Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **476,** 1120–1139 (2018) doi:10.1093/mnras/sty165

Estimates of the atmospheric parameters of M-type stars: a machine-learning perspective

L. M. Sarro,[1]★ J. Ordieres-Meré,[2] A. Bello-García,[3] A. González-Marcos[4] and E. Solano[5]

[1] *Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia, c/ Juan del Rosal 16, E-28040 Madrid, Spain*
[2] *Universidad Politécnica de Madrid (UPM), ETSII. PMQ Research Group, José Gutiérrez Abascal 2, E-28006 Madrid, Spain*
[3] *Universidad de Oviedo, Construction and Manufacturing Engineering Department, Campus de Viesques s/n, Gijón, E-33203 Asturias, Spain*
[4] *Universidad de la Rioja, P2ML Research Group, San José de Calasanz 31, E-26004 Logroño, La Rioja, Spain*
[5] *Centro de Astrobiología (CSIC-INTA), Ctra. Ajalvir km 4, E-28850 Torrejón de Ardoz, Madrid, Spain*

Accepted 2018 January 16. Received 2018 January 16; in original form 2017 October 5

ABSTRACT

Estimating the atmospheric parameters of M-type stars has been a difficult task due to the lack of simple diagnostics in the stellar spectra. We aim at uncovering good sets of predictive features of stellar atmospheric parameters ($T_{\rm eff}$, $\log(g)$, $[M/H]$) in spectra of M-type stars. We define two types of potential features (equivalent widths and integrated flux ratios) able to explain the atmospheric physical parameters. We search the space of feature sets using a genetic algorithm that evaluates solutions by their prediction performance in the framework of the BT-Settl library of stellar spectra. Thereafter, we construct eight regression models using different machine-learning techniques and compare their performances with those obtained using the classical χ^2 approach and independent component analysis (ICA) coefficients. Finally, we validate the various alternatives using two sets of real spectra from the NASA Infrared Telescope Facility (IRTF) and Dwarf Archives collections. We find that the cross-validation errors are poor measures of the performance of regression models in the context of physical parameter prediction in M-type stars. For $R \sim 2000$ spectra with signal-to-noise ratios typical of the IRTF and Dwarf Archives, feature selection with genetic algorithms or alternative techniques produces only marginal advantages with respect to representation spaces that are unconstrained in wavelength (full spectrum or ICA). We make available the atmospheric parameters for the two collections of observed spectra as online material.

Key words: methods: data analysis – methods: statistical – techniques: spectroscopic – stars: atmospheres – stars: fundamental parameters – stars: late-type – stars: statistics.

1 INTRODUCTION

M-type dwarfs constitute the largest contribution by number to the Galactic population (Bochanski et al. 2010). This Galactic component is very important as its properties convey crucial information about the Galactic structure and evolution (Bonfils et al. 2013). They are known to harbour super-Earth (Bonfils et al. 2013) and Earth-sized exoplanets (Dressing & Charbonneau 2015), and have recently become a major target in large-scale searches for habitable ones due, amongst other reasons, to the reduced star-to-planet mass and light ratios (Alonso-Floriano et al. 2015). In addition, the geometrical probability to observe a transit is significantly higher because the habitable zone is closer to the host star (Shields, Ballard & Johnson 2016), and the shorter orbital periods result in higher transit frequencies and larger expected numbers of transits for any given observation time span.

These stars span two orders of magnitude in luminosity and almost one order of magnitude in mass, from 0.075 M⊙ to 0.6 M⊙. At 0.35 M⊙, these stars become fully convective and, given their low internal pressures, this results in life spans that greatly exceed the age of the Universe (Adams, Bodenheimer & Laughlin 2005). Although much theoretical work has been invested in understanding this low-mass end of the Main Sequence (Browning 2008), there are still some discrepancies between models and observations (see e.g. Torres 2013, for an account of the observed inflated radii and cooler temperatures with respect to model predictions).

Scattering from the fine dust grains that form in the atmospheres of the coolest M dwarfs results in veiling of the spectra (Allard et al. 2012). Furthermore, the complexity of modelling the molecular bands (numerous transitions, frequency-dependent absorption coefficients, collisional transitions) and the difficulty in defining a

★ E-mail: lsb@dia.uned.es

true stellar continuum prevents the application of standard techniques (common for F-, G- and K-type stars) to the determination of physical parameters of M-type stars.

Given the prevalence of these late-type stars, it has become increasingly important to be able to estimate their atmospheric physical parameters with reproducible methods that provide homogeneous values for large samples of spectra. Rojas-Ayala et al. (2012) proposed the H_2O–K2, Na I and Ca I spectral indices to estimate spectral type, effective temperature ($T_{\rm eff}$) and metallicity from K-band spectra between 1.0 and 2.4 μm at a resolution $R \approx 2700$. They measured spectral types with a quoted accuracy of 0.6 subtypes, and metallicities with a root mean square error (RMSE) of 0.1 dex (0.14 dex for the iron abundance). The definition of the line/band spectral regions and pseudo-continuum is justified in terms of contiguity and the avoidance of other atomic features.

Neves et al. (2014), in contrast, concentrated on high-resolution ($R \approx 115\,000$) spectra in the optical range. They proposed a method based on a linear least-squares fit of the equivalent widths (EWs) of 4104 lines in the 530–690 nm spectral range to effective temperatures and metallicities derived from the scales by Neves et al. (2012) and Casagrande, Flynn & Bessell (2008). They show RMSE of 0.12 dex for the metallicity and 293 K for the effective temperature. Newton et al. (2015) developed a calibration of effective temperatures, radii and bolometric luminosities with Mg and Al spectral features measured in low-resolution near-infrared spectra (from the SpeX instrument at the NASA Infrared Telescope Facility (IRTF, Cushing, Rayner & Vacca 2005; Rayner, Cushing & Vacca 2009), the same instrument used and described in Section 3 of this work). They quoted residual standard deviations of the $T_{\rm eff}$ fit of 73 K.

Mann et al. (2015) calculated effective temperatures using the classical χ^2 minimization of the observed optical spectra with respect to the CIFIST2011 library of BT-Settl synthetic spectra. Uncertainties of temperatures are estimated around 60 K. Metallicities are estimated from the EWs of atomic spectral lines in near-infrared spectra (again SpeX) using calibrations obtained from wide binaries with FGK primaries. The estimated errors in metallicity are 0.08 dex.

The previous summary shows a lack of estimates for the surface gravity and a variety of methodologies for the estimation of temperatures and metallicities. In this work we are mainly concerned with estimating atmospheric physical parameters using spectral features and libraries of synthetic spectra. Our aim is to identify the best spectral features for estimating $T_{\rm eff}$, $\log(g)$ and $[M/H]$, not only for M dwarfs but also for other luminosity classes.

Cesetti et al. (2013) proposed sensitivity maps (the derivative of the monochromatic fluxes with respect to the atmospheric physical parameters) to rank spectral features. They searched the space of spectral features for predicting $T_{\rm eff}$, $\log(g)$ and $[M/H]$. In contrast, Mann et al. (2013) concentrated their efforts on a systematic search for spectral diagnostics of the metallicity in the domain of moderate resolution ($1300 < R < 2000$) visible and infrared spectra of late-K and M dwarfs. In this sense, our work represents an extension of the work by Mann et al. (2013) in that we (i) include other luminosity types other than dwarf stars in the analysis; (ii) search for spectral diagnostics of the effective temperature and surface gravity as well; (iii) include the definition of the continuum bands in the search for the optimal features, as opposed to using a fixed list of continuum definitions; (iv) train a regression model to predict the physical parameters from the set of optimal spectral features rather than a collection of linear models, one for each spectral feature; and (v) introduce the technique of cross validation in order to minimize the problem of overfitting (see e.g. Gelman et al. 2013) and the associated underestimation of the prediction errors. Our (comparatively more complex nonlinear) regression models integrate all the information from the selected features to produce a single estimate of the physical parameters. These more complex models come at the expense of abandoning the empirical approach of Mann et al. (2013) that used observed spectra of late-K and M dwarfs in binaries with solar-type primaries. Our more comprehensive approach requires a larger set of spectra to select the features and train the regression models and thus we must turn to libraries of synthetic spectra that cover more densely and homogeneously the space of parameters.

In this work we explore the validity of the features proposed by Cesetti et al. (2013) and propose and evaluate new features using standard machine-learning (ML) techniques. In Section 2 we describe the methodology used to define and evaluate the spectral features; in Section 3 we apply the methodology described in Section 2 in the context of the wavelength coverage and resolution of the IRTF collection of spectra; we describe the feature definition results and evaluate them for the task of predicting physical parameters on the actual observed spectra that make up the collection. Section 4 describes the same steps in the context of the Dwarf Archives collection of spectra. Finally, Section 5 summarizes the main results and conclusions of the paper.

2 METHODOLOGY

The objective addressed in this section is to develop an automated procedure to identify spectral bands that yield good atmospheric temperature, gravity and metallicity (hereafter physical parameters) diagnostics for M-type stars. Given the lack of a calibration set of benchmark stars with observed spectra and homogeneous coverage of the space of physical parameters, we must turn to synthetic libraries of spectra. Furthermore, only temperatures and gravities can be calibrated independently of the spectra (for example as in Ségransan et al. 2003, using interferometry): all metallicity estimates in the literature are based on collections of synthetic spectra, and therefore spectral synthesis codes are the only resource to construct regression models. Even in the case of interferometry, the estimates of radii (and therefore gravity) depend on stellar models (although less strongly) via limb-darkening corrections.

As an alternative to the methods based on genetic algorithms used in this work, the atomic or molecular line/band parameters can be used in principle to select the spectral features that are more sensitive to changes in the physical parameters, as in Passegger, Wende-von Berg & Reiners (2016). However, the suitability of spectral features as diagnostics of the stellar atmospheric properties depends not only on the individual behaviour of each line/band, but also on the relative properties of neighbouring features in the same spectral region, which may overlap depending on the spectral resolution. Furthermore, good spectral diagnostics at a given signal-to-noise ratio (SNR) may show a severely degraded predictive power in the low-SNR regime. Therefore, we propose an alternative selection approach that considers the resolution and SNR ratio, to assess the utility of spectral features for the task of inferring physical atmospheric parameters.

In the following, we adopt the BT-Settl CIFIST2011 library of synthetic spectra (Allard et al. 2013) as the framework where spectral diagnostics will be searched. This library does not include variations due to line-broadening mechanisms such as micro- or macro-turbulence or rotation and hence our results can be biased if their effects in the spectra are strong. The predictions from the models presented here do not take these broadening mechanisms

into account. These synthetic spectra were preprocessed in several steps, as described below.

2.1 Spectral preprocessing

First, and in order to define good temperature diagnostics, spectra between 2000 and 4200 K in steps of 100 K were selected, with $\log(g)$ in the range between 4 and 6 dex (when g is expressed in cm s^{-2}), in steps of 0.5 dex. The metallicity of the representative spectra was restricted to the set 0, 0.5 and −1 dex. This yields a total set size of 535 available spectra.

A series of preprocessing steps was then carried out in order to match the spectral resolution and wavelength coverage and sampling of the synthetic library to that of the collection of observed spectra collected from the Dwarf Archives or IRTF (see below). This required the definition of a common wavelength range present in all available observed spectra, and subsequent trimming to match that range. A unique wavelength sampling was also defined and all spectra (synthetic and observed) interpolated to match the sampling. Finally, all spectra, both synthetic and observed, were divided by the integrated flux in order to factor out the stellar distance. This is necessary in order to compare our feature selection proposal with the techniques that make use of the full spectrum (minimum χ^2 and independent component analysis (ICA) compression; see below).

In order to avoid selecting spectral features that work well only in the unrealistic SNR $= \infty$ regime, the search for optimal diagnostics of the atmospheric parameters of M stars was carried out for three SNR values (10, 50 and ∞) by degrading the synthetic spectra with Gaussian noise of zero mean. These values were found to be sufficient in a wide range of experiments carried out in parallel and described in González-Marcos et al. (2017). The special SNR $= \infty$ case has been retained for the sake of completeness although González-Marcos et al. (2017) show that training sets derived from noiseless spectra are at best unnecessary, and at worst damage performance severely.

2.2 Feature definition and selection

As mentioned in Section 1, defining good spectral diagnostics for the prediction of atmospheric physical parameters of M stars is a difficult task. The work in Cesetti et al. (2013) defined wavelength regions in the I and K bands optimal for determining physical parameters based on the sensitivity exhibited by the flux emitted in these segments to changes of the physical parameters. The sensitivity was measured in terms of the derivative of the flux with respect to the physical parameter. The approach adopted here is to select spectral features that yield the best accuracy when used as predictive variables in a regression model that estimates the stellar atmospheric physical parameters ($T_{\rm eff}$, $\log(g)$ and metallicity). The evaluation of the accuracy of the estimates produced from a subset of features is described further below. We consider the effective temperature as the dominant parameter influencing changes in the stellar spectra (a strong feature). Therefore, it was estimated first, and then used as input in the regression models for the gravity and metallicity.

