degree, on the cross validation fold. We report these numbers in Tab. 1. Since we are mostly interested in the properties of the SMC sample in this section, we will use the corresponding SMC fractional samples size (marked in boldface) as a reference in Fig. 7. The corresponding LMC sample sizes can be read-off from Tab. 1.

We quantify the performance of our methodology relative to the Cepheid P-W fit in Fig. 7. In the left panel we show the RMSE (Eq. 8) as a function of the culled fraction of SMC variable stars for the RGB and the AGB population and compare these values with the corresponding result for the P-W fit of fundamental mode Cepheids (F-Ceph) in the SMC. We note that both the RGB OSARG as well as the AGB OSARG population show a rapid decline in RMSE when more data is being culled. While the RGB OSARG sample has a smaller RMSE than the AGB OSARG sample for an original selection of 100%, both samples show quite similar results to each other and to F-Cepheids for an original selection of 20%. The middle panel of Fig. 7 shows the relative bias in the slope of the linear regression fits (Eq. 11) as a function of the culled fraction of the respective sample. We see that the OSARG AGB population performs much better in terms of this metric as compared with the RGB sample. For an original selection of 20%, the performance of the AGB subsample is comparable with the Cepheid sample reference within the 1σ errorbars.

The right panel concentrates on the AGB OSARG and the Cepheid sample. We show the offsets between the linear regression functions estimated on the SMC and the LMC samples $W_\delta = W_{\rm pred,SMC} - W_{\rm pred,LMC}$, evaluated at the per-

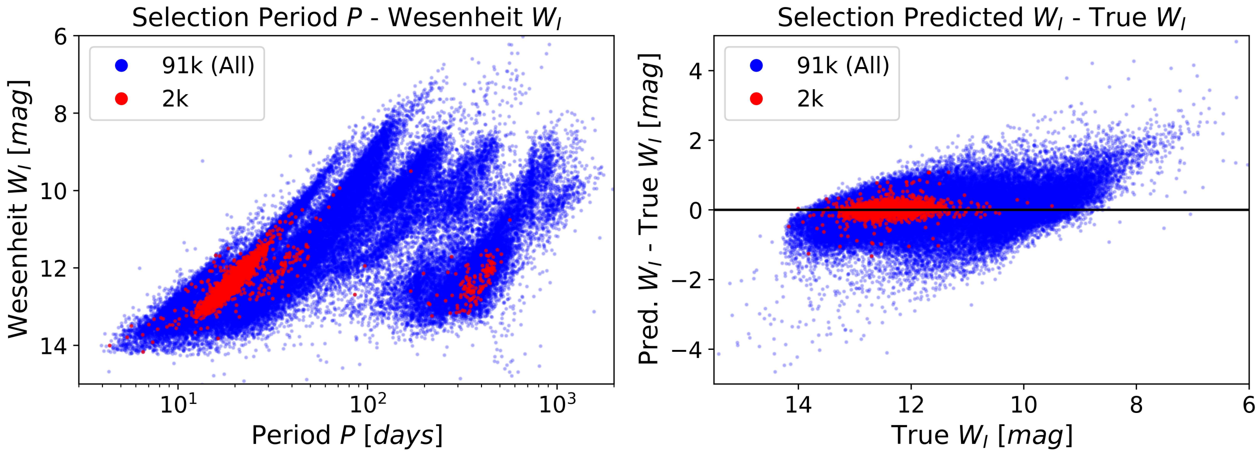

Figure 6. Period-Wesenheit relations and performance of our selected sample of long period variables in the LMC. *Left*: Period P against Wesenheit W_I for long period variables (LPV) in the LMC. Blue points show the full sample, red points show the 2k stars that have the smallest conditional standard deviation. *Right*: Corresponding plot of the true Wesenheit W_I against the difference between the predicted and true Wesenheit W_I. The red points again show the 2k predictions for which our model infers the smallest conditional standard deviation.

