Astronomy & Astrophysics manuscript no. ordieres ©ESO 2015
December 16, 2015

Physical parameter estimates of M-type stars: a machine learning perspective.

L. M. Sarro[1], J. Ordieres-Mere[2], A. Bello-Garcia[3], A. Gonzalez-Marcos[4], and M.B. Prendes-Gero[3]

[1] [1] Universidad Nacional de Educación a Distancia,
Department of Artificial Intelligence. e-mail: lsb@uned.es

[2] [2] Universidad Politécnica de Madrid (UPM), PMQ Research Group,
José Gutiérrez Abascal 2, 28006 Madrid, Spain. e-mail: j.ordieres@upm.es

[3] [3] Universidad de Oviedo, Construction and Manufacturing Engineering Department,
Campus de Viesques s/n, Gijón, Asturias, Spain. e-mail: {abello,mbprendes}@uniovi.es

[4] [4] Universidad de la Rioja, P2ML Research Group,
Luis de Ulloa 20, 26004 Logroño, La Rioja, Spain. e-mail: ana.gonzalez@unirioja.es

Received December 16, 2015; accepted

ABSTRACT

Key words. class M stars – dynamic feature selection – physical parameter identification – Temperature, gravity and metalicity Modelling – Learning from BT-Settl spectra library

Use \titlerunning to supply a shorter title and/or \authorrunning to suply a shorter list of author.

1. Introduction

2. Methodology.

The objective addressed in this Section is to develop a procedure to identify spectral bands that yield good temperature, gravity and metallicity diagnostics. Given the lack of a calibration set of observed spectra with homogeneous coverage of the space of physical parameters, we turn to synthetic libraries of spectra. The atomic or molecular line/band parameters could in principle indicate the spectral features that are more sensitive to changes in the physical parameters. The suitability of spectral features as diagnostics ofthe 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, that may overlap depending on the spectral resolution. Furthermore, good

spectral diagnostics at a given signal-to-noise ratio (SNR) may show a severy degraded predictive power in the low SNR regime. In the following we adopt the BT-Settl library of synthetic spectra (Allard et al. (2013)) as the framework where spectral diagnostics will be searched for. These synthetic spectra were pre-processed in several steps as described below.

2.1. Spectral preprocessing

First, and in order to define good temperature diagnostics, spectra between 2000 and 4200K 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 were 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 (IPAC or IRTF, see below). This required the definition of a common wavelength range present in all available observed spectra, and the 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.

In order to increase the density of examples in parameter space, we introduced interpolated spectra in the BT-Settl grid. Interpolation was obtained as a linear combination of spectra in the grid, weighted by the inverse square of the euclidean distance. **Aqui, la distancia euclidea deberia calcularse en parametros normalizados, porque si no la temperatura domina la distancia. Fue asi?** We compared a set of interpolated spectra with those produced using the PHOENIX code (Fuhrmeister et al. (2005)) to be sure that interpolation was a valid solution to infer new synthetic spectra. **Yo aqui daría el RMSE de reconstruccion, mejor que la figura comp-gen-inter**

A first interpolation stage allowed us to define a finer mesh step of 0.25 dex for both, $\log(g)$ and metallicity and 50K in temperature, yielding a total 1329 spectra. Then, a second interpolation stage refined the grid down to 25 K in temperature and 0.125 dex in $\log(g)$, keeping the metallicity step at 0.25 dex and producing a dataset with 25912 spectra. In spite of these, and in order to keep their knowledge closer to the original BT-Settl source, most of the analyses have been performed with the original 535 spectra dataset.**Habría que delimitar exactamente donde se han utilizado 535 y dónde 25912. Si la mayoria del analisis se ha realizado sobre 535, no se si tiene sentido incluir la parte de interpolacion.**

In order to avoid selecting spectral features that are good predictors only in the unrealistic SNR=∞ regime, the search for optimal diagnostics of the atmopheric paramters of M stars was carried out for three SNR values (10, 50 and ∞) by degrading the synthetic spectra with Gaussian noise of zero mean. **Quizás deberíamos citar el trabajo de Ana como in preparation**

2.2. Feature definition and selection

As mentioned in Sect. 1, it is well known the difficulty in defining good spectral diagnostics for M stars in the infrared.

The work in Cesetti et al. (2013) defined wavelength regions in the I and K bands optimal for the diagnostic of 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 in this work 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_{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) and thus, it was estimated first, and then used as in the regression models for the gravity and metalicity.

Here, a feature F is defined as

$$F = \int_{\lambda_1}^{\lambda_2} (1 - \frac{f(\lambda)}{F_{cont}} \cdot \mathrm{d}\lambda) \tag{1}$$

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

Aquí he omitido los detalles sobre las restricciones a las bandas porque no lo entiendo bien. ¿Podrías explicarlo con palabras en lugar de items?

Another type of features defined as

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

was considered, where $\lambda_1, \lambda_2, \lambda_3$, and $lambda_4$ delimit two spectral bands such that the ratio of the integrated fluxes in the two bands is hoped to be a good predictor (alone or in combination with other features) of the star atmospheric physical parameters. The results obtained with this alternative feature definition did not differ significantly on average from the ones observed with the one adopted in Eq. 1, and including them here would result in an excessively lengthy paper. In view of the equivalent global performances, we prefered the former because it allows direct comparison with the features proposed by Cesetti et al. (2013).

We used Genetic Algorithms 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 (R Core Team 2013) ??? package.

These methods have demonstrated to perform well (**falta cita**) **¿en qué circunstancias?**. However, in some cases they can be ineffective regardless of the classification method used. An obvious conceptual limitation of univariate approaches is also the lack of consideration that features works in the contexts of interconnected pathways and therefore it is their behavior as a group that may be predictive of the phenotypic variables. Multivariate selection methods are tested in combination to identify interactions between features. However, the extremely large number of models that can be constructed from different combination of thousands of features cannot be extensively evaluated using standard computational resources.