Here, a feature F is defined as

$$F = \int_{\lambda_1}^{\lambda_2} \left(1 - \frac{f(\lambda)}{F_{\rm cont}}\right) \cdot {\rm d}\lambda \qquad (1)$$

where $f(\lambda)$ denotes the normalized flux from the star at wavelength λ, and where $F_{\rm cont}$ is the average flux in a spectral band between $\lambda_{\rm cont;\,1}$ and $\lambda_{\rm cont;\,2}$. We explain below how we search for the band definitions that produce physical parameter predictions with the smallest errors.

We also studied features defined as

$$F' = \frac{\int_{\lambda_1}^{\lambda_2} f(\lambda) \cdot {\rm d}\lambda}{\int_{\lambda_3}^{\lambda_4} f(\lambda) \cdot {\rm d}\lambda} \qquad (2)$$

where λ_1, λ_2, λ_3 and λ_4 delimit two spectral bands such that the ratio of the integrated fluxes in the two bands is assumed to be a good feature for predicting the star atmospheric physical parameters, alone or in combination with other features. The results obtained with this alternative feature definition did not differ significantly on average from those observed with that adopted in equation (1), and including them here would result in an excessively lengthy article. In view of the equivalent global performances, we preferred the former because it allows direct comparison with the features proposed by Cesetti et al. (2013).

We used genetic algorithms (hereafter GAs) to solve the optimization problem described above, that is, the problem of finding the features (band boundaries) that minimize the prediction error of a regression estimate of the physical parameters. We used the implementation of genetic algorithms publicly available as the R GA package (Scrucca 2013). The concept of using *in silico* (that is, an algorithmic analogue of) evolution for the solution of optimization problems was introduced by Holland (1975). Although its application is now reasonably widespread (see e.g. Goldberg 1989), it became popular only when sufficiently powerful computers became available. GAs were presented to the astronomical community by Charbonneau (1995), and have been used extensively in the past (see De Geyter et al. 2013, for a significant application of GAs in astronomy).

For the sake of simplicity, let us define GAs as search algorithms that are based on the principle of evolution by natural selection. The procedure works by evolving (in the sense explained below) an initial random population of chromosomes, in our case defined as sets of spectral features defined by equation (1). Evolution proceeds via cycles of differential replication, recombination and mutation of the fittest chromosomes. The concept of fittest is context dependent, but in our case fitness is defined in relation to the accuracy with which a simple multivariate linear model trained on a given chromosome (a set $\{F_i\}$ of spectral features) predicts the physical parameters. This linear model may not capture instrinsic nonlinearities between the input space (the features) and the predicted physical parameters but it is significantly faster and simpler than the actual models (described in Section 2.3) that will eventually be used to predict the physical parameters. It would certainly be preferable to use as fitness criterion the accuracy of the models described in Section 2.3 but this turned out to be computationally prohibitive for the several millions of evaluations involved in the GA. The accuracy of the linear models is measured with the Akaike information criterion (AIC) of the fitted linear model. The AIC value of a given model is given by equation (3),

$${\rm AIC} = 2 \times k - 2 \times \ln(\hat{L}), \qquad (3)$$

where k is the number of model parameters and $\hat{L}$ is the maximum value of the likelihood function (the probability of the data given the linear model). Since k is fixed for all models, the AIC effectively reduces to the maximum likelihood or least-squares criterion.

The data set used to search for the optimal set of spectral features will be, as mentioned before, the BT-Settl collection of

synthetic spectra, where each spectrum is tagged with the effective temperature, gravity and metallicity of the model atmosphere from which the spectrum emerged.

The implementation of the GA comprises the following steps:

(i) **Stage 1**: Definition of the population of potential features as a set of chromosomes that are evolved by the genetic algorithm in order to increase the fitness.
(ii) **Stage 2**: Each chromosome in the population is evaluated by its ability to predict the physical parameters of each star in the data set (fitness function: equation 3).
(iii) **Stage 3**: Chromosome selection. The new generation of individuals is initialized by transferring a number of the fittest chromosomes in the previous generation. The percentage of individuals transferred is known as the degree of elitism.
(iv) **Stage 4**: The population of chromosomes is replicated. Chromosomes with higher fitness scores will generate more numerous offspring.
(v) **Stage 5**: The genetic information contained in the parent chromosomes is combined through genetic crossover (two randomly selected parent chromosomes are used to create two new chromosomes).
(vi) **Stage 6**: Mutations are then introduced in the chromosomes randomly. These mutations produce new genes by randomly redefining the gene components (that is, by changing the band boundaries λ_i that define the gene). Stages 5 and 6 are applied over the chromosomes established at Stage 4.
(vii) **Stage 7**: This process is repeated from Stage 2 until a target accuracy is achieved or the maximum number of iterations attained.

We test features defined by bands (the numerator and denominator of equation 1) that comprise 10 consecutive bins (fluxes) of a spectrum. The bands tested in different features may overlap by as much as five consecutive bins (which in practice implies that we define the first feature as the spectral chunk between wavelength bins $i = 1$ and $i = 10$, the second feature between bins $i = 6$ and $i = 15$, the third feature between bins $i = 11$ and $i = 20$, etc). The spectral bands in the numerator and denominator of a test feature cannot overlap.

It would be possible to evaluate the predictive performance of individual features defined with equation (1). An obvious conceptual limitation of this univariate approach (considering chromosomes that code a single predictive feature) would be the lack of consideration that features work in the context of interconnected pathways and therefore it is their behaviour as a group that has to be evaluated in terms of the predictive accuracy. In other words, a single feature can yield a poor predictive performance alone, but improve very significantly the prediction accuracy when used in combination with other features. Multivariate selection methods thus seem more suitable for the analysis of the regressors since variables are tested in combination to identify interactions between features. In this work we define a chromosome as a set of 10 individual genes, and each gene codes a pair of non-overlapping spectral bands, the ratio of which is the feature defined by equation (1).

The population size was set to 8000 individuals and the maximum number of accepted iterations set to 4000. We produced three randomly started populations so as to provide enough initial variety. The crossover and mutation probabilities were set to 0.85 and 0.35, respectively. Elitism (the fraction of the population copied into the next generation and composed of the fittest individuals) was fixed to 15 per cent. We used a binary codification of the chromosomes and a parallel implementation of the GA in a farm of fifteen computers per physical parameter.[1]

Feature fitness was defined in terms of the RMSE of a linear regression model trained with the chromosome features. It is important to stress that the regression model used to evaluate the fitness of the feature sets (chromosomes) is not the same model that will be used in practice to predict physical parameters for observed spectra, as described in Section 2.3 below. For fitness evaluation in the GA we used a simple multilinear model for the sake of speed, given the extreme size of the search space of all possible combinations of 10 spectral features. In the IRTF context, these 10 features in each chromosome are selected from among the roughly 6000 potential features. This is 600010, which has an order of magnitude of 10^{24}.

The GA procedure provides us with a large collection of chromosomes, each one consisting of 10 spectral features. We choose 10 as a compromise. On the one hand, we have the intuition that the physical parameters that we intend to predict can be formulated as nonlinear combinations of several interacting features (that is, the predictive power of the set of features is higher that the sum of the individual predictive powers of the individual features). On the other hand, we need to limit the complexity of the models in order to attain reasonable computation times. Although these collections of chromosomes resulting from the GA are all potential solutions of the problem, it is not immediately clear which one should be selected for the final regression model. In this work we have selected the most frequent features amongst the fittest chromosomes as predictive variables of the physical parameters in regression models. Features appearing in fewer than five chromosomes were initially discarded as they cannot be relevant by themselves and just arise randomly by combination with other stronger chromosomes.

Once the GA has generated a proposal set of features for predicting each of the physical parameters, the next step consists in training the regression model based on these features. This is described in the next section.

2.3 Regression models

Once a feature set has been selected from the output of the GA, we construct regression models to predict the physical parameters (response or predicted variables) from it. In the context of machine learning, constructing a regression model consists in using a training set (a set of cases defined by the selected variables for which the physical parameters are available) to infer the parameters of a mapping between the feature set and predicted variables. The regression model parameters should not be confused with the physical parameters of the atmosphere we aim at inferring.

In this case, our training set is again the BT-Settl collection of synthetic spectra, for which we already computed the feature set as part of the GA selection procedure. Of course, for each spectrum we also have available the effective temperature, the gravity and the metallicity. Once the model is trained, we will apply it to observed spectra as described in Sections 3 and 4.

Several regression models are trained for the prediction of each physical parameter in order to evaluate their performance:

(i) Bagging with multiadaptive spline regression models (hereafter MARS).

[1] All computations needed for this work were carried out in the CeSViMa (http://www.cesvima.upm.es/) power7 HPC characterized by processors with eight cores and four threads per core, running at 3.3 GHz and with 32 Gb of RAM each.

(ii) Random forest regression models (RF).
(iii) *k*-nearest neighbours (KNN).
(iv) Generalized boosted regression models (GBM).
(v) Support vector regression with Gaussian kernel (SVR).
(vi) Multi-layer perceptron neural networks (NNET).
(vii) Kernel partial least-squares regression (KPLS).
(viii) Rule regression models (RR).

In order to assess the validity of our feature sets we also compare the predictions based on them with other input spaces. In particular, we also compute physical parameters that yield the the minimum χ^2, and train a projection pursuit regression model with the independent components (Hyvärinen 1998) derived from each spectrum.

Including here a sufficient description of each and every regression model that we trained would render the manuscript excessively lengthy but interested readers can find additional information in Baraud (2002), Geman, Bienenstock & Doursat (1992), Elith, Leathwick & Hastie (2008), Meyer, Leisch & Hornik (2003), Svetnik et al. (2003). Suffice it to say that each one of them can be thought of as a parametric model that predicts one physical parameter from an input vector. The input vector can be the full normalized spectrum, the ICA lower-dimensional representation of the full spectrum, the spectral features selected by Cesetti et al. (2013) or those selected by the GA. The regression model parameters are inferred (using strategies that differ from one regression model to another) from a set of examples. As explained above, this set of examples (spectra of stars for which we know the physical parameters) is called the training set, and the process by which the model parameters are determined from the training set is called training of the model. In the next paragraph we give minimal details of each regression model trained, and references for the interested reader.

In order to avoid the well known problem of overfitting (see e.g. Dietterich 1995), we use five-fold cross validation to estimate the prediction errors. *n*-fold cross validation consists in dividing the training set into *n* disjoint subsets and training different regression models, each one of them with $(n-1)$ of the *n* subsets. The *n*th subset not used for training is used instead to estimate the errors.

As every type of model has its own set of tuneable parameters as well as its own training procedure, we have used a common R (R Core Team 2016) wrapper for all models named CARET (short for Classification And REgression Training, Kuhn 2008). This wrapper enables a common interface, as well as the use of the same set of training/set samples for the adopted five-fold cross-validation error estimation. As explained above, each regression model has its own set of model parameters. For each model we have searched for the parameter set that minimized the root mean square error (RMSE) in a grid of values defined *ad hoc* for each technique.

The adopted procedure for learning the models can then be summarized as the pseudocode 2.1.

Algorithm 2.1: MODEL LEARNING($DataSet$, $ParRanges$)

$\mathcal{S}_{ModelParameters} \leftarrow ParRanges$
$\mathcal{S}_{DataFolders} \leftarrow Preprocess(DataSet)$
for each $x \in \mathcal{S}_{ModelParameters}$
do
  for each $z \in \mathcal{S}_{DataFolders}$
  do
    $HDS(z) \leftarrow$ Hold-out specific samples
    $Model(z) \leftarrow Fits(\mathcal{S}_{DataFolders} \setminus HDS(z))$
    $Perf(z) \leftarrow Predicts(Model(z), HDS(z))$
  $Perf(x) \leftarrow Average(Perf(z)) \quad \forall z \in \mathcal{S}_{DataFolders}$
$OPS \leftarrow argmax(Perf(y)) \quad \forall y \in \mathcal{S}_{Modelparameters}$
$Model \leftarrow Fits(DataSet, OPS)$

In the description of Algorithm 2.1 *ParRanges* represents the set of available parameter ranges, which are organized into different sets named $S_{ModelParameters}$. Similarly, the available *DataSet* will be used to create the five disjoint data folders named $S_{DataFolders}$. Then, the learning procedure sequentially combines all the data folders excluding one ($HDS(z)$). By fitting the models according to the particular parameter set and training data, the algorithm produces models $Model(z)$. These models can now be scored against the unseen data folder to yield $Perf(z)$. The selection of the most suitable model configuration (*Model*) will be based on the parameter set *OPS*, which maximizes the average performance for the available training data *DataSet*.

During the preprocessing stage (2.1) the spectral resolution of the BT-Settl library was degraded to the IRTF resolution (R $\sim$ 2000) by convolving with a Gaussian. Then, the spectra were trimmed to produce valid segments between 8145.92 and 24 106.85Å, which is the spectral range common to all M stars in the IRTF library. Finally, all spectra were divided by the total integrated flux in this range in order to factor out the stellar distance.