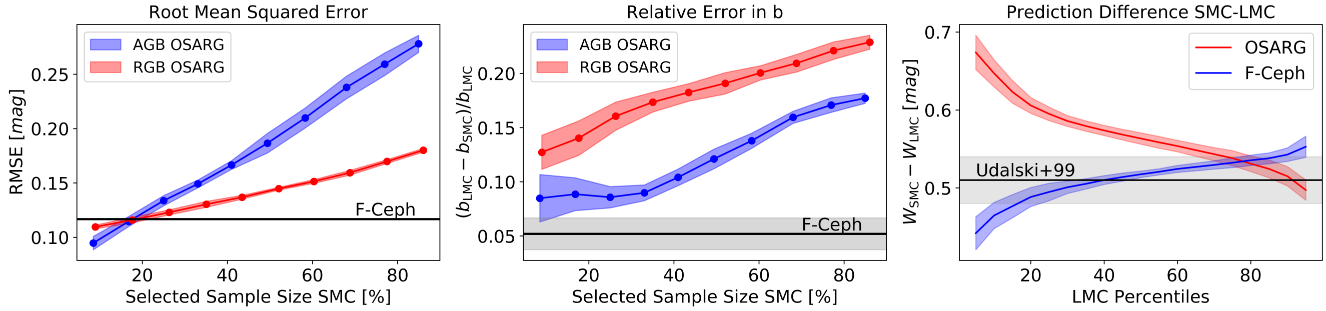

Figure 7. Metrics to quantify the generalization performance of the Random Forest predictions. The model is trained on LMC data and applied to the respective SMC data. Then a linear regression is fitted between the predicted and true Wesenheit magnitudes. The errorbars shown in these plots are the $\pm 1\sigma$ standard deviations from the 10 cross validation folds as described in §4.2. The quoted fractional sample sizes correspond to the 'SMC AGB' and 'SMC RGB' columns of Tab. 1. *Left*: Root mean squared error for the linear regression obtained on the SMC predictions as a function of the remaining fraction of SMC data (§4.2). We show the result for the AGB OSARG (blue), the RGB OSARG (red) and the corresponding result from the fundamental mode Cepheids (F-Ceph). *Middle:* Relative difference in the slope parameter of the linear regression functions fitted on the LMC and SMC samples, as a function of the remaining fraction of LMC and SMC data respectively. For the F-Ceph result, we plot the absolute value of this metric for easier comparison. *Right*: Offset between the linear regression functions fitted on the LMC and SMC samples. The result is shown for the OSARG AGB stars that correspond to an original selection of 20% (see Tab. 1) and the sample of fundamental mode Cepheids. The horizontal axis grid is given by the LMC percentiles of predicted Wesenheit magnitude for the OSARG AGB and fundamental mode Cepheid sample respectively. The horizontal grey area corresponds to the $\pm 1\sigma$ error contour of the measured difference of distance moduli between LMC and SMC using OGLE fundamental mode Cepheids by Udalski et al. (1999).

centiles of the predicted Wesenheit distribution of the respective LMC OSARG AGB and LMC fundamental mode Cepheid sample. This allows us to compare both samples, that cover a different range in Wesenheit index, on the same scale. We note that the slightly larger relative error in the regression slope b_Δ obtained on OSARG AGB stars propagates into a larger variation in W_δ, compared with the Cepheid reference. However both results are comparable and overlap at the faint end, i.e. at the 80th percentile. We compare these results with measurements of the LMC-SMC distance modulus by Udalski et al. (1999) that used a comparable sample of OGLE Cepheids. The corresponding horizontal grey 1σ error contours are consistent with our Cepheid results and also with the results obtained using AGB OSARGs at the faint end. Comparing with Fig. 4, we note that the error induced by biases in the regression slope (≈ 0.1 mag) is an order of magnitude larger than the mean error in the LMC distance anchor (≈ 0.01 mag). Controlling the bias in the regression slope is therefore the most important challenge for obtaining more accurate distance measurements.

Fig. 8 shows the resulting $W_{\rm pred}$-$W_{\rm true}$ relations for the Cepheid population (left panel) and the population of OSARG AGB stars (right panel) selected using the best 20% elements from the original selection based on their conditional standard deviation[8]. For simplicity, in this plot, we merge