Joaquín: please confirm that the previous paragraph is correct. I remember that we discussed this in your office. I recall myself advocating for a GA that considered sets of features rather than individual features. I do not remember how this all ended up.

For the sake of simplicity let us define Genetic Algorithms (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) sets of variables (chromosomes) from an initial random population. 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 with the accuracy with which a given chromosome (set of spectral features F_i) predicts the physical parameters. The concept of using in-silico evolution for the solution of optimization problems was introduced by Holland (1975). Although its application is now reasonably widespread (Goldberg et al. 1989, see e.g.)), they became very popular only when sufficiently powerful computers became available. **Aquí hay que citar trabajos en astrofísica que utilicen GA y, en particular, un artículo de Charbonneau http://adsabs.harvard.edu/abs/1995ApJS..101..309C en 1995 que fue como la presentación en sociedad.**

The implementation of the GA comprises the following steps:

Stage 1: Definition of the population of potential features (chromosomes).

Stage 2: Each chromosome in the population is evaluated by its ability to predict the physical paramters of each star in the dataset (fitness function).

Stage 3: Chromosome selection, when a chromosome has a score higher than a predefined value.

Stage 4: The population of chromosomes is replicated. Chromosomes with higher fitness scores will generate more numerous offspring.

Stage 5: The genetic information contained in the replicated parent chromosomes is combined through genetic crossover. Two randomly selected parent chromosomes are used to create two new chromosomes.

Stage 6: Mutations are then introduced in the chromosome randomly. These mutations produce new genes used in chromosomes. Steps 5 and 6 are applied over the chromosomes established at Step 4.

Stage 7: This process is repeated from Stage 2 until a target accuracy is achieved or the maximum number of iterations is attained.

Es muy importante definir la codificación del cromosoma. ¿Es la codificación de un único feature? ¿un subconjunto de features? ¿de qué tamaño?

There are different statistics that can be used to identify features that are differentially expressed between two or more groups of samples **hay que explicar a qué nos referimos con differential expression, samples y groups of samples aquí** and then uses the most differentially expressed ones to construct a statistical model.

The population size was set to 1000 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 was fixed to 0.15 **No hemos mencionado elitismo; hay que mencionarlo y definirlo antes**. Feature fitness was defined in terms of the Akaike Information Criterion (AIC) for linearity between the potential feature against the physical parameter. **No entiendo esta última frase. La linealidad... ¿se refiere al modelo de regresión lineal que utilizamos para medir el fitness de una feature? Creo que hay que añadir un párrafo en el que expliquemos con detalle el regresor utilizado para medir la fitness. Y sobre todo, aclarar si el cromosoma codifica sólo una feature o un conjunto de features.** The most frequent and efficient features were selected as candidates to predictive variables of the physical parameters in regression models. We used a binary codification of the chromosomes and a parallel implementation of the GA in a farm of fifteen computers per physical parameter. **Here a bit more detail is needed: what processors, number of cores, etc. Just one additional sentence.**

The GA procedure provides us with a large collection of chromosomes. Although these are all potential solutions of the problem, it is not immediately clear which one should be selected for the final regression model. This single regression model should, to some extent, be representative of the population. The simpler strategy would be to use the frequency of the chromosome in the population as criterion for inclusion in a forward selection strategy. However we prefered to select the features based on their highest fitness. **How many? Do we select the top 10 fittest chromosomes? Why 10?**

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 to predict them based in these features. The GA generates a large set of proposals **here we need to explain how we go from the output of the GA to a list of 10 features. They are ordered by fitness and number of replicates in the pool, I believe. We keep the top fittest features with many replications, but can we describe this more quantitatively?**

In order to assess the performance of the regression models, we compare their predictions with i) values of the physical parameters from the literature (when available); ii) the predictions from the popular *minimum*χ^2 distance to spectra in the BT-Settl library; iii) parameter predictions based on a projection pursuit regression model **Is this correct, Joaquín? Somewhere in your first version of the paper it was stated that the ICA components were fed to an SVM with C=10 and**

epsilon=0.001 for temperature, and different values for logg and metallicity trained with projections of the BT-Settl spectra onto the set of vectors resulting from an Independent Component Analysis (ICA); and finally, iv) predictions from a regression model trained with the features proposed by Cesetti et al. (2013) (only for the IRTF spectra) **Joaquín, aquí necesitamos explicar qué tipo de modelo entrenamos con las features de cesetti.**.

2.3. Models considered.

For the models to be built, the same strategy was used for all the three physical parameters ($T_{eff}, log(g), met$) and it was to use non linear methods for modellization. As a classical regression problem several linear and non-linear modelling techniques with specific research for adequate parameters per method when required, were considered: **Joaquín, no entiendo este párrafo. Pareces decir primero que utilizas modelos no lineales, para luego indicar que utilizas varios modelos lineales y no lineales. Los GAMs son lineales ¿no?**

- Generalized Additive Models *(GAM)*.
- Bagging with Multiadaptative Spline Regression Models *(MARS)*.
- Random Forest Regression Models *(RF)*.
- Gradient Boosting with Regression Trees *(BOOSTING)*.
- Generalized Boosted Regression Models *(GBM)*.
- Support Vector Regression with Gaussian Kernel *(SVM)*.
- MLP Neural Networks *(NNET)*.
- Kernel Partial Least Squares Regression *(KPLS)*.

Including here a sufficient description of each and every regression model that we trained would render the manuscript excessively lengthy. 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 normalised spectrum, the ICA lower-dimensional representation of the full spectrum, or the spectral features selected by Cesetti et al. (2013) or by the GA. The model parameters are infered (using algorithms that differ from one regression model to the other) from a set of examples. 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.

Aquí haría falta describir muy mínimamente cada uno de los modelos y dar una referencia que los describa en detalle. Luego, explicar cómo se determina el valor óptimo de los parámetros de cada modelo. Me pareció entender que caret lo hace automáticamente, pero en cualquier caso habría que escribirlo explícitamente y citar caret (si este es el caso).