As mentioned above, the training set was constructed from the BT-Settl library of stellar spectra. The interested reader may find different approaches in the literature to the problem of finding an optimal set of training examples. Ness et al. (2015), for example, prefer to use real observed spectra rather than synthetic libraries to create a generative model in which the individual spectral fluxes are modelled as second-degree polynomials with the physical parameters as arguments. The real observed spectra have physical parameters taken from the literature, which in turn are almost always inferred using synthetic spectral libraries. In our opinion, this approach does not solve the dependence of the predicted parameters on the necessarily imperfect synthetic libraries, but has the advantage that the relative frequencies of examples in the training set represent better the biases naturally encountered in surveys than the uniform sampling of parameter space found in synthetic libraries. Recently, Heiter et al. (2015) have started a program to compile a set of stars with accurate physical parameter determinations inferred independently of spectroscopic measurements and atmospheric models (as much as possible). Unfortunately, this ambitious program only contains 34 stars of spectral types F, G and K. In the M regime we find similar approaches in Boyajian, van Belle & von Braun (2014) and references therein, where the atmospheric parameters are derived using interferometric measurements of stellar radii. Again, this only amounts to a very small number (21 K and M stars) of examples and a very sparse sampling of the parameter space.

All efforts to compile training sets of stars with accurate, homogeneous, and reliable physical parameters derived independently of spectroscopic measurements are valuable not only because they allow for the improvement of the stellar atmospheric models but also because they help increase the reliability of the regression models by making them independent of these atmospheric models. But until these training sets with sufficient and homogeneous sampling of the parameter space are available, we must turn to the use of synthetic libraries.

3 PHYSICAL PARAMETERS OF THE IRTF COLLECTION OF SPECTRA

In the following, we will summarize the results obtained for the IRTF data set. We deal with the different physical parameters in separate sections. We start by reporting the root mean/median square errors (RMSE/RMDSE) with respect to the parameters gathered

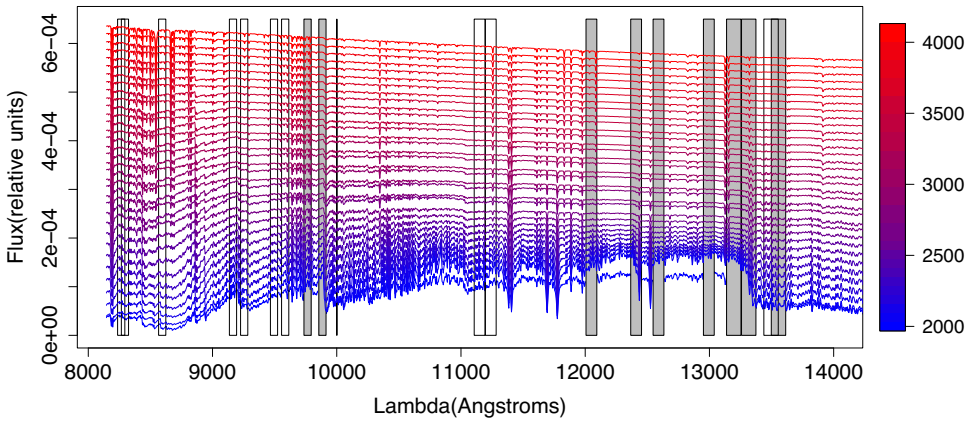

Figure 1. Features selected by the GA for predicting $T_{\rm eff}$ using noiseless BT-Settl synthetic spectra in the IRTF wavelength range and resolution. The BT-Settl spectra are plotted in a colour scale that ranges from blue (2000 K) to red (4100 K). The empty boxes correspond to the selected features and the grey boxes to the continuum bands.

from the literature by Cesetti et al. (2013) and included in their table 3.

We report both the mean and the median square errors because the accuracy estimates are often dominated by a small subset of the spectra that produce errors outlying the overall distribution. Whenever the mean and median square errors differ significantly, we can deduce that this is the case and hence the mean is not representative of the typical errors.

3.1 Spectral bands selected

The GAs were applied to the selection of features for the prediction of effective temperature from both noiseless and noisy spectra. For the IRTF wavelength range and resolution, results in the features have been included in Table A1. Features are ordered by the fitness value (according to the AIC criterion, as explained in equation 3) and we only consider features that are present in at least five sets.

Table A1shows a very wide variety of features with very few repetitions. Only spectral features 4, 5, 6 and 9 in the SNR = 50 experiment are also found in the SNR = ∞ and SNR = 10 feature sets (albeit with different continuum definitions). This reinforces the impression that the information useful for the estimation of the effective temperatures is spread over the entire IRTF spectrum.

For gravity estimation (on a logarithmic scale) and metallicity, the GA search procedure produces the features presented in Tables A2 and A3, respectively. Fig. 1 shows a graphical representation of the bands selected for the determination of the effective temperature superimposed on a set of noiseless BT-Settl spectra.

3.2 Regression models

3.2.1 Effective temperature models

Table C1 summarizes the RMSE/RMDSE for the complete set of models: the minimum χ^2 estimate based on the full spectrum (χ^2), the projection pursuit regression based on the ICA components (PPR-ICA) and models trained on the spectral features proposed by the GA (GA-RF, GA-GBM, GA-SVR, GA-NNET, GA-MARS, GA-KPLS, GA-RR). For each model, we report the RMSE/RMDSE obtained for several noise levels of the training sets. SNR = ∞ corresponds to noiseless spectra. In the GA cases, models are trained with the spectral features found by the genetic algorithms when applied to BT-Settl spectra of the corresponding SNR.

Table C1 shows that the performance of classifiers based on the full spectrum (or in a compressed version in the form of ICA components) and the best classifier based on features derived from limited spectral bands is equivalent. The Bartlett test shows that the variances are homogeneous with a Bartlett' s K^2 of 8.5 with two degrees of freedom and a p-value of 0.014 26. The Fligner–Killeen test shows that homoscedasticity is verified at the $p = 0.005\,886$ level. Finally, the F-ANOVA test clearly shows that there is no significant difference between models. Thus, we conclude that the quality of features from the two approaches (full spectrum and selected features) is equivalent in predictive performance. The difference between the performances of the best classifier (GA-KNN; best on average over SNR), the minimum χ^2 classifier, and the PPR-ICA classifiers are not statistically significant. In any case, it is evident that the RMSE is significantly above the grid spacing in temperature. We interpret the small differences as an indication that there is as much information spread over the entire spectrum shape as can be distilled from a few spectral bands.

The comparison with the effective temperatures compiled by Cesetti et al. (2013) shows, however, some significant differences across models when evaluated not by the RMSE/RMDSE, but by the average bias (see Table C2).

In general, all regression models tend to predict lower effective temperatures than those in the literature except in the noiseless scenario. The models trained with noiseless spectra tend to overestimate $T_{\rm eff}$, suggesting that the optimal SNR is between SNR = 50 and ∞. The minimum χ^2 approach and the GA-KNN model systematically underestimate $T_{\rm eff}$ for all SNR regimes. This shared behaviour is not surprising since minimum χ^2 is a single-nearest-neighbour method applied in the space of the entire spectrum as opposed to the space-selected features.

We have found in previous studies that, at least for input spaces constructed from ICA compressions of the spectra, it is not necessary to adapt the training set SNR to match exactly that of the prediction set. On the contrary, we find that two regimes are sufficient to obtain acceptable results. The two regimes are separated at SNR = 10. The model trained with SNR = 50 spectra gives close to optimal results for spectra with SNRs above 10, while below that limit the same situation holds for the model trained with SNR = 10 spectra (González-Marcos et al. 2017).

Fig. 2 shows the correlation between the $T_{\rm eff}$ estimates of the best (in the RMDSE sense) regression models and the effective temperatures in table 3 of Cesetti et al. (2013). It is worth noting that in the M-star regime, there are 63 effective temperatures available

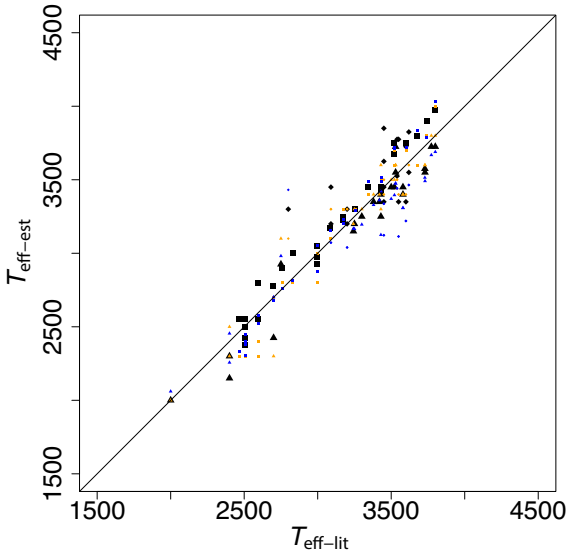

Figure 2. Comparison of the effective temperatures from the literature included in Cesetti et al. (2013) and those inferred from the KNN model trained with the GA features (black). In orange and blue we show the estimates from the minimum χ^2 estimate and from the PPR model based on the ICA components, respectively.

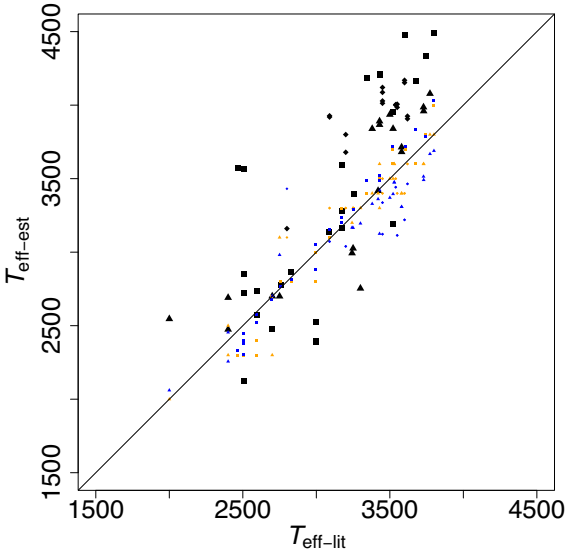

Figure 3. Comparison of the effective temperatures from the literature included in Cesetti et al. (2013) and those inferred from the NNET model trained with the features introduced in Cesetti et al. (2013) (black). In orange and blue we show the estimates from the minimum χ^2 estimate and from the PPR model based on the ICA components, respectively.

in Cesetti et al. (2013), and 46 of the 63 were estimated from the spectral types using the calibrations of Ostlie & Carroll (2007). We have substituted them with a spline fit to the combined data sets of M dwarfs in Rajpurohit et al. (2013) and Boyajian et al. (2012) as it removes systematic biases in the temperatures between 2500 and 3500 K.

It is not evident that the GA-KNN model performs significantly better than the minimum χ^2 estimate, but in the following we will retain the former for further analysis. Fig. 2 shows that the GA features can be used to estimate effective temperatures with an accuracy equivalent to that yielded by full wavelength range spectra of the same resolution.

We have trained the same nonlinear regression models discussed above using the features suggested by Cesetti et al. (2013). The performance of the models based on these features is included in Table C3. From the comparison of Tables C1 and C3 we can draw the following conclusions:

(i) The RMSE for SNR = 10 and 50 is equivalent for the regression models trained on GA features and those recommended in Cesetti et al. (2013).

(ii) However, the RMDSE from Cesetti et al. (2013) is significantly higher in the case of the features for all SNR values.

(iii) In the unrealistic case of noiseless spectra, the features proposed by Cesetti et al. (2013) produce RMSE and RMDSE significantly worse than the GA features.

However, the cross-validation errors are far from informative with respect to the true performance when the models are applied to real data. In the case of the features defined by Cesetti et al. (2013), we find that the best RMSE/RMDSE (obtained not from cross validation but from the comparison with the effective temperatures in table 3 of Cesetti et al. (2013)) is attained by the CES-NNET model. Fig. 3 shows a graphical comparison of the CES-NNET predictions with the table 3 reference values. Again, and in the rest of this work, we substitute the effective temperatures in that table, which were estimated using the Ostlie & Carroll (2007) calibration, with those from the spline fit described above.

Fig. 3 shows that the features found by the GA are to be preferred to the ones proposed by Cesetti et al. (2013). We would like to emphasize that Figs 2 and 3 compare the predictions of the best regression models trained with the features derived by the GA or by the sensitivity maps described in Cesetti et al. (2013) with the effective temperatures gathered from the literature. These literature estimates are not free of errors and therefore the comparison of the two figures only reveals the better adequacy of the GA features to reproduce these literature values.

3.2.2 *Surface gravity models*

For the validation of our models, we only have 10 literature values of the surface gravity available in table 3 of Cesetti et al. (2013). Unfortunately, this is too small a number to draw significant conclusions on the comparison of methodologies from external data. Hence, we are left only with plausibility arguments for the selection of models. In this section $\log(T_{\rm eff})$–$\log(g)$ diagram comparisons will be used to select the most plausible model results.

An important difference with respect to the models discussed above is that we use the $T_{\rm eff}$ estimated in the previous stage as input in our models. It introduces an average improvement in the RMSE/RMDSE of 20 per cent with respect to the models without input information on the effective temperature although it represents a risk if the $T_{\rm eff}$ estimate is in gross error.