[8] This corresponds to the maximum sample size in the left and

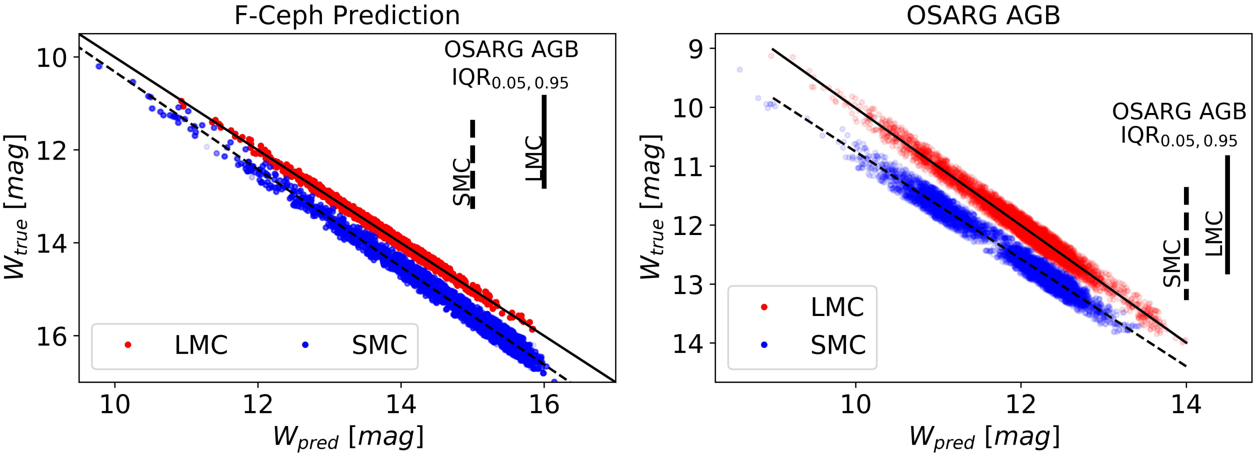

Figure 8. Performance when training on LMC data and applying the model to both LMC and SMC data (§4.2) for an original selection of 20%, i.e. when beginning the analysis with 20% of the sample based on the cut on the conditional standard deviation, before further outlier rejection. *Left*: Wesenheit prediction for fundamental mode Cepheids (F-Ceph) in the LMC (red) and SMC (blue) datasets. The dashed (solid) vertical lines show the inter-quantile range of true Wesenheit values covered by the OSARG AGB sample for the SMC (LMC) datasets. *Right*: Wesenheit predictions for an original selection of 20% OSARG AGB stars for the LMC (red) and SMC (blue).

the datasets in the 10 cross validation folds before applying our methodology. Comparing both panels of Fig. 8, we note that the OSARG AGB sample is at the bright end of the fundamental mode Cepheid (F-Ceph) sample as shown by the vertical lines that indicate the inter-quantile range IQR$_{0.05,0.95}$ between the 0.05 and the 0.95 quantile of the selected true OSARG AGB Wesenheit distribution. We see that the regression lines between the LMC and the SMC are slightly biased for both the Cepheid and the OSARG AGB prediction. We also note that this bias is complementary for both samples, i.e. the offset between the respective SMC and LMC linear regression lines is larger at the bright end for the OSARG AGB subsample and decreases towards the faint end, whereas the contrary is the case for the Cepheid sample. This is also highlighted in the right panel of Fig. 7 that shows the distance between the LMC and SMC linear regression lines as a function of the predicted LMC Wesenheit percentiles. High (low) percentiles indicate the faint (bright) end of the OSARG AGB and F-Ceph predicted LMC Wesenheit distributions. This complementarity of systematic biases suggests that combining distance measurements using both samples might compensate for their respective systematic errors. This will however require that the respective samples are sufficiently complete to avoid introducing additional sample selection biases. We leave a detailed study of this for future work.

4.3 Selected Population

The previous sections demonstrated that our methodology has potential to incorporate LPVs into distance measurements. In §2 we discussed that LPVs, including OSARG variables, follow tight P-W relations and exhibit characteristic period ratios that will help the machine learning algorithm to distinguish between them. The incorporation of oscillation amplitudes will also help to improve the performance of both the selection as well as the final prediction, as they relate to the growth rate of oscillation modes and the luminosity of the star.

In this section we discuss which stars are preferentially selected by our methodology, on the example of the OSARG AGB variables. The left (right) panel of Fig. 9 shows the P-W relation of the 20% best LMC (SMC) OSARG variables selected by our methodology. We note that no outlier rejection is applied here, as we want to show the selection based purely on our methodology. The plot shows that the selected stars cluster around tight 'Wood-sequences' (Wood et al. 1999; Wood 2000; Ita et al. 2004; Wood 2015; Trabucchi et al. 2017), that correspond to different overtones of stellar oscillation. Overplotting the analytical sequences obtained on the LMC OGLE-III dataset by Soszynski et al. (2007), we can attribute our selected sequences, starting from the lower period end, to the A', A, B/C' and D_c/E relations. Trabucchi et al. (2017) analyzed these sequences using a pulsation and stellar population synthesis model tuned to resemble the population of red giants in the LMC. Following their discussion the selected sequences A' and A are attributed to the third (O3) and second (O2) overtone, and the splitted sequences B/C' to the first overtone (O1). The long period population can be roughly associated with the sequences D_c and E in Soszynski et al. (2007), which could potentially be caused by stellar pulsation (Wood et al. 2004; Saio et al. 2015; Trabucchi et al. 2017) or binarity[9] (Wood et al. 2004; Soszyński 2007). The right panel of Fig. 9 shows the corresponding selection for the SMC OSARG AGB sample. We note that the selections between both Magellanic clouds populate the same Wood sequences. Accordingly we can assume that our selection generalizes well to the SMC, as stars with very similar oscillation patterns are selected. This explains

middle panel of Fig. 7, where the RMSE and bias in the slope b_Δ for the OSARG AGB sample is comparable with the Cepheid result.