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 represents 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 infered 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 **?**, **?**, and **?**, where the atmospheric parameters are derived using interferometric measurements of stellar radii. Again, this only amounts to a very small number of examples and a very sparse sampling of the parameters space.

We believe that 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 same atmospheric models. But until these training sets with sufficient and homogeneous sampling of the parameter space are available, we turn to the use of synthetic libraries.

3. Physical parameters of the IRTF collection of spectra.

3.1. Spectral bands selected

During the preprocessing stage (described in Sect. 2) the spectral resolution of the BT-Settl library was degraded to the IRTF resolution (R 2000) by convolving with a Gaussian. Then, the spectra were trimmed to produce valid segments between 8145.92 and 24106.85Å, 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.

3.2. Spectral features for estimation of effective temperatures.

The application of the GAs to the selection of features for the prediction of effective temperature from noiseless spectra with the IRTF wavelength range and resolution results in the features included in Table 12. Features are ordered by the fitness value (the AIC) and we only consider features that are present in at least 5 sets.

TBD by Luis: interpret the features.

When noise is added to the BT-Settl spectra, we obtain

Tables 12 and **??** show a very wide variety of features with very few repetitions. Only spectral features 4, 5, 6, and 9 in the SNR=50 experiment are found too in the SNR=∞ and SNR=10

λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
9225.86	9283.94	9736.02	9793.96
11106.48	11193.56	13497.81	13613.95
13438.08	13554.08	12006.54	12093.56
9135.89	9193.91	10002.04	9999.92
9555.93	9614.06	12951.62	13038.62
9466.08	9523.82	13137.94	13253.96
11196.56	11283.24	12546.46	12633.49
8566.08	8624.07	13258.32	13374.32
8266.11	8324.03	9856.06	9913.91
8235.96	8294.04	12366.32	12453.33

Table 1: Features selected by the GA for predicting T_{eff} using BT_Settl noiseless synthetic spectra.

SNR = 10				SNR=50			
λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$	λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
8235.96	8294.04	12681.62	12768.68	8145.92	8204.03	12636.48	12723.57
8505.89	8563.93	13378.12	13494.13	8895.95	8953.95	11331.57	11418.65
9376.07	9433.92	12951.62	13038.62	8176.03	8234.13	10611.36	10698.46
8145.92	8204.03	12366.32	12453.33	13438.08	13554.08	12546.46	12633.49
9195.86	9253.93	9135.89	9193.92	8235.96	8294.04	11961.44	12048.54
9585.95	9644.12	10002.04	9999.92	9376.07	9433.92	10002.04	9999.92
8385.99	8443.94	11826.48	11913.28	9406.09	9463.96	13258.32	13374.32
9135.89	9193.92	9225.86	9283.94	9346.13	9403.92	13086.46	13194.09
13618.20	13734.15	11376.63	11463.51	11106.48	11193.56	13438.08	13554.08
9105.87	9163.91	8865.98	8923.94	9255.86	9314.01	8865.98	8923.94

Table 2: Recommended features and Continuum bandpass for predicting T_{eff} by using BT_Settl with SNR= 10 and 50.

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.

A closer look at features 4, 5, 6, and 9

As a reference, Table 3 lists the features found by **?** using sensitivity maps.

For gravity (in the form of $\log(g)$) estimation, the GA search procedure produces the features presented in Tables 14 and 17 for the pure synthetic signal and signal-to-noise ratios of 10 and 50, respectively.

Finally, the best features found by the GA for metallicity estimation are listed in Table 16 for the noiseless BT-Settl spectra, and in Table **??** for signal-to-noise ratios equal to 10 and 50.

When signal-to-noise ratios equal to 10 and 50 are considered, the GA finds the selected features listed in Table **??**.

3.3. Regression models

In the following, we will summarise the results obtained for the IRTF data set. We deal with the different physical paramenters in separate Sections. We start by reporting the cross validation Root Mean Square Errors (RMSE) for the five-fold cross-validation strategy, and subsequently discuss the accuracy of the predictions with respect to literature values where available.

Index	Element	Signal_from	Signal_To	Cont1_From	Cont1_To	Cont2_From	Cont2_To
Pa1	H I	8461	8474	8474	8484	8563	8577
Ca1	Ca II	8484	8513	8474	8484	8563	8577
Ca2	Ca II	8522	8562	8474	8484	8563	8577
Pa2	H I	8577	8619	8563	8577	8619	8642
Ca3	Ca II	8642	8682	8619	8642	8700	8725
Pa3	H I	8730	8772	8700	8725	8776	8792
Mg	Mg I	8802	8811	8776	8792	8815	8850
Pa4	H I	8850	8890	8815	8850	8890	8900
Pa5	H I	9000	9030	8983	8998	9040	9050
FeClTi	Fe I, Cl I, Ti I	9080	9100	9040	9050	9125	9135
Pa6	H I	9217	9255	9152	9165	9265	9275
Fe1	Fe I	1.9297	1.9327	1.9220	1.9260	2.0030	2.0100
Brδ	H I (n=4)	1.9425	1.9470	1.9220	1.9260	2.0030	2.0100
Ca1	Ca I	1.9500	1.9526	1.9220	1.9260	2.0030	2.0100
Fe23	Fe I	1.9583	1.9656	1.9220	1.9260	2.0030	2.0100
Si	Si I	1.9708	1.9748	1.9220	1.9260	2.0030	2.0100
Ca2	Ca I	1.9769	1.9795	1.9220	1.9260	2.0030	2.0100
Ca3	Ca I	1.9847	1.9881	1.9220	1.9260	2.0030	2.0100
Ca4	Ca I	1.9917	1.9943	1.9220	1.9260	2.0030	2.0100
Mg1	Mg I	2.1040	2.1110	2.1000	2.1040	2.1110	2.1150
Brγ	H I (n=4)	2.1639	2.1686	2.0907	2.0951	2.2873	2.2900
Na_d	Na I	2.2000	2.2140	2.1934	2.1996	2.2150	2.2190
FeA	Fe I	2.2250	2.2299	2.2133	2.2176	2.2437	2.2479
FeB	Fe I	2.2368	2.2414	2.2133	2.2176	2.2437	2.2479
Ca_d	Ca I	2.2594	2.2700	2.2516	2.2590	2.2716	2.2888
Mg2	Mg I	2.2795	2.2845	2.2700	2.2720	2.2850	2.2874
^{12}CO	^{12}CO(2,0)	2.2910	2.3070	2.2516	2.2590	2.2716	2.2888

Table 3: Recommended features and continuum bandpasses recommended in **?** as relevant for the estimation of the effective temperature in bands I and K.

λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
10245.88	10304.02	11241.29	11328.54
8415.91	8473.96	11511.51	11598.51
12906.56	12993.61	13041.48	13133.82
8716.00	8773.99	10425.90	10484.13
8805.93	8863.97	12816.72	12903.73
10126.02	10183.93	13086.46	13194.09
8176.03	8234.13	10971.57	11058.46
8626.02	8683.99	10746.43	10833.57
8536.03	8594.06	10215.95	10274.10
12951.62	13038.62	11196.56	11283.24

Table 4: Recommended features and continuum bandpasses for predicting log(g) obtained using noiseless BT_Settl spectra.

3.4. Effective temperature models

Table 8 summarises the RMSE 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 obtained for several noise levels of the training sets. We use the following notation: RMSE represents the RMSE obtained for a model trained and tested on BT Settl spectra. SNR=∞ corresponds to noiseless

SNR = 10				SNR=50			
λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$	λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
8176.03	8234.13	9165.87	9223.91	11151.63	11238.46	13086.46	13194.09
10485.99	10563.41	10002.04	9999.92	8385.99	8443.94	13618.20	13734.14
8656.09	8714.047	10926.46	11013.60	8176.03	8234.13	11241.29	11328.54
9525.89	9584.059	10002.04	9999.92	8536.03	8594.06	13041.48	13133.82
8205.98	8263.967	13041.48	13133.82	12771.70	12858.73	10306.03	10363.88
10275.97	10333.96	11376.63	11463.51	13378.12	13494.13	10002.04	9999.92
10306.03	10363.88	11151.63	11238.46	8626.02	8683.99	10926.46	11013.60
9165.87	9223.91	8385.99	8443.94	9826.05	9883.91	10006.07	10064.01
9645.82	9704.16	13137.94	13253.96	10521.56	10608.46	11736.71	11823.49
8326.00	8383.94	12726.69	12813.71	8205.98	8263.96	9796.09	9853.94

Table 5: Recommended features and continuum bandpasses for predicting $log(g)$ obtained using BT_Settl with SNR= 10 and 50.

λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
12096.68	12183.66	12051.50	12096.68
9525.89	9584.05	12321.33	12408.32
8205.98	8263.96	10126.02	10183.93
8566.08	8624.07	12276.52	12363.34
11196.56	11283.24	11151.63	11196.56
11151.639	11238.46	11466.35	11553.33
9555.93	9614.06	8205.98	8263.96
11016.62	11103.37	10791.44	10878.40
9766.16	9823.94	12681.62	12768.68
9942.14	9999.92	9555.93	9614.06

Table 6: Feature and Continuum bandpasses selected for predicting *Metallicity* using noiseless BT_Settl spectra.

SNR = 10				SNR=50			
λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$	λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
8235.96	8294.04	11331.57	11418.65	9255.86	9314.01	13197.94	13313.92
9376.07	9433.92	10566.33	10653.62	8385.99	8443.94	9376.07	9433.92
10306.03	10363.88	9942.14	9999.92	8716.00	8773.99	9585.95	9644.12
11286.42	11373.45	11241.29	11286.42	8235.96	8294.04	13086.46	13194.09
9676.00	9734.02	13086.46	13194.09	9676.00	9734.02	10791.44	10878.40
8775.95	8833.94	8415.91	8473.96	8415.91	8473.96	12411.34	12498.41
12411.34	12498.41	10245.88	10304.02	8446.03	8503.94	9406.09	9463.96
8476.01	8534.03	12276.52	12363.34	8205.98	8263.96	8955.88	9013.95
12636.48	12723.57	12051.50	12138.72	8985.93	9043.98	12186.62	12273.48
8415.91	8473.96	13618.20	13734.14	9015.98	9073.98	11241.29	11328.54

Table 7: Feature and Continuum bandpasses selected for predicting *Metallicity* using noiseless BT_Settl spectra with signal-to-noise ratios equal to 10 and 50.

spectra. The RMDSE means the root median square error, as a way to provide robust estimation against outliers.

The comparison with the effective temperatures compiled by Cesetti et al. (2013) does not show significant differences either. In general, all classifiers tend to predict lower effective temperatures than those in the literature.

Most models show remarkably similar distributions of the predictions when trained with different SNR levels. In the cases of GA-MARS, GA-KNN, and GA-SVR this is the case even in the unrealistic scenario of SNR=∞.

	$SNR = 10$		$SNR = 50$		$SNR = \infty$	
RegressionModels	*RMSE*	*RMDSE*	*RMSE*	*RMDSE*	*RMSE*	*RMDSE*
$\chi^2 BTSettl$	232.55	100.00	235.25	120.00	232.45	100.00
ICA + ppr	241.93	127.75	241.83	**99.14**	279.70	161.90
rf	308.46	183.03	247.65	136.48	**167.21**	135.23
gbm	287.21	159.92	247.93	149.01	233.22	112.59
svr	**221.46**	122.36	281.21	150.59	299.05	160.22
nnet	283.44	191.58	264.16	**114.35**	326.35	211.81
knn	238.18	120.00	**232.36**	137.50	**219.14**	**100.00**
mars + bagging	253.14	113.47	254.00	**95.44**	**226.14**	133.49
kpls	275.48	120.00	299.88	**118.61**	387.06	217.53

Table 8: RMSE and RMDSE for the various regression models that predict T_{eff} (K).