Table C4 shows the RMSE and RMDSE of the cross-validation experiments for the $\log(g)$ regression models and the same SNR regimes discussed for the estimation of $T_{\rm eff}$. We have assessed the models according to plausibility arguments relative to the distribution of the model predictions in $T_{\rm eff}$–$\log(g)$ diagrams. Fig. 4 shows this distribution for four models selected based on these plausibility criteria: GA-RR, GA-PLS, GA-KNN (all three of them for SNR = 50), and PPR-ICA (clockwise, starting at the top left). These models that produce the most plausible $\log(T_{\rm eff})$–$\log(g)$ diagrams are not amongst the best performing in terms of RMSE or RMDSE except

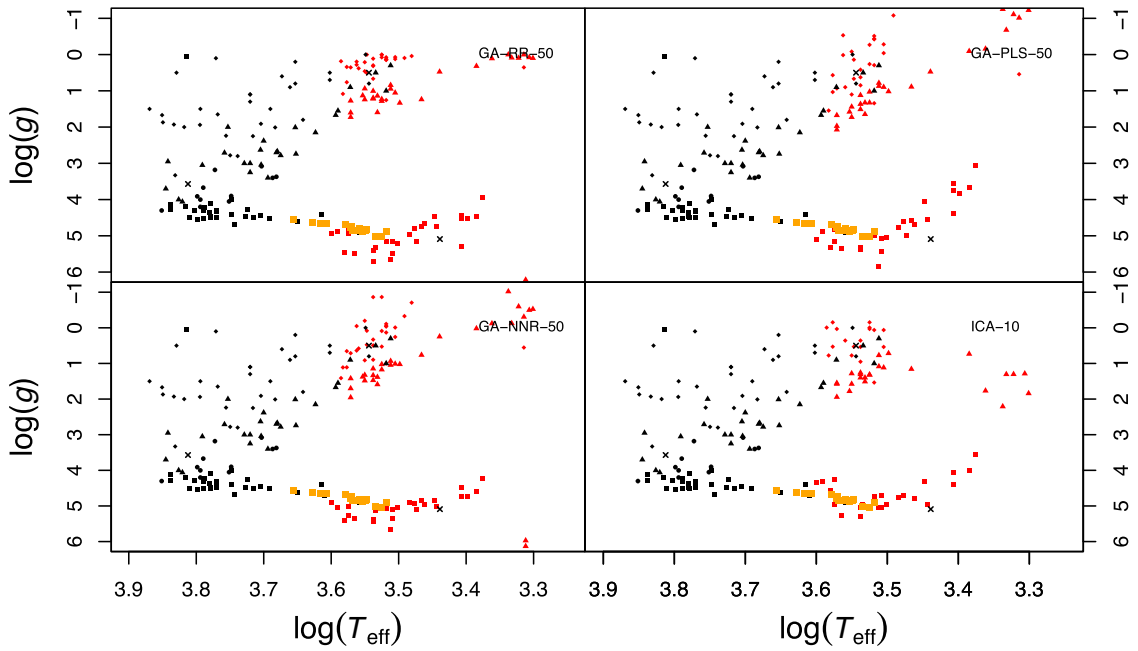

Figure 4. log ($T_{\rm eff}$)–log (g) diagrams produced by the GA-KNN (SNR = ∞) effective temperatures and gravities derived with the GA-RR (SNR = 50), GA-PLS (SNR = 50), GA-NNET (SNR = 50) and PPR-ICA-10 models (clockwise, starting from the top left). Squares represent dwarfs stars, triangles represent luminosity class III and circles represent luminosity classes I and II. Black symbols represent values taken from Cesetti et al. (2013), orange symbols correspond to values derived from high-resolution infrared spectra by Lindgren et al. (2016) and Lindgren & Heiter (2017), and red symbols are our own predictions.

in the case of the PPR-ICA model. This is another indication that the cross-validation errors are not good predictors of the true performance on real spectra. All four panels show a tendency towards lower surface gravities at the coolest end, and a reasonable capability to separate dwarfs from giants, and giants from supergiants. We include as orange squares the recent predictions by Lindgren, Heiter & Seifahrt (2016) and Lindgren & Heiter (2017) for a set of sources with high-resolution infrared spectra. The estimates are in reasonable agreement with the extrapolation of the distributions that can be guessed from the values in table 3 of Cesetti et al. (2013) (represented by black filled symbols). Since we cannot propose quantifications of this plausibility argument valid for all luminosity classes, we let the reader decide which estimate (GA-RR, GA-PLS, GA-KNN or ICA-10) is to be preferred. If we judge only by the concordance with the locus defined by the M dwarfs in the studies by Lindgren et al. (2016) and Lindgren & Heiter (2017), then ICA-10 is to be preferred.

Fig. E1 shows the equivalent diagram for predictions obtained from the features selected in Cesetti et al. (2013). Again, we select the regression models that yield the most plausible log ($T_{\rm eff}$)–log (g) distributions with no quantitative criteria defined to select the models. The superiority of the GA-based features over those defined by Cesetti et al. (2013) is again evident.

3.2.3 *Metallicity models*

Finally, the same regression models are trained to infer the metallicity, again considering the effective temperature as an input feature as in the log (g) regression models. Table C5 shows the RMSE and RMDSE obtained for the cross-validation experiments of each regression model. The minimum cross-validation errors are consistently obtained with the minimum χ^2, PPR-ICA and GA-KNN (with some exceptions). The differences from these cross-validation experiments are only marginal, but we see that even at these intermediate resolutions the reduction of dimensionality (either with ICA or GA) produces an improvement in the predictions.

This is even more evident if we compare our predictions with more recent metallicity estimates not included in Cesetti et al. (2013). We have gathered estimates for stars in both the IRTF collection and a series of recent metallicity catalogues by Rojas-Ayala et al. (2012), Neves et al. (2013), Newton et al. (2014), Gaidos et al. (2014), and Mann et al. (2015). All of the aforementioned references provide us with estimates of the iron abundance ratio [Fe/H] except Rojas-Ayala et al. (2012), which provides both the overall metallicity [M/H] and the [Fe/H] ratio. Our estimates, coming from the BT-Settl library, are for the [M/H] ratios, so some offset could be expected from the different nature of the quantities compared. Hence, when comparing our estimates with those from the literature, we compute the RMSE or RMDSE after subtracting any difference in the mean. It turns out that, after correcting for these different scales, PPR-ICA trained with SNR = 10 examples yields the lowest RMSE/RMDSE. Fig. 5 represents the estimates of [M/H] obtained from the PPR-ICA-based regressor, as a function of the values taken from these references for the sources in common. The black empty circles represent values from Cesetti et al. (2013); orange filled circles, values from Neves et al. (2013); green filled squares, values that the Vizier catalog entry for table 8 of Neves et al. (2013) links to Jao et al. (2005), although we find no evidence that Jao et al. (2005) contains estimates of metallicities; cyan and blue filled squares, values of [M/H] and [Fe/H], respectively, in Rojas-Ayala et al. (2012); red filled squares, values from Mann et al. (2015); yellow filled squares, values from Newton et al. (2014); and, finally, black filled squares, values from Gaidos et al. (2014).

It is remarkable that the minimum χ^2 predictions result in a 50 per cent increase in the RMDSE with respect to the ICA-10

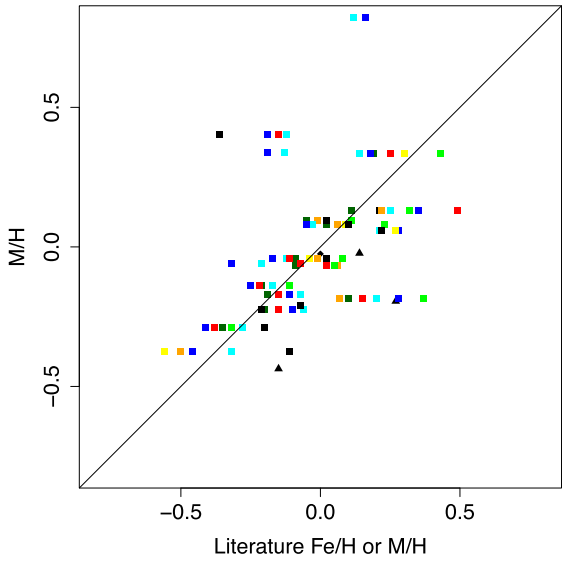

Figure 5. Comparison between metallicity estimates from the literature and predictions from the PPR-ICA (SNR = 10) model. Black empty circles represent values from Cesetti et al. (2013); orange filled circles, values from Neves et al. (2013); green filled squares, values that the Vizier catalog entry for table 8 of Neves et al. (2013) links to Jao et al. (2005), although we find no evidence that Jao et al. (2005) contains estimates of metallicities; cyan and blue filled squares, the values of [M/H] and [Fe/H], respectively, in Rojas-Ayala et al. (2012); red filled squares, values from Mann et al. (2015); yellow filled squares, values from Newton et al. (2014); and, finally, black filled squares, values from Gaidos et al. (2014).

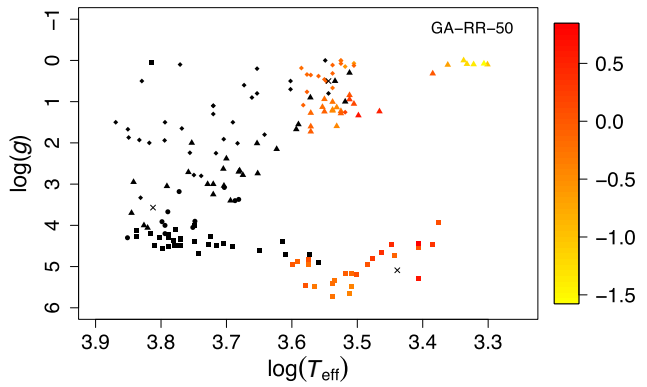

Figure 6. Plane of predictions for $T_{\rm eff}$ (from GA-KNN-∞) and $\log(g)$ (from GA-RR-50) with metallicity predictions from PPR-ICA-10.

models and the best-performing GA-based models. Therefore, we advise against their use in the context of metallicity estimations.

Fig. 6 summarizes the predictions from a set of selected regression models in $T_{\rm eff}$ (from the GA-NN-∞ model), $\log(g)$ (from the GA-RR-50 model) and metallicity (from the PPR-ICA-10 model). Again, plausibility arguments such as the lower metallicity of the supergiants apparently give evidence supporting the good performance of our models, but the lack of extensive good-quality estimates of the physical parameters of the IRTF collection of stars prevents us from a more quantitative assessment of the predictions.

The equivalent plot for predictions based on the features defined by Cesetti et al. (2013) and the RF model trained with SNR = 50 spectra is included as Fig. E2.

4 PHYSICAL PARAMETERS OF THE DWARF ARCHIVES COLLECTION OF SPECTRA

4.1 Spectral bands selected

As for the IRTF spectra, the spectral resolution of the BT-Settl library was degraded to match the average resolution of the spectra in the Dwarf Archives.[2] Then, the spectra were trimmed to produce segments in the spectral range common to all spectra of M stars in the archive, to avoid missing data in the input variables. Finally, all spectra were divided by the total integrated flux in this range in order to factor out the stellar distance.

There is little hope a priori for reasonable accuracies with regression models that predict the surface gravity and metallicity from such wavelength-limited, low-/intermediate-resolution spectra. Anyhow, we provide the results obtained applying the same methodology as in Section 3 (and described in Section 2) to show the limitations.

The application of the GA to the selection of features for the prediction of effective temperature within the Dwarf Archives wavelength range and resolution results in the features included in Table B1. Table B2 shows the spectral features selected for predicting the surface gravity, and Table B3 those for predicting metallicity.

4.2 Regression models

In the following, we will summarize the results obtained for the Dwarf Archives data set. We deal with the different physical parameters in separate sections. We start by reporting the cross-validation root mean square errors (RMSE) and root median square error (RMDSE) for the five-fold cross-validation strategy, and we subsequently discuss the accuracy of the predictions with respect to literature values where available.

4.2.1 Effective temperature models

Table D1 summarizes the RMSE/RMDSE for the complete set of models: the minimum χ^2 estimate based on the full spectrum (χ^2), the projection pursuit regression based on the ICA components (PPR-ICA) and some models trained on the spectral features proposed by the GA (GA-RF, GA-GBM, GA-SVR, GA-NNET, GA-MARS, GA-KPLS). For each model, we report the RMSE/RMDSE obtained for several noise levels of the training sets.

Again, as in the IRTF case, we see that the compression of the spectra results in a clear performance degradation with respect to the χ^2 minimization technique results. Fig. 7 shows a comparison between the effective temperatures derived from a spectral type calibration (x axis) and the predictions of the best regression models (y axis). In particular, we have converted the spectral types available in the DwarfArchives.org collection to effective temperatures using the same spline fit described in the IRTF section. This shows that the best results based on the 10 features selected by the GA are barely equivalent to the prediction accuracy of the χ^2 estimates. The decrease in the number of predictive variables results in a simpler and faster model, but the χ^2 model is already simple, so our conclusion is that the feature selection in this context of low-resolution spectra in the optical range is unnecessary.