[9] A low mass companion 'drags' a dust cloud ejected from the central red giant star, which disturbs its lightcurve.

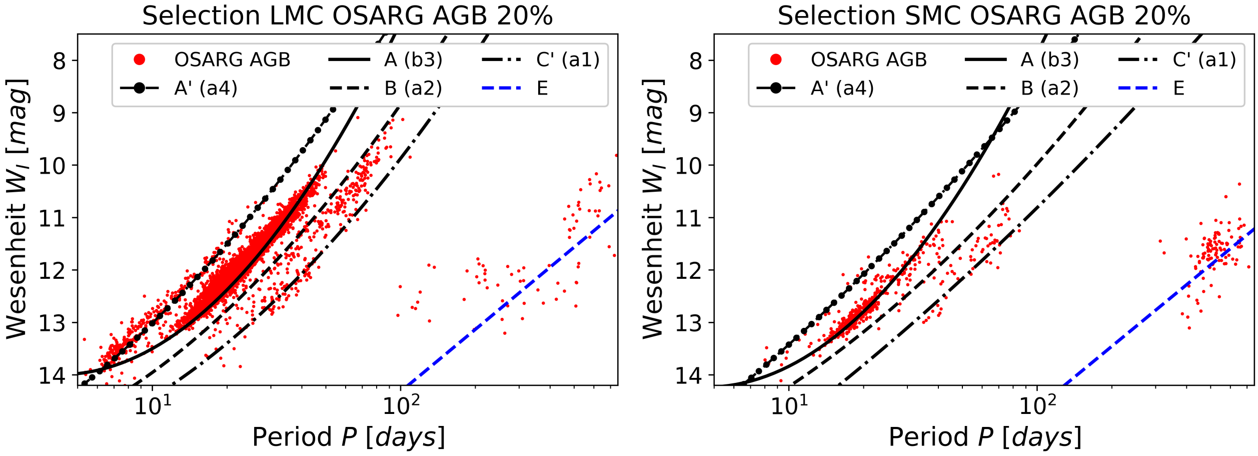

Figure 9. P-W relation of the selected AGB OSARGs in the LMC and SMC. *Left*: P-W relation of the 20% best OSARG variables selected in the LMC (red points). We overplot the Wood sequences obtained from Soszynski et al. (2007, table 1) that best fit our selection. The legend quotes the sequences according to the two conventions in Trabucchi et al. (2017) and, in parenthesis, Soszynski et al. (2007). *Right*: Corresponding plot for the SMC selection and Wood sequences extracted from Soszynski et al. (2007, Tab. 2). We note that no outlier rejection has been applied.

the good performance of our model on the SMC data as demonstrated in the previous section.

5 SUMMARY AND CONCLUSIONS

This paper introduced a novel methodology to select variable stars based on the width of the posterior distribution of their Wesenheit magnitude given their lightcurve parameters. Our selection procedure uses the Random Forest algorithm to estimate this conditional predictive distribution $p(W_I|\mathbf{f})$ of the Wesenheit W_I given a set of features $\mathbf{f}$ extracted from the lightcurve of the variable star. Our feature set $\mathbf{f}$ consists of the first three periods and oscillation amplitudes from the fourier lightcurve fit. We then select variable stars that have a small standard deviation in the conditional predictive distribution.

This selection procedure constructs a sample of variable stars, that show a very tight correlation between the extracted lightcurve features and their Wesenheit magnitude, which is an important requirement to use them as cosmological standard candles. We demonstrate the effectiveness of this methodology using samples of variable stars in the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC), observed in the photometric bands I and V by the OGLE collaboration. We show that our method is able to select a subsample of variable stars within the LMC for which highly accurate predictions of the Wesenheit magnitude, as quantified by the root mean squared error (RMSE), can be derived. For the sample of OGLE Small Amplitude Red Giant (OSARG) stars, we show that the RMSE of these predictions are comparable to the results obtained on fundamental mode Cepheids in the LMC. However the sample of OSARGs with comparable RMS error is larger by a factor of 3-4 and brighter by ≈ 2 Wesenheit magnitudes on average. Accordingly, we can select a sample of variable stars with comparable systematics to Cepheids, that is both more numerous and brighter. This provides exciting prospects to extend the distance ladder to extragalactic galaxies by utilizing these variable stars to improve the calibration of local supernovae samples in distance anchors such as the LMC or M31.