	$SNR = 10$	$SNR = 50$	$SNR = \infty$
χ^2	-77.45	-86.88	-85.00
RuleRegression	-101.78	-38.38	169.74
rf	-172.61	-127.12	-5.32
gbm	-140.87	-108.53	31.70
svr	-57.62	-2.93	91.68
nnet	-146.57	-36.12	39.10
knn	-75.57	-109.85	-66.60
mars + bagging	-56.96	-87.88	98.22
pls	-120.32	-4.23	213.68

Table 9: Bias for T_{eff} (K) over the Cesseti estimation.

In general, models tends to produce better behaved solutions (with smaller biases and less scatter) for SNR=50. We interpret this value as representative of the SNR of the majority of spectra in the IRTF collection. 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 proper 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 applies holds for the model trained with SNR=10 spectra. **I should move this explanation to the beginning of the section or to the methodology section.**

We then compare the predicted effective temperatures with the spectral types listed in the IRTF spectral library. We attempted a direct comparison with the literature values gathered in Cesetti et al. (2013) but it only returns 57 estimates of effective temperature for M stars. We converted the spectral types into effective temperatures using the calibration of Stephens et al. (2009).

TODO: Luis, cambiar spectral libraries por stellar atmosphere models o synthetic spectral libraries.

	$SNR = 10$		$SNR = 50$		$SNR = \infty$	
RegressionModels	*RMSE*	*RMDSE*	*RMSE*	*RMDSE*	*RMSE*	*RMDSE*
$\chi^2 BTSettl$	0.82	0.45	0.93	0.61	3.5	3.48
ICA + ppr	**0.54**	0.48	0.34	**0.17**	**0.72**	**0.57**
rf	**0.64**	**0.38**	**0.77**	0.72	**0.53**	**0.39**
gbm	**0.48**	0.45	**0.61**	**0.47**	**0.49**	**0.41**
svr	**0.66**	**0.40**	**0.63**	**0.58**	**0.46**	**0.21**
nnet	**0.78**	0.61	**0.47**	**0.44**	1.2	0.97
mars + bagging	0.84	0.57	**0.54**	**0.37**	0.99	0.76
knn	1.23	0.83	1.39	1.44	1.60	1.32
kpls	0.99	0.99	**0.51**	**0.49**	0.96	0.77
RuleRegression	**0.74**	0.57	**0.50**	**0.47**	**0.57**	**0.41**

Table 10: RMSE and RMDSE for the various regression models predicting $Log(G)$ [dex].

Forecast quality of models was tested by the error against the temperature estimated based on the Spectral Subtype for each of the IRTF available spectra (see **??**). Both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) where calculated and it is presented in the table **??**.

From this comparison several things arise:

- The behavior of χ^2 distance is quite stable against SNR in the original dataset (BT_Settl) with a slightly better global performace in fovour of SNR=50.
- Models trained with different SNR=∞ have similar performance but heavy differences appear when SNR features are considered.
- When synchronous behavior is observed FT00, FT11, FT55, the better SNR is 10.
- Best set of features to be used for forecast are those from SNR=∞ (FT0b).
- As a conclusion the better performance was produced by FT01, followed by the FT51.

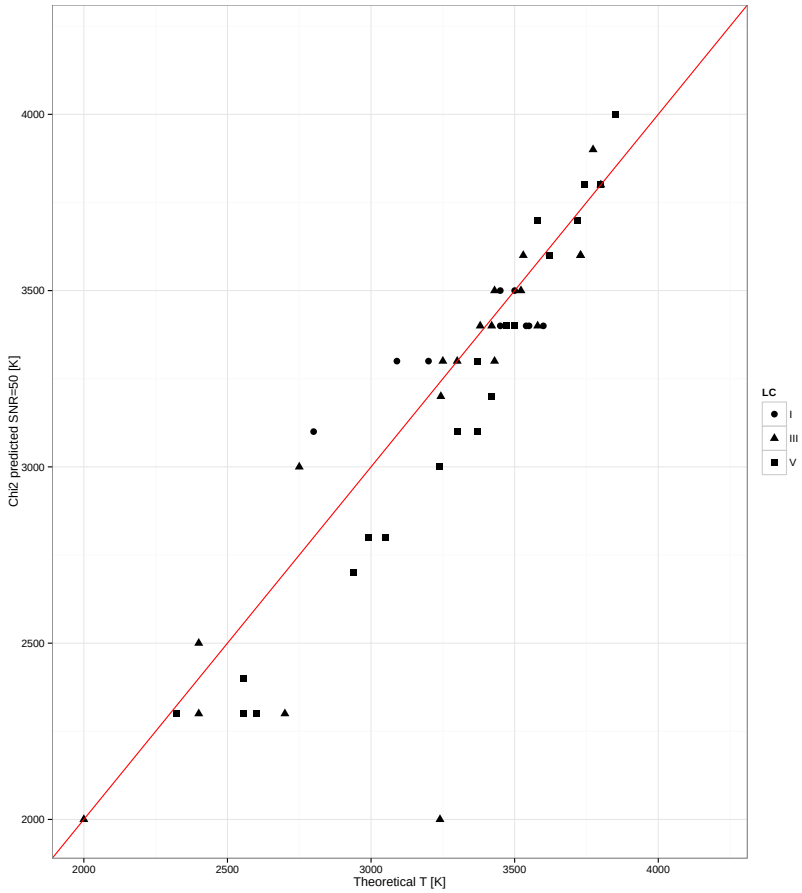
In Figures **?? ??** the relationship between Temperature estimated from the GA technique proposed features and modeled with different technqiues and the chi^2 with SNR=50 against the estimations provided by Cesetti et al. (2013) can be seen.