Having shown that the feature selection with GAs degrades the performance of regression models, one can wonder whether a

[2] http://spider.ipac.caltech.edu/staff/davy/ARCHIVE/index.shtml

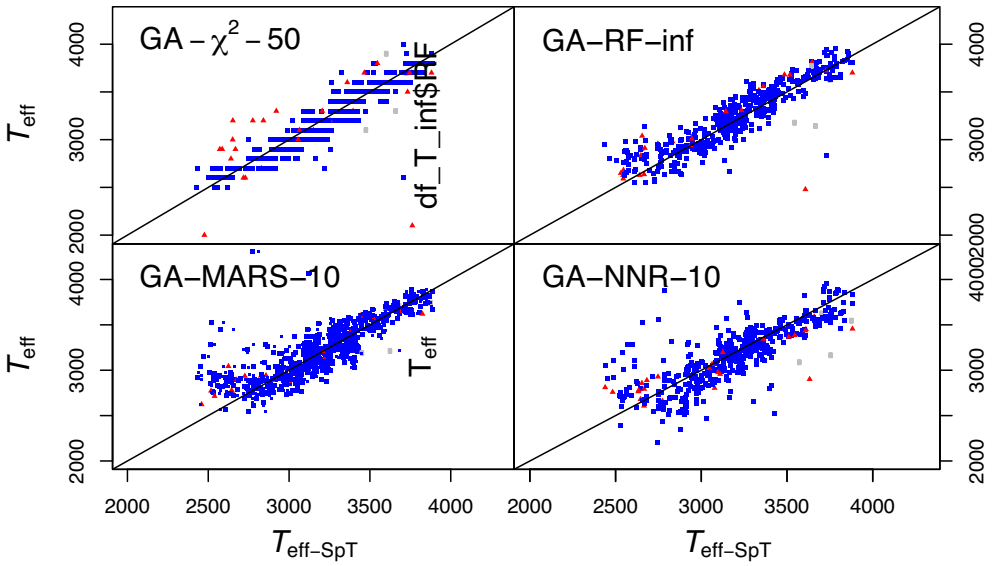

Figure 7. Comparison between the effective temperatures derived from the tabulated spectral types in Cesetti et al. (2013) (x axis) and those inferred by the various regression models (y axis): χ^2 model (top left, SNR = 50), random forest regression model (top right, SNR = ∞), GA-MARS model (bottom left, SNR = 10), and the neural network model (bottom right, SNR = 10). Blue squares denote Main Sequence dwarfs and red triangles denote giant stars (luminosity class III) according to Cesetti et al. (2013).

different feature selection procedure would produce better results. In particular, we investigate the possibility that the features proposed by Cesetti et al. (2013) result in a performance equal to or even better than the one achieved with χ^2.

We train the same types of regression models to the features selected in Cesetti et al. (2013), again learning from BT-Settl spectra of various SNRs and predicting over the Dwarf Archives set. A summary of the results can be found in Table D2, where we use CS- to indicate that the model was trained using the features by Cesetti et al. (2013).

For SNR = 10, the best GA models (GA-KPLS in RMDSE or GA-RF in RMSE) outperform the best CS model (GA-GBM). For SNR = 50 the situation depends on the figure-of-merit used to compare the regression models: in RMSE the best model is CS-GBM while in RMDSE GA-GBM outperforms all CS models. Finally, for the unrealistic case of noiseless spectra, Table D2 shows an overwhelming degradation of the prediction accuracy from CS features. But even in the only case where the CS features outperform those selected by the GA, the performance is below the one achieved by the minimum χ^2 approach. It is important to remark here that the features selected in Cesetti et al. (2013) were fit for the IRTF wavelength range and resolution, and not all of them can be extracted from Dwarf Archives spectra. Hence, unlike in the case of the IRTF spectra, the comparison of GA- and CS-based performances with Dwarf Archives spectra is not fair and the results are only included for the sake of completeness.

The relationship between the GA predicted effective temperature and the one measured by Rojas-Ayala et al. (2012) (blue for the predictions by the GA-MARS SNR = 10 model and black for the GA-RF SNR = ∞ model), and by Lindgren et al. (2016) and Lindgren & Heiter (2017) (orange and green for the same models as before) can be found in Fig. 8.

4.2.2 *Surface gravity models*

As in the IRTF exercise, we attempt to select features for surface gravity estimation from BT-Settl spectra using GAs despite the much lower spectral resolution and smaller wavelength coverage of the Dwarf Archives spectra. Since there is no substantive compilation of surface gravities that we could cross-match with the IPAC list of M stars in the Dwarf Archive, we are left with the same plausibility arguments used in the IRTF study, which are based on the log ($T_{\rm eff}$)–log (g) diagram.

We again use the effective temperatures as input of the regression models. Table D3 shows the cross-validation RMSE and RMDSE for the same set of regression models used throughout this article.

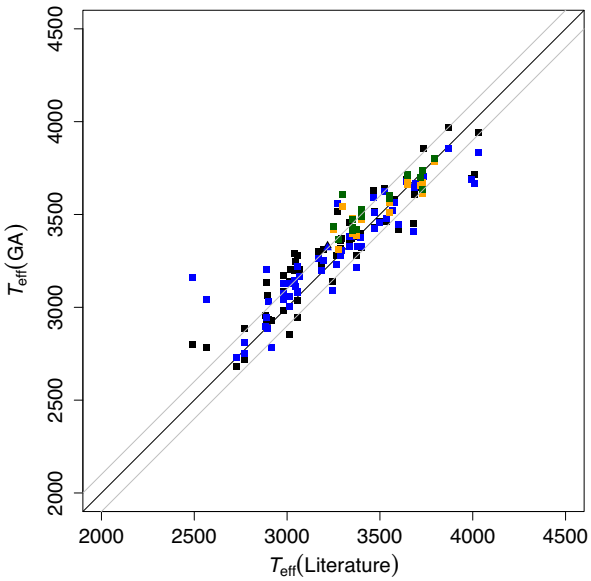

Figure 8. Relationship between log ($T_{\rm eff}$) from Rojas-Ayala et al. (2012) on the x axis and log ($T_{\rm eff}$) as predicted by the GA-RF model with SNR = ∞ (black symbols, squares for dwarfs and a triangle for the only giant in the sample) and the GA-MARS trained with SNR = 10 data (blue symbols). Green and orange symbols correspond to sources in common with Lindgren et al. (2016) and Lindgren & Heiter (2017) and the predictions by the GA-RF (SNR = ∞) and GA-MARS (SNR = 10) models, respectively.

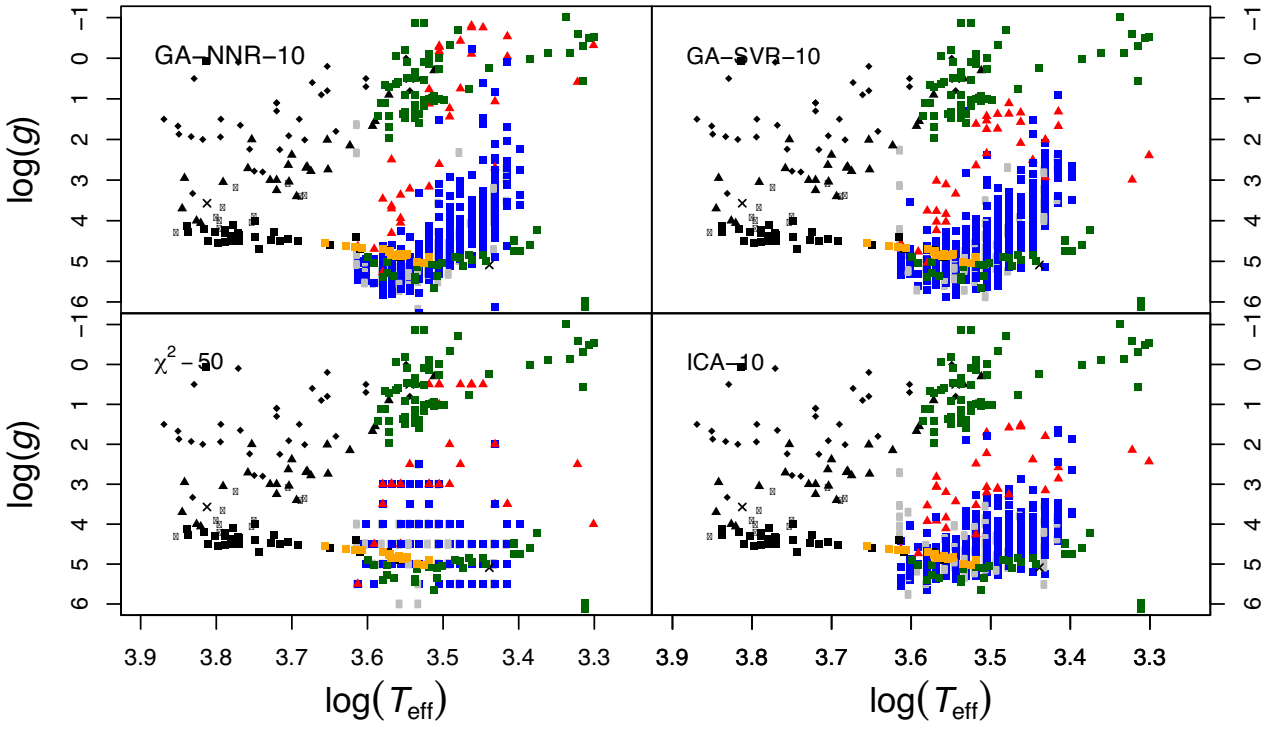

Figure 9. $\log(T_{\rm eff})$–$\log(g)$ planes obtained using the χ^2 (SNR = 50) $\log(T_{\rm eff})$ predictions and the $\log(g)$ values from the GA-NNET (SNR = 10, top left), GA-SVR (SNR = 10, top right), χ^2 (SNR = 50, bottom left) and PPR-ICA (SNR = 10, bottom right) regression models. Black symbols correspond to objects with physical parameters in table 3 of Cesetti et al. (2013); green symbols correspond to the predictions shown in Section 3.2.1 for the IRTF spectra; blue squares correspond to predictions for dwarf stars according to the DwarfArchives.org luminosity classes; red triangles correspond to giant stars according to DwarfArchives.org; orange symbols correspond to values derived from high-resolution infrared spectra by Lindgren et al. (2016) and Lindgren & Heiter (2017); and, finally, empty grey circles correspond to sources with no luminosity class in DwarfArchives.org.

It shows that the GA-RF model outperforms all others in all SNR regimes, giving a consistent RMDSE of 1.0 dex. Obviously, this is barely enough for classification in luminosity classes. Furthermore, and as has been the case in the evaluation of all previous regression models, the cross-validation errors are poor estimates of the true performances on real observed spectra. Fig. 9 shows the $\log(T_{\rm eff})$–$\log(g)$ diagram for the two best-performing $\log(g)$ regression models (GA-NNET and GA-SVR, both trained with SNR = 10 BT-Settl spectra) and the two reference models based on the minimization of the χ^2 (SNR = 50) and the PPR-ICA (SNR = 10). Three of the four panels (all except the χ^2 predictions) show a strong spurious correlation in the gravities of the M dwarfs in the sense that the coolest dwarfs have unreasonable values of $\log(g)$ around 2 dex. The PPR-ICA model shows this trend too, but with a much shallower slope ($\log(g) \approx 4$ at $\log(T_{\rm eff}) \approx 3.4$). Only the GA-NNET model (and to a lesser extent the χ^2 model) places giant stars (red triangles) in the locus expected, judging by the values in table 3 of Cesetti et al. (2013) (black filled symbols) and the predictions for the IRTF data set (green filled symbols). The various symbols (squares, triangles, circles) reflect the luminosity classes found in either table 3 of Cesetti et al. (2013) (for the IRTF values) or the DwarfArchives table.

Fig. E3 shows the $\log(T_{\rm eff})$–$\log(g)$ diagram for predictions obtained from the features selected in Cesetti et al. (2013) and the Dwarf Archives data. As in the IRTF case, we select the regression models that yield the most plausible $\log(T_{\rm eff})$–$\log(g)$ distributions with no quantitative criteria defined to select the models. It shows the superiority of the GA-based features over those defined by Cesetti et al. (2013).

4.2.3 *Metallicity models*

Finally, the same analysis is performed for metallicities, again using the previously inferred temperature as a fixed input feature. Table D4 shows a summary of the cross-validation performance of the different models.

In general, models trained with SNR = ∞ show much poorer performance except for the GA-RF and GA-GBM cases. The best χ^2 model produces errors almost a factor of two larger than the GA-RF-∞ model (although it has to be borne in mind that, while our regression models are capable of predicting metallicities that are intermediate in the grid, the minimum χ^2 can only yield values in the grid, which has a step size of 0.5 dex). Models trained with SNR = 10 and 50, in contrast, show a more consistent behaviour for the entire set of regressors, with poorer performances than the apparently optimal GA-RF-∞, but also smaller differences between models.

In order to select the best model, we again compare our model predictions with the reference catalogues used in Section 3.2.3. We select the random forest trained with noiseless synthetic spectra as the best model, which renders the minimum RMSE (0.3 dex). Fig. 10 shows the comparison of our estimates with the reference catalogues, using the same symbols and colours as in Fig. 5.

Our value of the RMSE contrasts with the differences between estimates for the same star in the literature. We obtain a mean difference of 0.1 dex, a factor of three smaller than our RMSE.