To demonstrate the generalization performance on unseen data from a different galaxy, we train our model on the OSARG sample in the LMC and apply the trained model to the corresponding OSARG sample in the SMC. Using our selection methodology, we obtain RMSE values consistent with the corresponding results obtained by fitting the Period-Wesenheit relation (P-W) on fundamental mode Cepheids. Furthermore we investigate how the bias in the slope of the resulting linear regression between the predicted and the true Wesenheit differs between the samples. This quantity is of specific interest as it contributes to the systematic error budget in the distance calibration of local supernovae. We find that for the AGB subsample we can obtain biases that are comparable with the results obtained using fundamental mode Cepheids. The distance modulus between the LMC and the SMC regressions obtained using fundamental mode Cepheids and OSARG AGB variables is consistent at the faint end of the covered range of Wesenheit magnitudes, despite the slightly larger bias in the regression slope as measured on the selected OSARG AGB stars. Notably, the sign of this systematic is complementary between the OSARG AGB and fundamental mode Cepheid sample. While at the bright end the regression lines are farther apart for the selected OSARG AGB sample, the contrary is true for the fundamental mode Cepheids.

This result indicates that there is potential to combine distance measurements obtained using multiple types of variable stars, like Cepheids and OSARG variables, to better control systematic errors. The inclusion of additional photometric bands and lightcurve features will likely improve these results but also allow the incorporation of other types of variable stars like Mira variables that show tight PL relations when observed with deep near infrared photometry (Huang et al. 2018). The inclusion of Gaia parallaxes can naturally substitute the apparent Wesenheit magnitudes as regression targets, which will improve the calibration of distances to the LMC and SMC. While the presented methodology does not require prior variable star classification, it will still be useful to generate additional features that help the methodology to better separate the different subclasses. For example we found that AGB OSARGs in general yielded better results in the considered performance metrics, as compared with the other types of LPV.

It has to be noted however, that using OSARG variables in different anchor samples like M31 will require better photometry than Cepheid observations. The small amplitude of OSARG variability sets stringent requirements on the quality of the photometry. The median amplitudes for the selected LMC population in Fig. 9 range from 0.006−0.01 [mag]. Thus, the photometric error of the observations has to be of that order to avoid sample selection biases. The observation of long period variables will also require long time series to obtain accurate period estimates. While OSARG variables can have very long oscillation periods of 10^3 days, the bulk of the stars selected by our methodology have periods < 100 days. As a result, the majority of variable stars that would be interesting for distance measurements have period lengths that are comparable to the long period tail of Cepheid samples observed in possible anchor galaxies like M31 (e.g. Kodric et al. 2018). While OSARGs have a more complicated oscillation pattern than these Cepheids, we still expect that the main observational challenge will be their detection at larger distances. Assuming that these observational requirements can be met, we found that the prediction accuracy is mostly sensitive to accurate measurements of the periods and relatively robust against errors in the amplitudes. This can be explained by the tightly spaced period ratio structure of OSARG variables shown in Fig. 2 and the comparatively broad correlation between amplitude and Wesenheit (see Fig. 3). We refer for a more discussion to the appendix. In future work we will further investigate which type of LPV stars can be used for distance calibration in fainter samples. For the calibration of local Cepheid distances, OSARG variables already appear as a viable option.

In a practical application it will also be important to optimize the sample not only with respect to the width of the conditional predictive distribution, but also such that differences in the linear regression slope between the different anchor samples are reduced. This procedure will naturally include a more advanced outlier rejection methodology, that was not optimized in this work. In this context we want to reiterate that the methodology presented in this paper does not require a prior classification of the sample into different types of variable stars. Samples of variable stars that are ‘good standard candles’ will be automatically selected in high-dimensional feature space based on their small RMS error and their robust regression functions across different anchor samples. This not only reduces the need to obtain large, accurately labeled training sets to use in variable star classification pipelines, but also avoids biases due to misclassification errors. In future work we will apply our methodology to additional anchor samples and optimize our selection criteria towards the specific science goal of calibrating local supernovae samples.

We conclude that the usage of other types of variable stars besides Cepheids as ‘standard candles’ has great potential to improve distance measurements and the calibration of local supernovae samples, which will ultimately lead to a better understanding of sources of systematic error in H_0 measurements.