3.5. Surface gravity models

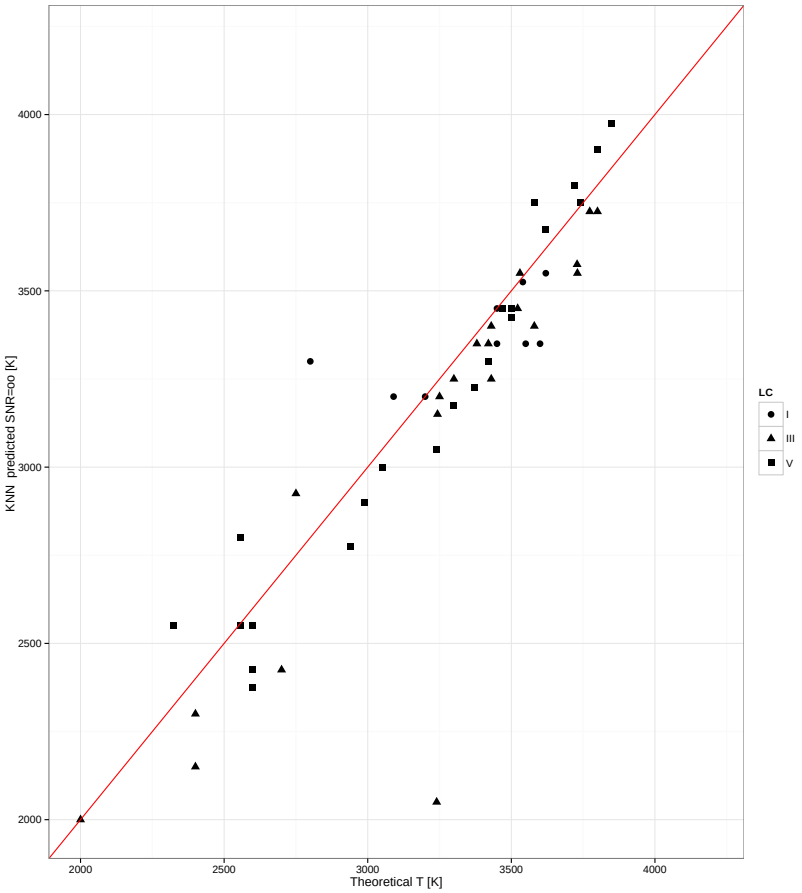
The same approach can become useful to produce $log(G)$ estimations. Here comparisons can only be possible between GA based features, the global spectra based approach with χ^2 distance to be minimized and those stars with gravity was estimated in **?**.

The only difference with the methodology presented above is because Temperature has been considered a fixed feature in the estimation of Gravity.

In Table 10 we can see the analysis of performance between the χ^2 identificacion and the one based on features from the spectrum depending on several classes of features. All of them have been carried out against the **?** estimations but, unfortunatelly those authors do not provide excessive number of IRTF class M stars to validate (just eight items).

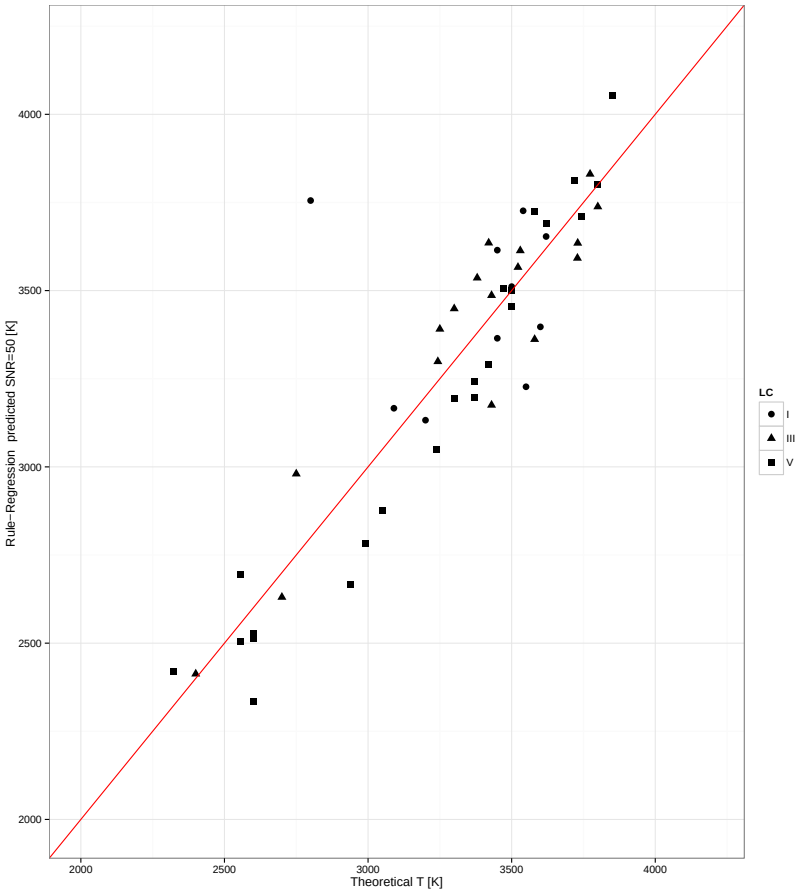


(a) Comparison between Temperature estimations from Cesetti in x axis and the closest BT_Settl spectra by χ^2 at SNR=50 on y-axis