It is interesting to note that our predictions extend to metallicities as low as [M/H] = −2.1. Fig. 11 shows a histogram of the metallicities predicted by the GA-RF-∞ model for the Dwarf Archives set of spectra. We find predictions below −1.5 for 11 sources, six

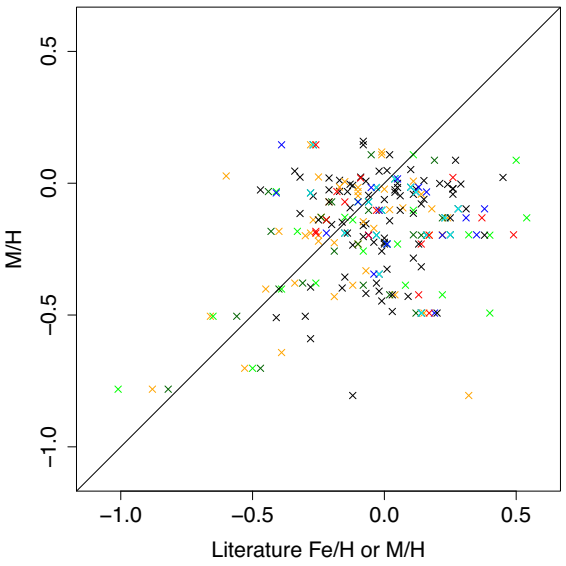

Figure 10. Relationship between the RF-∞ predictions for metallicity and values from the literature for stars in the Dwarf Archives. Black symbols represent values from Cesetti et al. (2013); orange symbols, values from Neves et al. (2013); green symbols, values that the Vizier catalog entry for table 8 of Neves et al. (2013) links to Jao et al. (2005), although we find no evidence that Jao et al. (2005) contains estimates of metallicities; cyan and blue symbols, the values of [M/H] and [Fe/H], respectively, in Rojas-Ayala et al. (2012); red symbols, values from Mann et al. (2015); yellow symbols, values from Newton et al. (2014); and, finally, black symbols, values from Gaidos et al. (2014).

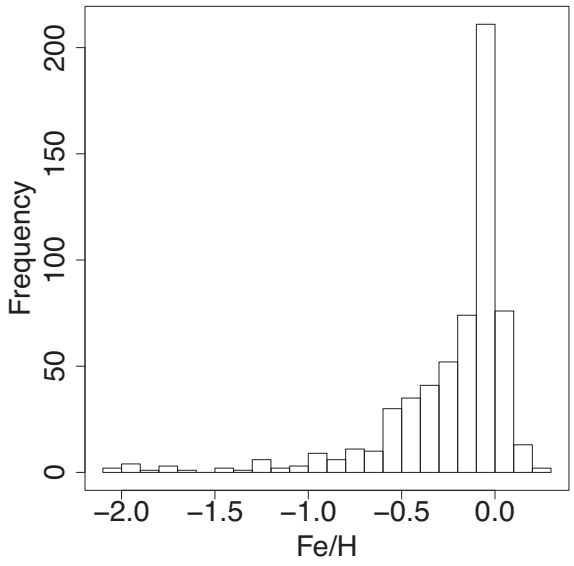

Figure 11. Predictions of the GA-RF-∞ regression model for the metallicity of Dwarf Archives stars.

of which have been previously identified as subdwarfs of different categories (see Table 1).

The remaining five stars with metallicities below −1.5 are 2MASS J17275631−3240430 (−2.0 dex); LHS 1625 (−1.97 dex); 2MASS J19215188+2802275 (−1.9 dex); 2MASS J19004675+2806462 (−1.7 dex), classified as K7III by Kirkpatrick et al. (1994); and 2MASS J14465233−5320580 (−1.7 dex).

Table 1. Previously known subdwarfs in the Dwarf Archives collection of spectra and the corresponding GA-RF-∞ predictions.

Identifier	Classification	Reference	GA-RF-∞
LHS 3768	usdM3	Kirkpatrick, Henry & Simons (1995)	− 2.1
LHS 2352	esd	Kirkpatrick et al. (1995)	− 2.0
LHS 1691	usdM2	Lépine, Rich & Shara (2007)	− 1.95
LHS 2023	esdM6	Riaz, Gizis & Samaddar (2008)	− 1.95
LHS 515	esdM5	Reid & Gizis (2005)	− 1.8
LP471-17	sdM	Kirkpatrick et al. (1995)	− 1.7

5 SUMMARY AND CONCLUSIONS

In this work we have attempted to construct regression models to predict physical parameters of M-type-star atmospheres. We have tried several representation spaces or sets of predictive variables: the full spectrum, the ICA compression coefficients, and several sets of features (pseudo-equivalent widths) optimized using GAs to predict effective temperatures, surface gravities and metallicities. The main conclusions for this extensive study can be summarized as follows:

(i) The cross-validation root mean square errors based on a training set of synthetic spectra are poor estimates of the true performance of a regression model applied to true observed spectra. As stated in Section 3.2.1, the errors estimated with the cross-validation experiments are much lower than those found by comparison with the literature values collected in Cesetti et al. (2013). We interpret this as a direct consequence of (i) the differences between the BT-Settl library of synthetic spectra and the real spectra of M stars, and (ii) the internal errors in the literature values collected in Cesetti et al. (2013). There may also be a contribution due to noise excursions not properly accounted for by the Gaussian random noise added to the synthetic spectra, but we do not expect this contribution to be the major contributor to the difference between cross-validation error estimates and external validation with literature values.

(ii) The features selected by Cesetti et al. (2013) based on sensitivity maps (the gradient of the monochromatic fluxes as a function of the physical parameters) have sub-optimal performances when used for prediction purposes.

(iii) In the context of IRTF spectra ($R \approx 2000$ between 8146 and 24 107 Å), our feature set for predicting effective temperatures combined with a nearest-neighbour regression model produces similar results to those obtained from the χ^2 classical technique and a projection pursuit regression model based on the ICA compression coefficients. Hence, there is no apparent gain in reducing the dimensionality of the representation space other than the simplicity, interpretability and computation speed of the models.

(iv) For the prediction of the IRTF star surface gravities, and based on plausibility arguments, we find a significant improvement in the predictions obtained from machine-learning models (mainly rule regression and artificial neural networks) and the GA features with respect to minimizing the χ^2 of the full spectrum. While ICA remains a competitive alternative, it fails to produce predictions for the coolest giants in the sample that are consistent with the literature luminosity class. However, ICA shows the best agreement with the M dwarf stellar parameters derived from high-resolution infrared spectra in Lindgren et al. (2016) and Lindgren & Heiter (2017). In the case of metallicities, the ICA coefficients remain as the optimal representation space although, at these reduced resolutions, the accuracy of the predictions is low.

(v) In the context of predicting $T_{\rm eff}$ for Dwarf Archives optical spectra, dimensionality reduction is not necessary and may indeed be counterproductive as it seems to induce a bias for the lowest temperatures. The prediction of surface gravities seems hopeless in the representation spaces tested in this work, whether it is the full spectrum in a χ^2 minimization scheme, or a machine-learning algorithm applied to ICA coefficients or GA features.

(vi) Finally, although the typical dispersion of the predictions for metallicities of Dwarf Archives stars is large ($\approx$0.25 dex) we find that our model based on GA-selected features and a random forest regression model can detect subdwarfs known in the literature and we produce a list of five new candidates that need to be confirmed with higher-resolution spectra.

The models developed in this work and the tools to preprocess the spectra are available upon request to the first author as RData files.

ACKNOWLEDGEMENTS

This research has benefited from the M, L, T and Y dwarf compendium of spectra housed at DwarfArchives.org. The authors would like to acknowledge funding by the Spanish Ministry of Economy and Innovation through the grant AyA2014-55216. The IRTF library is provided by the University of Hawaii under Cooperative Agreement no. NNX-08AE38A with the National Aeronautics and Space Administration, Science Mission Directorate, Planetary Astronomy Program. This research has made extensive use of the R software (R Core Team 2016) and the CARET R package. Finally, the authors acknowledge the computer resources and technical assistance provided by the Centro de Supercomputación y Visualización de Madrid (CeSViMa).

REFERENCES

Adams F. C., Bodenheimer P., Laughlin G., 2005, Astron. Nachrichten, 326, 913
Allard F., Homeier D., Freytag B., Sharp C. M., 2012, in Reylé C., Charbonnel C., Schultheis M., eds, EAS Publications Series Vol. 57, Low-Mass Stars and the Transition Stars/Brown Dwarfs – EES2011. EAS Publications Series, p. 3 (arXiv:1206.1021)
Allard F., Homeier D., Freytag B., Schaffenberger W., Rajpurohit A. S., 2013, Memorie della Soc. Astron. Ital. Suppl., 24, 128
Alonso-Floriano F. J. et al., 2015, A&A, 577, A128
Baraud Y., 2002, ESAIM: Probability and Statistics, 6, 127
Bochanski J. J., Hawley S. L., Covey K. R., West A. A., Reid I. N., Golimowski D. A., Ivezić Ž., 2010, AJ, 139, 2679
Bonfils X. et al., 2013, A&A, 556, A110
Boyajian T. S. et al., 2012, ApJ, 757, 112
Boyajian T. S., van Belle G., von Braun K., 2014, AJ, 147, 47
Browning M. K., 2008, ApJ, 676, 1262
Casagrande L., Flynn C., Bessell M., 2008, MNRAS, 389, 585
Cesetti M., Pizzella A., Ivanov V. D., Morelli L., Corsini E. M., Dalla Bontà E., 2013, A&A, 549, A129
Charbonneau P., 1995, ApJS, 101, 309
Cushing M. C., Rayner J. T., Vacca W. D., 2005, ApJ, 623, 1115
De Geyter G., Baes M., Fritz J., Camps P., 2013, A&A, 550, A74
Dietterich T., 1995, ACM Comput. Surv., 27, 326
Dressing C. D., Charbonneau D., 2015, ApJ, 807, 45
Elith J., Leathwick J. R., Hastie T., 2008, J. Animal Ecology, 77, 802
Gaidos E. et al., 2014, MNRAS, 443, 2561
Gelman A., Carlin J., Stern H., Dunson D., Vehtari A., Rubin D., 2013, Bayesian Data Analysis, 3rd edn. Taylor & Francis, Abingdon
Geman S., Bienenstock E., Doursat R., 1992, Neural Comput., 4, 1
Goldberg D. E., 1989, Genetic Algorithms in Search, Optimization and Machine Learning, 1st edn. Addison-Wesley Longman, Boston, MA
González-Marcos A., Sarro L. M., Ordieres-Meré J., Bello-García A., 2017, MNRAS, 465, 4556
Heiter U., Jofré P., Gustafsson B., Korn A. J., Soubiran C., Thévenin F., 2015, A&A, 582, A49
Holland J. H., 1975, Adaptation in Natural and Artificial Systems. Univ. Michigan Press, Ann Arbor, MI
Hyvärinen A., 1998, Proc. 1997 Conf. Advances in Neural Information Processing Systems 10: NIPS '97. MIT Press, Cambridge, MA, p. 273
Jao W.-C., Henry T. J., Subasavage J. P., Brown M. A., Ianna P. A., Bartlett J. L., Costa E., Méndez R. A., 2005, AJ, 129, 1954
Kirkpatrick J. D., McGraw J. T., Hess T. R., Liebert J., McCarthy D. W., Jr, 1994, ApJS, 94, 749
Kirkpatrick J. D., Henry T. J., Simons D. A., 1995, AJ, 109, 797
Kuhn M., 2008, J. Statistical Software, 28, 1
Lépine S., Rich R. M., Shara M. M., 2007, ApJ, 669, 1235
Lindgren S., Heiter U., 2017, A&A, 604, A97
Lindgren S., Heiter U., Seifahrt A., 2016, A&A, 586, A100
Mann A. W., Brewer J. M., Gaidos E., Lépine S., Hilton E. J., 2013, AJ, 145, 52
Mann A. W., Feiden G. A., Gaidos E., Boyajian T., von Braun K., 2015, ApJ, 804, 64
Meyer D., Leisch F., Hornik K., 2003, Neurocomputing, 55, 169
Ness M., Hogg D. W., Rix H.-W., Ho A. Y. Q., Zasowski G., 2015, ApJ, 808, 16
Neves V. et al., 2012, A&A, 538, A25
Neves V., Bonfils X., Santos N. C., Delfosse X., Forveille T., Allard F., Udry S., 2013, A&A, 551, A36
Neves V., Bonfils X., Santos N. C., Delfosse X., Forveille T., Allard F., Udry S., 2014, A&A, 568, A121
Newton E. R., Charbonneau D., Irwin J., Berta-Thompson Z. K., Rojas-Ayala B., Covey K., Lloyd J. P., 2014, AJ, 147, 20
Newton E. R., Charbonneau D., Irwin J., Mann A. W., 2015, ApJ, 800, 85
Ostlie D., Carroll B., 2007, An Introduction to Modern Stellar Astrophysics. Pearson Addison-Wesley, Boston, MA
Passegger V. M., Wende-von Berg S., Reiners A., 2016, A&A, 587, A19
R Core Team, 2016, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. Available at: https://www.R-project.org/
Rajpurohit A. S., Reylé C., Allard F., Homeier D., Schultheis M., Bessell M. S., Robin A. C., 2013, A&A, 556, A15
Rayner J. T., Cushing M. C., Vacca W. D., 2009, ApJS, 185, 289
Reid I. N., Gizis J. E., 2005, PASP, 117, 676
Riaz B., Gizis J. E., Samaddar D., 2008, ApJ, 672, 1153
Rojas-Ayala B., Covey K. R., Muirhead P. S., Lloyd J. P., 2012, ApJ, 748, 93
Scrucca L., 2013, J. Statistical Software, 53, 1
Ségransan D., Kervella P., Forveille T., Queloz D., 2003, A&A, 397, L5
Shields A. L., Ballard S., Johnson J. A., 2016, Phys. Rep., 663, 1
Svetnik V., Liaw A., Tong C., Culberson J. C., Sheridan R. P., Feuston B. P., 2003, J. Chemical Inf. Comput. Sci., 43, 1947
Torres G., 2013, Astron. Nachr., 334, 4

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Please note: Oxford University Press is 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.