ACKNOWLEDGEMENTS

MMR thanks Ben Hoyle and Kerstin Paech for useful discussions and comments on the draft.
MMR is supported by DOE grant DESC0011114. HT and RM are partially supported by NSF grant IIS-1563887. SK is partially supported by NSF grant AST-1813881.

REFERENCES

Armstrong D. J., et al., 2016, MNRAS, 456, 2260
Bernal J. L., Verde L., Riess A. G., 2016, J. Cosmology Astropart. Phys., 10, 019
Bishop C. M., 2006, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA
Breiman L., 2001, Mach. Learn., 45, 5
Breiman L., Friedman J., Stone C., Olshen R., 1984, Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis, https://books.google.com/books?id=JwQx-WOmSyQC
Conley A., et al., 2011, ApJS, 192, 1
Di Valentino E., Melchiorri A., Silk J., 2016, Physics Letters B, 761, 242
Dubath P., et al., 2011, MNRAS, 414, 2602
Dvorkin C., Wyman M., Rudd D. H., Hu W., 2014, Phys. Rev. D, 90, 083503
Efstathiou G., 2014, MNRAS, 440, 1138
Feeney S. M., Mortlock D. J., Dalmasso N., 2018, MNRAS, 476, 3861
Frank E., Bouckaert R. R., 2009, in Advances in Machine Learning, First Asian Conference on Machine Learning, ACML 2009, Nanjing, China, November 2-4, 2009. Proceedings. pp 65–81, doi:10.1007/978-3-642-05224-8_7, https://doi.org/10.1007/978-3-642-05224-8_7
Freedman W. L., et al., 2001, ApJ, 553, 47
Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA
Huang C. D., et al., 2018, ApJ, 857, 67
Ita Y., et al., 2004, MNRAS, 347, 720
Kiss L. L., Bedding T. R., 2003, MNRAS, 343, L79
Kodric M., et al., 2015, ApJ, 799, 144
Kodric M., et al., 2018, AJ, 156, 130
Kügler S. D., Gianniotis N., Polsterer K. L., 2015, MNRAS, 451, 3385
Leistedt B., Peiris H. V., Verde L., 2014, Physical Review Letters, 113, 041301
Madore B. F., 1982, ApJ, 253, 575
Naul B., Bloom J. S., Pérez F., van der Walt S., 2018, Nature Astronomy, 2, 151
OGLE 2018, OGLE-III On-line Catalog of Variable Stars, ogledb.astrouw.edu.pl/~ogle/CVS/
Palaversa L., et al., 2013, AJ, 146, 101
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Planck Collaboration et al., 2016a, A&A, 594, A13
Planck Collaboration et al., 2016b, A&A, 594, A14
Rau M. M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, MNRAS, 452, 3710

Richards J. W., et al., 2011, ApJ, 733, 10

Riess A. G., et al., 2016, ApJ, 826, 56

Saio H., Wood P. R., Takayama M., Ita Y., 2015, MNRAS, 452, 3863

Schapire R. E., Stone P., McAllester D., Littman M. L., Csirik J. A., 2002, in Proceedings of the Nineteenth International Conference on Machine Learning. `http://www.cs.utexas.edu/users/ai-lab/?ICML02-tac`

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Sesar B., et al., 2017, AJ, 153, 204

Smolec R., Dziembowski W., Moskalik P., Netzel H., Prudil Z., Skarka M., Soszynski I., 2017, in European Physical Journal Web of Conferences. p. 06003 (`arXiv:1703.03029`), doi:10.1051/epjconf/201715206003

Solà J., Gómez-Valent A., de Cruz Pérez J., 2017, Physics Letters B, 774, 317

Soszyński I., 2007, ApJ, 660, 1486

Soszynski I., Udalski A., Kubiak M., Szymanski M., Pietrzynski G., Zebrun K., Szewczyk O., Wyrzykowski L., 2004, Acta Astron., 54, 129

Soszynski I., et al., 2007, Acta Astron., 57, 201

Soszyński I., et al., 2008, Acta Astron., 58, 163

Soszyński I., et al., 2009, Acta Astron., 59, 239

Soszyński I., et al., 2010, Acta Astron., 60, 17

Soszyński I., et al., 2011, Acta Astron., 61, 217

Stello D., et al., 2007, MNRAS, 377, 584

Tabur V., Bedding T. R., Kiss L. L., Giles T., Derekas A., Moon T. T., 2010, MNRAS, 409, 777