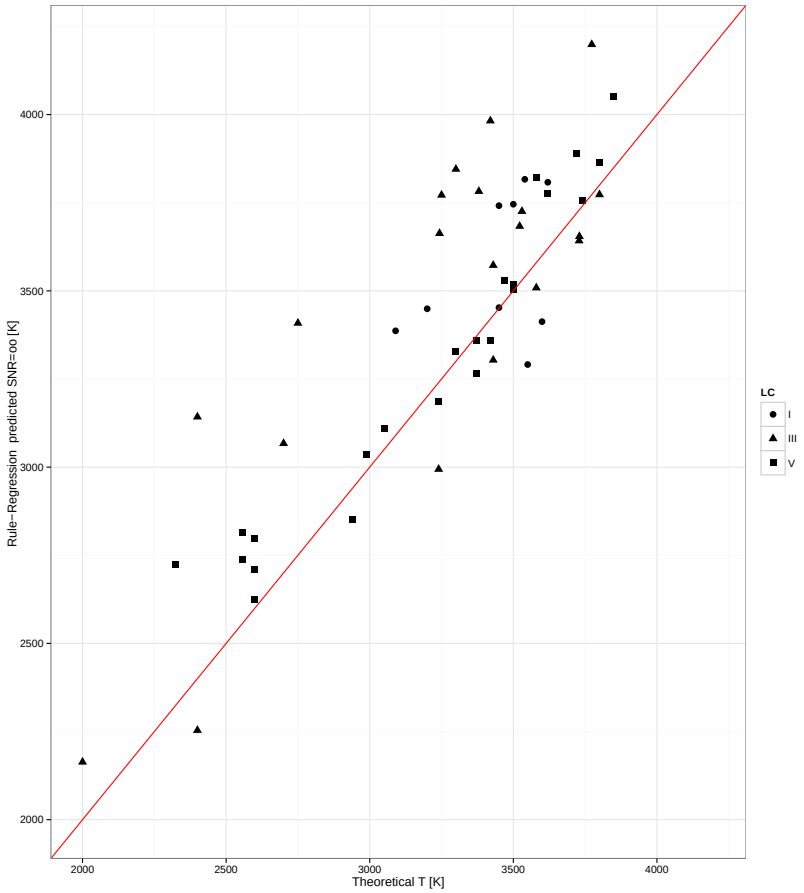


(b) Comparison between Temperature estimations from Cesetti in x axis and the KNN model for GA based features at SNR=∞ on y-axis

Fig. 1: Performance comparison between the χ^2 based selection and the band oriented features



(a) Comparison between Temperature estimations from Cesetti in x axis and the Rule Regression model for GA based features at SNR=∞ on y-axis



(b) Comparison between Temperature estimations from Cesetti in x axis and the Rule Regression model for GA based features at SNR=50 on y-axis

Fig. 2: Performance comparison between two Rule regression based models with different SNR

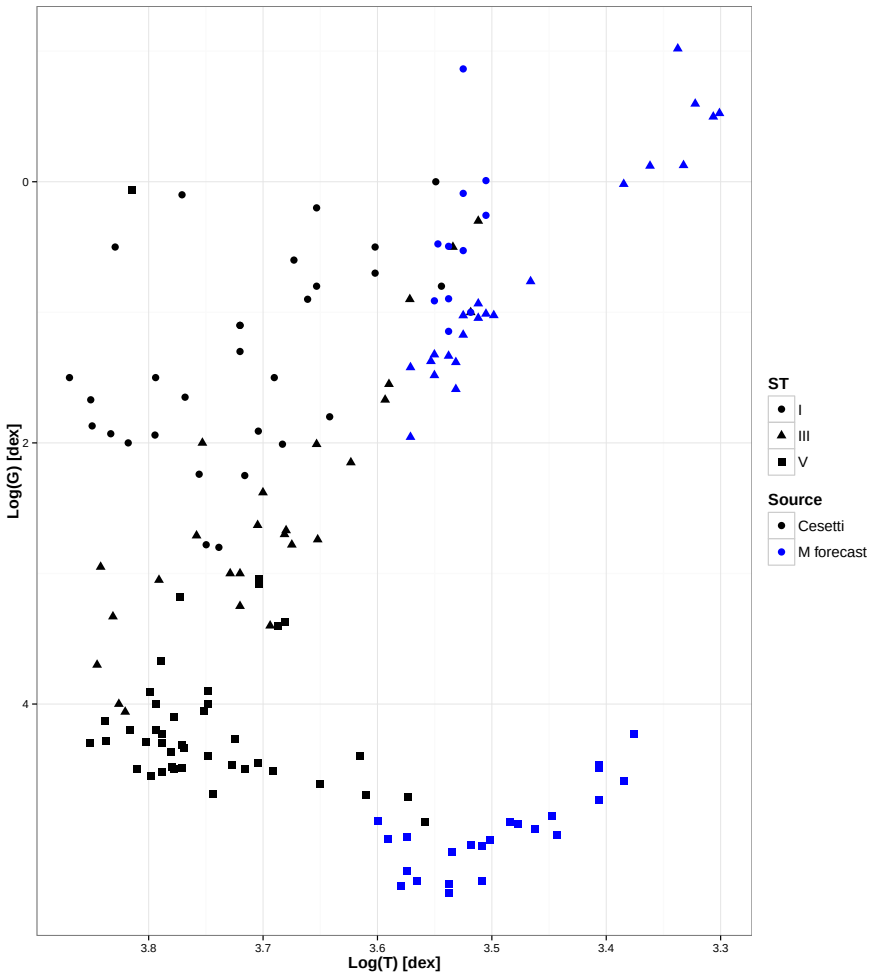


Fig. 3: Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for KNN model based on the GA provided bandpass features with SNR=∞

It is possible to present relationships between $log(g)$ and $log(T_{eff})$ as a matter of congruence analysis between predictions. In the Figure 3 such relationship is presented for models based on artificial intelligence selected features.

3.6. Metallicity models

Finally, the same analysis is performed for the Metalicty parameter, again by considering Temperature as a fixed feature. In Table 11 we can see the analysis of performance of different classes of models and cosidering a variety in features, even though just six IRTF stars where estimated by Cesetti et al. (2013), which strongly reduces the validation technique.

	$SNR = 10$		$SNR = 50$		$SNR = \infty$	
RegressionModels	*RMSE*	*RMDSE*	*RMSE*	*RMDSE*	*RMSE*	*RMDSE*
$\chi^2 BTSettl$	0.76	0.22	0.36	0.18	0.36	0.18
ICA + ppr	**0.24**	**0.13**	**0.31**	0.22	0.43	0.27
rf	**0.33**	0.25	0.73	0.41	0.61	0.36
gbm	**0.27**	**0.19**	0.70	0.52	0.63	0.35
svr	**0.33**	0.22	0.45	0.32	0.92	0.89
nnet	**0.37**	0.30	**0.33**	0.37	0.95	0.81
knn	**0.69**	0.55	**0.23**	**0.15**	**0.21**	**0.15**
mars + bagging	**0.36**	**0.16**	0.49	0.41	0.83	0.85
RuleRegression	**0.31**	**0.17**	**0.30**	0.24	0.78	0.23

Table 11: RMSE and RMDSE for the various regression models predicting *Met* [dex].