APPENDIX A: IRTF FEATURES

In this Appendix we list the features selected by the Genetic Algorithm for the IRTF wavelength range and resolution.

Table A1. Recommended features and continuum bandpasses for predicting $T_{\rm eff}$ using BT-Settl with SNR = ∞, 10 and 50 and the IRTF wavelength range and resolution.

SNR = ∞				SNR = 10				SNR = 50			
λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$
9225.86	9283.94	9736.02	9793.96	8235.96	8294.04	12 681.62	12 768.68	8145.92	8204.03	12 636.48	12 723.57
11 106.48	11 193.56	13 497.81	13 613.95	8505.89	8563.93	13 378.12	13 494.13	8895.95	8953.95	11 331.57	11 418.65
13 438.08	13 554.08	12 006.54	12 093.56	9376.07	9433.92	12 951.62	13 038.62	8176.03	8234.13	10 611.36	10 698.46
9135.89	9193.91	10 002.04	9999.92	8145.92	8204.03	12 366.32	12 453.33	13 438.08	13 554.08	12 546.46	12 633.49
9555.93	9614.06	12 951.62	13 038.62	9195.86	9253.93	9135.89	9193.92	8235.96	8294.04	11 961.44	12 048.54
9466.08	9523.82	13 137.94	13 253.96	9585.95	9644.12	10 002.04	9999.92	9376.07	9433.92	10 002.04	9999.92
11 196.56	11 283.24	12 546.46	12 633.49	8385.99	8443.94	11 826.48	11 913.28	9406.09	9463.96	13 258.32	13 374.32
8566.08	8624.07	13 258.32	13 374.32	9135.89	9193.92	9225.86	9283.94	9346.13	9403.92	13 086.46	13 194.09
8266.11	8324.03	9856.06	9913.91	13 618.20	13 734.15	11 376.63	11 463.51	11 106.48	11 193.56	13 438.08	13 554.08
8235.96	8294.04	12 366.32	12 453.33	9105.87	9163.91	8865.98	8923.94	9255.86	9314.01	8865.98	8923.94

Table A2. Recommended features and continuum bandpasses for predicting $\log(g)$ using BT-Settl spectra of SNR = ∞, 10 and 50 in the IRTF wavelength range and resolution.

SNR =∞				SNR = 10				SNR = 50			
λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$
10 245.88	10 304.02	11 241.29	11 328.54	8176.03	8234.13	9165.87	9223.91	11 151.63	11 238.46	13 086.46	13 194.09
8415.91	8473.96	11 511.51	11 598.51	10 485.99	10 563.41	10 002.04	9999.92	8385.99	8443.94	13 618.20	13 734.14
12 906.56	12 993.61	13 041.48	13 133.82	8656.09	8714.047	10 926.46	11 013.60	8176.03	8234.13	11 241.29	11 328.54
8716.00	8773.99	10 425.90	10 484.13	9525.89	9584.059	10 002.04	9999.92	8536.03	8594.06	13 041.48	13 133.82
8805.93	8863.97	12 816.72	12 903.73	8205.98	8263.967	13 041.48	13 133.82	12 771.70	12 858.73	10 306.03	10 363.88
10 126.02	10 183.93	13 086.46	13 194.09	10 275.97	10 333.96	11 376.63	11 463.51	13 378.12	13 494.13	10 002.04	9999.92
8176.03	8234.13	10 971.57	11 058.46	10 306.03	10 363.88	11 151.63	11 238.46	8626.02	8683.99	10 926.46	11 013.60
8626.02	8683.99	10 746.43	10 833.57	9165.87	9223.91	8385.99	8443.94	9826.05	9883.91	10 006.07	10 064.01
8536.03	8594.06	10 215.95	10 274.10	9645.82	9704.16	13 137.94	13 253.96	10 521.56	10 608.46	11 736.71	11 823.49
12 951.62	13 038.62	11 196.56	11 283.24	8326.00	8383.94	12 726.69	12 813.71	8205.98	8263.96	9796.09	9853.94

Table A3. Feature and continuum bandpasses selected for predicting metallicity using noisy BT-Settl spectra with signal-to-noise ratios equal to ∞, 10 and 50 in the IRTF wavelength range and resolution.

SNR = ∞				SNR = 10				SNR = 50			
λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$
12 096.68	12 183.66	12 051.50	12 096.68	8235.96	8294.04	11 331.57	11 418.65	9255.86	9314.01	13 197.94	13 313.92
9525.89	9584.05	12 321.33	12 408.32	9376.07	9433.92	10 566.33	10 653.62	8385.99	8443.94	9376.07	9433.92
8205.98	8263.96	10 126.02	10 183.93	10 306.03	10 363.88	9942.14	9999.92	8716.00	8773.99	9585.95	9644.12
8566.08	8624.07	12 276.52	12 363.34	11 286.42	11 373.45	11 241.29	11 286.42	8235.96	8294.04	13 086.46	13 194.09
11 196.56	11 283.24	11 151.63	11 196.56	9676.00	9734.02	13 086.46	13 194.09	9676.00	9734.02	10 791.44	10 878.40
11 151.639	11 238.46	11 466.35	11 553.33	8775.95	8833.94	8415.91	8473.96	8415.91	8473.96	12 411.34	12 498.41
9555.93	9614.06	8205.98	8263.96	12 411.34	12 498.41	10 245.88	10 304.02	8446.03	8503.94	9406.09	9463.96
11 016.62	11 103.37	10 791.44	10 878.40	8476.01	8534.03	12 276.52	12 363.34	8205.98	8263.96	8955.88	9013.95
9766.16	9823.94	12 681.62	12 768.68	12 636.48	12 723.57	12 051.50	12 138.72	8985.93	9043.98	12 186.62	12 273.48
9942.14	9999.92	9555.93	9614.06	8415.91	8473.96	13 618.20	13 734.14	9015.98	9073.98	11 241.29	11 328.54

APPENDIX B: DWARF ARCHIVES FEATURES

In this Appendix we list the features selected by the Genetic Algorithm for the Dwarf Archives wavelength range and resolution.

Table B1. Spectral features and continuum bandpasses selected by the GA for predicting $T_{\rm eff}$ using BT-Settl spectra with SNR $= \infty$, 10 and 50 in the Dwarf Archives wavelength range and resolution.

SNR $= \infty$				SNR $= 10$				SNR $= 50$			
λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$
7062	7094.4	7314	7346.4	7692	7724.4	6936	6968.4	7062	7094.4	7296	7328.4
7116	7148.4	7782	7814.4	6990	7022.4	7998	8030.4	7026	7058.4	7044	7076.4
7134	7166.4	7872	7904.4	6900	6932.4	7548	7580.4	7080	7112.4	7926	7958.4
6900	6932.4	7764	7796.4	7854	7886.4	7710	7742.4	6900	6932.4	7548	7580.4
7170	7202.4	7890	7922.4	7116	7148.4	7908	7940.4	7134	7166.4	7836	7868.4
7080	7112.4	7926	7958.4	7278	7310.4	7926	7958.4	7296	7328.4	7962	7994.4
7188	7220.4	7548	7580.4	7152	7184.4	7746	7778.4	6936	6968.4	7728	7760.4
7800	7832.4	7962	7994.4	7134	7166.4	7764	7796.4	6972	7004.4	6900	6932.4
6990	7022.4	7008	7040.4	6918	6950.4	6900	6932.4	6990	7022.4	7944	7976.4
7026	7058.4	6990	7022.4	7224	7256.4	7962	7994.4	6918	6950.4	7782	7814.4

Table B2. Spectral features and continuum bandpasses selected by the GA for predicting $\log(g)$ using BT-Settl spectra of SNR $= \infty$, 10 and 50 in the Dwarf Archives wavelength range and resolution.

SNR $= \infty$				SNR $= 10$				SNR $= 50$			
λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$
7134	7166.4	7044	7076.4	6990	7022.4	6918	6950.4	6918	6950.4	6936	6968.4
6954	6986.4	7152	7184.4	6900	6932.4	7278	7310.4	6936	6968.4	7836	7868.4
7512	7544.4	7890	7922.4	7062	7094.4	7242	7274.4	7656	7688.4	7890	7922.4
7062	7094.4	7224	7256.4	7692	7724.4	7008	7040.4	6900	6932.4	7872	7904.4
6936	6968.4	7854	7886.4	7656	7688.4	7998	8030.4	7008	7040.4	7044	7076.4
6900	6932.4	7746	7778.4	6936	6968.4	7836	7868.4	7512	7544.4	7656	7688.4
6918	6950.4	7800	7832.4	7206	7238.4	7062	7094.4	7440	7472.4	7332	7364.4
7008	7040.4	7134	7166.4	7512	7544.4	7926	7958.4	7800	7832.4	7692	7724.4
7872	7904.4	7008	7040.4	7764	7796.4	7710	7742.4	7404	7436.4	7548	7580.4
7962	7994.4	7980	8012.4	7404	7436.4	7548	7580.4	7080	7112.4	7152	7184.4

Table B3. Spectral features and continuum bandpasses selected by the GA for predicting metallicities using BT-Settl spectra of SNR $= \infty$, 10 and 50 in the Dwarf Archives wavelength range and resolution.

SNR $= \infty$				SNR $= 10$				SNR $= 50$			
λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$	λ_1	λ_2	$\lambda_{\rm cont;\,1}$	$\lambda_{\rm cont;\,2}$
7188	7220.4	7854	7886.4	7692	7724.4	7026	7058.4	7098	7130.4	7926	7958.4
7080	7112.4	7926	7958.4	6900	6932.4	7008	7040.4	7188	7220.4	7962	7994.4
7116	7148.4	7098	7130.4	7350	7382.4	7908	7940.4	7368	7400.4	7980	8012.4
7422	7454.4	7836	7868.4	6918	6950.4	6900	6932.4	7116	7148.4	7872	7904.4
7350	7382.4	7998	8030.4	7098	7130.4	7314	7346.4	7062	7094.4	7206	7238.4
7224	7256.4	7818	7850.4	7440	7472.4	7872	7904.4	7584	7616.4	7170	7202.4
7710	7742.4	7062	7094.4	7134	7166.4	7962	7994.4	6936	6968.4	6918	6950.4
7476	7508.4	7944	7976.4	7368	7400.4	7926	7958.4	7692	7724.4	7890	7922.4
7134	7166.4	7584	7616.4	7080	7112.4	7044	7076.4	7134	7166.4	7548	7580.4
7836	7868.4	7278	7310.4	7044	7076.4	7980	8012.4	7494	7526.4	7998	8030.4

APPENDIX C: IRTF RMSE REGRESSION MODELS

In this appendix we include evaluation measures for the regression models trained with spectra with the IRTF wavelength range and resolution and three SNR levels.

Table C1. Cross-validation RMSE and RMDSE for the various regression models that predict $T_{\rm eff}$ (K) in the IRTF wavelength range and resolution.

	SNR =10		SNR =50		SNR =∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
χ^2	232	**100**	235	120	232	**100**
PPR-ICA	242	128	242	99	280	162
GA-RR	260	115	270	128	333	170
GA-RF	308	183	248	136	**167**	135
GA-GBM	287	160	248	149	233	113
GA-SVR	**221**	122	281	151	299	160
GA-NNET	283	192	264	114	326	212
GA-KNN	238	120	**232**	137	219	**100**
GA-MARS	253	113	254	**95**	226	133
GA-KPLS	275	120	300	119	387	218

Table C2. Average bias in the $T_{\rm eff}$ (K) estimates computed (IRTF wavelength range and resolution), with respect to the reference values in table 3 of Cesetti et al. (2013).

	SNR = 10	SNR = 50	SNR = ∞
χ^2	−77	−87	−85
PPR-ICA	−104	−55	−130
GA-RR	−102	−39	170
GA-RF	−173	−127	−5
GA-GBM	−141	−109	32
GA-SVR	−58	−3	92
GA-NNET	−147	−36	39
GA-KNN	−76	−110	−67
GA-MARS	−57	−88	98
GA-KPLS	−120	−4	214

Table C3. Regression model performance based on the features proposed by Cesetti et al. (2013).

	SNR = 10		SNR = 50		SNR = ∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
CS-RR	252	140	532	322	606	537
CS-RF	234	180	**264**	218	**321**	265
CS-GBM	**232**	195	268	254	325	246
CS-SVR	268	227	293	257	432	364
CS-NNET	357	255	357	**204**	552	435
CS-KNN	249	172	293	256	327	**230**
CS-MARS	289	**98**	676	245	1570	590
CS-KPLS	351	162	856	456	1086	535

Table C4. RMSE and RMDSE for the various log (g) regression models [dex] in the IRTF wavelength range and resolution.