Trabucchi M., Wood P. R., Montalbán J., Marigo P., Pastorelli G., Girardi L., 2017, ApJ, 847, 139

Udalski A., Szymanski M., Kubiak M., Pietrzynski G., Soszynski I., Wozniak P., Zebrun K., 1999, Acta Astron., 49, 201

Whitelock P. A., Feast M. W., Van Leeuwen F., 2008, MNRAS, 386, 313

Wood P. R., 2000, Publ. Astron. Soc. Australia, 17, 18

Wood P. R., 2015, MNRAS, 448, 3829

Wood P. R., et al., 1999, in Le Bertre T., Lebre A., Waelkens C., eds, IAU Symposium Vol. 191, Asymptotic Giant Branch Stars. p. 151

Wood P. R., Olivier E. A., Kawaler S. D., 2004, ApJ, 604, 800

Wyman M., Rudd D. H., Vanderveld R. A., Hu W., 2014, Phys. Rev. Lett., 112, 051302

Yuan W., He S., Macri L. M., Long J., Huang J. Z., 2017, AJ, 153, 170

Zhang B. R., Childress M. J., Davis T. M., Karpenka N. V., Lidman C., Schmidt B. P., Smith M., 2017, MNRAS, 471, 2254

Zhao G.-B., et al., 2017, Nature Astronomy, 1, 627

APPENDIX A: FEATURE IMPORTANCE AND ROBUSTNESS

The selection methodology presented in this paper is quite general and able to identify samples of variable stars with better prediction accuracy across a wide range of variable star types as shown in Fig. 4. However we found that Ogle Small Amplitude Red Giant variables show especially tight Period-Wesenheit relations. To understand the sensitivity of these predictions on the input feature set, this appendix performs a feature importance study and tests the robustness of our methodology against feature noise.

A1 Feature Importance

We use the Gini criterion (Breiman et al. 1984) (Eq. 7) as implemented in the scikit-learn package (Pedregosa et al. 2011), as a measure of feature importance. Considering a single tree, we associate a node with a split on a certain variable. The split is selected such that the Gini criterion is decreased. Furthermore, we can attribute a weight to each node defined by the fraction of samples that reach this node. Using these weights, we can average the decrease in the Gini criterion across all trees in the Random Forest to obtain a measure of how important a certain feature is.

Focusing on the LMC OSARG AGB dataset for simplicity, we run the 10 fold cross validation procedure described in §4.1 to obtain 10 estimates of feature importance as shown in the left panel of Fig. A1. The errobars on the histograms denote the 3σ errors across the 10 folds.

We see that the sets of periods are significantly more important than the sets of amplitudes. The most important features are the first P_1 and third P_3 period and the secondary amplitude A_2. The least important one is the third amplitude A_3. We note however that especially the amplitudes are correlated and we therefore expect a level of redundancy in the information contained in the amplitude features. Nonetheless, this result indicates that amplitude information is important for the prediction accuracy, even for the single population of LMC OSARG AGB stars, for which small oscillation amplitudes is a common feature. The much larger importance of periods for the prediction accuracy is to be expected, based on the fine grained period ratios apparent in the Petersen Diagram (see Fig. 2) and the strong Period-Wesenheit correlation in the Wood sequences (see §4.3).

A2 Feature Noise

We test the robustness of our Machine Learning methodology with respect to feature noise by artificially imposing Gaussian random noise on the features and subsequently testing the prediction accuracy. For this we perform an analysis similar to §4.1. We split the full dataset of LMC OSARG AGB stars randomly into 2 disjoint, similarly sized parts: the training set with 50000 stars and the test set with 41374 stars. We then add Gaussian noise to the set of periods, the set of amplitudes and both sets of features respectively, train our model on the training set and estimate conditional density functions for all stars in the test set. We do this for the noiseless dataset, as well as the three aforementioned combinations of noisy data and perform the subsequent analysis separtely on each of them. We select the 7% (3038) stars[10] from the test set predictions that have the smallest predicted conditional standard deviation. For this selected subset, we estimate the relative increase in the root mean squared error $\Delta_{\rm rel}$ over the noiseless case defined as

$$\Delta_{\rm rel} = \frac{\mathrm{RMSE}_{\rm noisy} - \mathrm{RMSE}_{\rm noiseless}}{\mathrm{RMSE}_{\rm noiseless}}, \tag{A1}$$

where $\mathrm{RMSE}_{\rm noisy}$ and $\mathrm{RMSE}_{\rm noiseless}$ denote the root mean squared error (RMSE) defined in Eq. 8 for the noisy and noise free datasets. The standard deviation of the Gaussian noise that is imposed on the features is given as $\sigma = \frac{f}{\mathrm{S/N}}$, where S/N denotes the signal to noise ratio. We note that

[10] This number amounts approximately to the sample size, where the LMC AGB OSARG prediction accuracy begins to significantly deviate from the Cepheid baseline as shown in Fig. 4.