4. Physical parameters of the IPAC collection of spectra.

4.1. Spectral bands selected

During the preprocessing stage (and in a similar procedure as used in the case of the IRTF spectra) the spectral resolution of the BT-Settl library was degraded to match the average resolution of spectra in the Dwarf Archives by convolving with a Gaussian. Then, the spectra were trimmed to produce valid segments between *** and *** Å, which is the spectral range common to all M stars in the archive. Finally, all spectra were divided by the total integrated flux in this range in order to factor out the stellar distance.

4.2. Spectral features for estimation of effective temperatures.

The application of the GAs to the selection of features for the prediction of effective temperature from noiseless spectra within the IPAC wavelength range and resolution results in the features included in Table 12. Features are ordered by the fitness value (the AIC) and we only consider features that are present in at least 5 sets.

λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
7062	7094.4	7314	7346.4
7116	7148.4	7782	7814.4
7134	7166.4	7872	7904.4
6900	6932.4	7764	7796.4
7170	7202.4	7890	7922.4
7080	7112.4	7926	7958.4
7188	7220.4	7548	7580.4
7800	7832.4	7962	7994.4
6990	7022.4	7008	7040.4
7026	7058.4	6990	7022.4

Table 12: Features selected by the GA for predicting T_{eff} using BT_Settl noiseless synthetic spectra.

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

Table 13: Recommended features and Continuum bandpass for predicting T_{eff} by using BT_Settl with SNR= 10 and 50.

λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
7134	7166.4	7044	7076.4
6954	6986.4	7152	7184.4
7512	7544.4	7890	7922.4
7062	7094.4	7224	7256.4
6936	6968.4	7854	7886.4
6900	6932.4	7746	7778.4
6918	6950.4	7800	7832.4
7008	7040.4	7134	7166.4
7872	7904.4	7008	7040.4
7962	7994.4	7980	8012.4

Table 14: Recommended features and continuum bandpasses for predicting log(g) obtained using noiseless BT_Settl spectra.

TBD by Luis: interpret the features.

When noise is added to the BT-Settl spectra, we obtain

For gravity (in the form of log(g)) estimation, the GA search procedure produces the features presented in Tables 14 and 17 for the pure synthetic signal and signal-to-noise ratios of 10 and 50, respectively.

Finally, the best features found by the GA for metallicity estimation are listed in Table 16 for the noiseless BT-Settl spectra, and in Table **??** for signal-to-noise ratios equal to 10 and 50.

When signal-to-noise ratios equal to 10 and 50 are considered, the GA finds the selected features listed in Table **??**.

4.3. Regression models

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

Table 15: Recommended features and continuum bandpasses for predicting $log(g)$ obtained using BT_Settl with SNR= 10 and 50.

λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
7188	7220.4	7854	7886.4
7080	7112.4	7926	7958.4
7116	7148.4	7098	7130.4
7422	7454.4	7836	7868.4
7350	7382.4	7998	8030.4
7224	7256.4	7818	7850.4
7710	7742.4	7062	7094.4
7476	7508.4	7944	7976.4
7134	7166.4	7584	7616.4
7836	7868.4	7278	7310.4

Table 16: Feature and Continuum bandpasses selected for predicting *Metallicity* using noiseless BT_Settl spectra.

SNR = 10				SNR=50			
λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$	λ_1	λ_2	$\lambda_{cont;1}$	$\lambda_{cont;2}$
7692	7724.4	7026	7058.4	7098	7130.4	7926	7958.4
6900	6932.4	7008	7040.4	7188	7220.4	7962	7994.4
7350	7382.4	7908	7940.4	7368	7400.4	7980	8012.4
6918	6950.4	6900	6932.4	7116	7148.4	7872	7904.4
7098	7130.4	7314	7346.4	7062	7094.4	7206	7238.4
7440	7472.4	7872	7904.4	7584	7616.4	7170	7202.4
7134	7166.4	7962	7994.4	6936	6968.4	6918	6950.4
7368	7400.4	7926	7958.4	7692	7724.4	7890	7922.4
7080	7112.4	7044	7076.4	7134	7166.4	7548	7580.4
7044	7076.4	7980	8012.4	7494	7526.4	7998	8030.4

Table 17: Feature and Continuum bandpasses selected for predicting *Metallicity* using noiseless BT_Settl spectra with signal-to-noise ratios equal to 10 and 50.

5. Conclusions

Acknowledgements. This research has benefitted from the M, L, T, and Y dwarf compendium housed at DwarfArchives.org. The authors also thanks to the Spanish Ministery for Economy and Innovation because of the support obtained through the project with ID: AyA2011-24052. IRTF library 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.

References

Allard, F., Homeier, D., Freytag, B., et al. 2013, Memorie della Societa Astronomica Italiana Supplementi, 24, 128

Cesetti, M., Pizzella, A., Ivanov, V. D., et al. 2013, A&A, 549, A129

Fuhrmeister, B., Schmitt, J., & Hauschildt, P. 2005, arXiv preprint astro-ph/0505375

Goldberg, D. E. et al. 1989, Genetic algorithms in search, optimization, and machine learning, Vol. 412 (Addison-wesley Reading Menlo Park)

Heiter, U., Jofré, P., Gustafsson, B., et al. 2015, ArXiv e-prints

Holland, J. H. 1975, Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. (U Michigan Press)

Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16

R Core Team. 2013, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria

Stephens, D. C., Leggett, S. K., Cushing, M. C., et al. 2009, ApJ, 702, 154

List of Objects