	SNR = 10		SNR = 50		SNR = ∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
χ^2	0.82	0.45	0.93	0.61	3.5	3.48
PPR-ICA	0.54	0.48	**0.3**	**0.17**	0.72	0.57
GA-RR	0.74	0.57	0.50	0.47	0.57	0.41
GA-RF	0.64	**0.38**	0.77	0.72	0.53	0.39
GA-GBM	**0.48**	0.45	0.61	0.47	0.49	0.41
GA-SVR	0.66	0.40	0.63	0.58	**0.46**	**0.21**
GA-NNET	0.78	0.61	0.47	0.44	1.2	0.97
GA-MARS	0.84	0.57	0.54	0.37	0.99	0.76
GA-KNN	1.23	0.83	1.39	1.44	1.60	1.32
GA-KPLS	0.99	0.99	0.51	0.49	0.96	0.77

Table C5. RMSE and RMDSE for the various regression models (IRTF wavelength range and resolution) predicting metallicity [dex].

	SNR = 10		SNR = 50		SNR = ∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
χ^2	0.76	0.22	0.36	0.18	0.36	0.18
PPR-ICA	0.24	**0.13**	0.31	0.22	0.43	0.27
GA-RR	0.31	0.17	0.30	0.24	0.78	0.23
GA-RF	0.33	0.25	0.73	0.41	0.61	0.36
GA-GBM	0.27	0.19	0.70	0.52	0.63	0.35
GA-SVR	0.33	0.22	0.45	0.32	0.92	0.89
GA-NNET	0.37	0.30	0.33	0.37	0.95	0.81
GA-MARS	0.36	0.16	0.49	0.41	0.83	0.85
GA-KNN	0.69	0.55	0.23	**0.15**	0.21	**0.15**
GA-KPLS	0.49	0.50	0.52	0.48	1.06	1.01

APPENDIX D: DWARF ARCHIVES RMSE REGRESSION MODELS

In this appendix we include evaluation measures for the regression models trained with spectra with the Dwarf Archives wavelength range and resolution.

Table D1. RMSE and RMDSE for the various regression models that predict $T_{\rm eff}$ (K) in the Dwarf Archives wavelength range and resolution.

	SNR = 10		SNR = 50		SNR = ∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
χ^2	**147**	79	**121**	**56**	**126**	**57**
PPR-ICA	188	126	164	95	191	130
GA-RR	189	102	287	103	378	239
GA-RF	160	97	196	103	145	94
GA-GBM	175	105	225	99	185	94
GA-SVR	203	112	285	106	368	154
GA-NNET	221	84	313	111	395	202
GA-MARS	222	76	361	103	374	157
GA-KNN	183	119	193	109	224	110
GA-KPLS	227	**72**	331	123	409	208

Table D2. Performances of regression models trained on the features selected by Cesetti et al. (2013) and applied to BT-Settl spectra in the Dwarf Archives wavelength range and resolution.

	SNR = 10		SNR = 50		SNR = ∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
CS-RR	211	128	400	239	828	774
CS-RF	203	140	243	**121**	**306**	**172**
CS-GBM	**188**	**120**	**161**	138	337	222
CS-SVR	197	135	379	194	840	688
CS-NNET	207	135	514	296	719	489
CS-MARS	252	124	789	186	3464	784
CS-KNN	235	158	246	137	314	175
CS-KPLS	250	201	741	361	2247	1424

Table D3. RMSE and RMDSE for the various regression models predicting log (g) [dex] in the Dwarf Archives wavelength range and resolution.

	SNR = 10		SNR = 50		SNR = ∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
χ^2	2.2	1.6	2.2	1.4	2.2	1.6
PPR-ICA	2.1	1.8	1.8	1.4	4.3	4.2
GA-RR	2.0	1.8	2.1	1.8	3.7	3.2
GA-RF	**1.3**	**1.0**	**1.6**	**1.1**	**1.4**	**0.9**
GA-GBM	1.6	1.1	1.7	1.4	1.7	1.2
GA-SVR	2.0	1.8	2.1	1.9	2.3	1.6
GA-NNET	2.0	1.8	2.2	1.9	3.2	2.8
GA-MARS	1.8	1.5	2.0	1.7	2.0	1.5
GA-KNN	2.0	1.5	2.2	1.7	1.7	1.2
GA-KPLS	1.8	1.4	2.0	1.7	2.7	2.3

Table D4. RMSE and RMDSE for the various regression models predicting metallicity [dex] in the Dwarf Archives wavelength range and resolution.

	SNR = 10		SNR = 50		SNR = ∞	
Regression models	RMSE	RMDSE	RMSE	RMDSE	RMSE	RMDSE
χ^2	0.55	0.27	0.51	0.29	0.43	0.29
PPR-ICA	0.48	0.27	0.70	0.39	0.85	0.71
GA-RR	0.47	0.29	0.50	0.36	1.18	1.18
GA-RF	0.55	0.38	0.71	0.61	0.23	0.16
GA-GBM	0.64	0.43	0.87	0.84	0.31	0.23
GA-SVR	0.46	0.26	0.57	0.44	3.38	2.33
GA-NNET	0.52	0.45	0.66	0.54	2.03	1.88
GA-MARS	0.71	0.47	0.80	0.69	1.15	0.68
GA-KNN	0.37	0.28	0.99	0.78	0.56	0.32
GA-KPLS	0.67	0.61	0.63	0.55	1.17	1.02

APPENDIX E: RESULTS WITH FEATURES FROM CESETTI ET AL. (2013)

In this appendix we include plots with the predictions of the best performing regression models with the features proposed by Cesseti et al (2013).

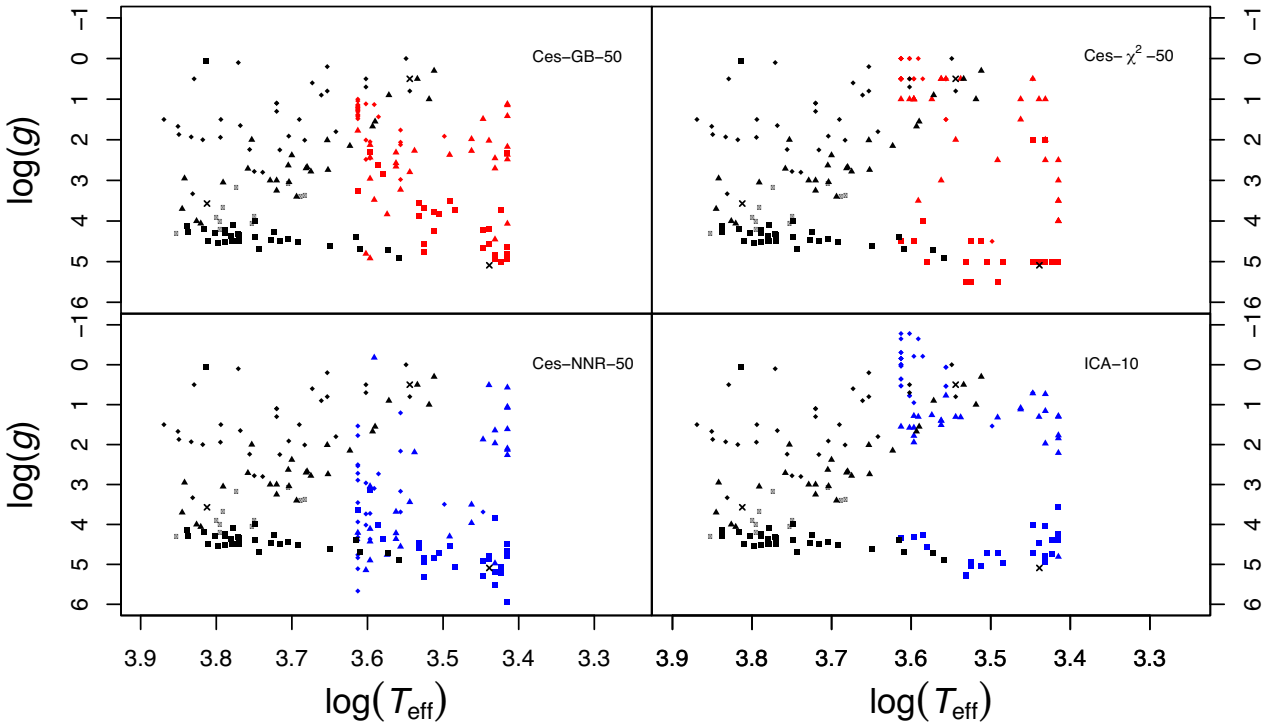

Figure E1. log ($T_{\rm eff}$)–log (g) diagrams produced by the CES-KNN (SNR = ∞) effective temperatures, and gravities derived for the IRTF collection of spectra with the CES-GBM (SNR = 50), CES-χ^2 (SNR = 50), CES-NNET (SNR = 50) and ICA-10 models (clockwise, starting from the top left).

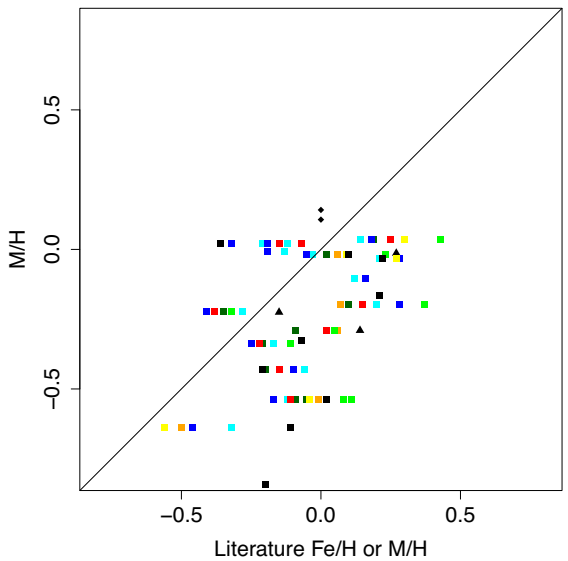

Figure E2. Comparison of the CES-RF (SNR = 50) regression model predictions for the IRTF collection of spectra with the estimates of the metallicity in the literature. The symbols and colours are the same as in Fig. 5.

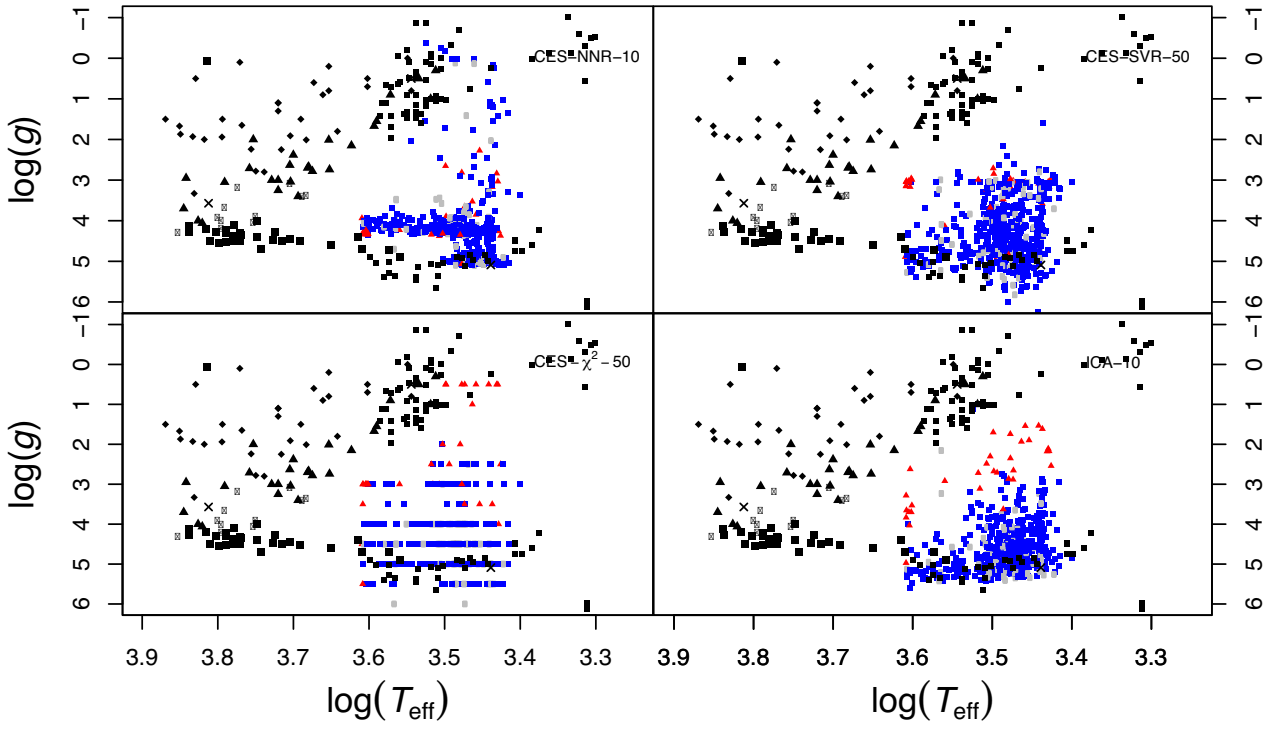

Figure E3. log ($T_{\rm eff}$)–log (g) diagrams produced by the CES-KNN (SNR = ∞) effective temperatures, and gravities derived for the Dwarf Archives collection of spectra with the CES-NNET (SNR = 10), CES-SVR (SNR = 50), CES-χ^2 (SNR = 50) and ICA-10 models (clockwise, starting from the top left).

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