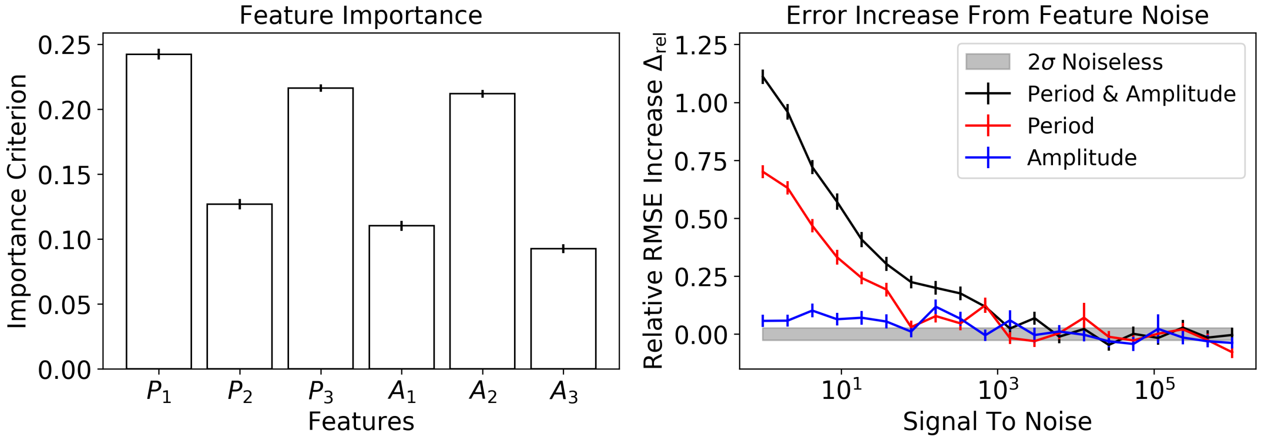

Figure A1. Feature importance and robustness to input noise. *Left:* Feature importance for the full LMC OSARG AGB sample for the period P_{1-3} and amplitude A_{1-3} features used in the analysis. We show the 3σ errors on the histograms to quantify the statistical noise. *Right:* Signal to noise imposed on the respective feature combination against the relative RMSE increase over the noiseless performance for a selected sample of 3038 LMC OSARG AGB stars. The grey horizontal contour and the error bars show the 2σ error to be expected from the statistical variance. We show the case of imposing noise on the first three periods (red), Amplitudes (blue) and both features (black).

this validation procedure is a slightly simplified version of the 10 fold cross validation approach, as it considers only a single train/test set split. This simplification is justified, as the LMC OSARG AGB sample is quite large even after selecting the best variable stars. For each prediction we obtain errors by propagating the accuracy in the mean squared error MSE through Eq. 8 and Eq. A1.

The right panel of Fig. A1 plots the signal to noise ratio against the relative increase in the RMSE for the three scenarios, i.e. degrading the three amplitudes (blue), the three periods (red) and both the periods and amplitudes (black). The errorbars show the 2σ errors. The grey horizontal band shows the 2σ error of the baseline noiseless case.

We see that degrading the periods has the largest effect on $\Delta_{\rm rel}$, where the model is quite robust against a degradation in the amplitudes. As expected, we obtain the largest performance reduction, if all features are degraded. For S/N $\approx 10^3$, we obtain consistent results with the noiseless case.

This apparent difference in robustness between periods and amplitudes can be explained by noting that the sequences shown in the Petersen diagrams (Fig. 2) can only be resolved, if the periods can be determined accurately. As a result, we can assume that quite precise period determinations are needed to robustly separate the Wood sequences, which will be essential for accurate predictions. The correlation between amplitude and Wesenheit shown in Fig. 3 are much broader and thus less sensitive to inaccurate measurements of the amplitudes. Greater sensitivity to period measurements compared with measurements of amplitudes, can also be explained by the high feature importance of periods as discussed in §A1.

This analysis considered the case of the same input feature noise across the training and test sets. We leave the case of different noise levels between training and test sets for future work. However we would like to note that the training sample will most likely have better photometry than the test sample, if nearby samples are used as anchors. This suggests that we can then artificially degrade this training sample to match the noise properties of the test sets, even without a more advanced correction.

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