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

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

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

[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 January 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.

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.

The spectral resolution was degraded to the IRTF resolution factor (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.

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**

2.1. 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.

The kind of spectral features that were evaluated took the form of flux ratios between two spectral bands (from λ_1 to λ_2, and from λ_3 to λ_4):

$$F = \frac{\int_{\lambda_1}^{\lambda_2} f(\lambda) \cdot \mathrm{d}\lambda}{\int_{\lambda_3}^{\lambda_4} f(\lambda) \cdot \mathrm{d}\lambda} \tag{1}$$

where $f(\lambda)$ denotes the normalized flux. We will search for the band boundaries that produce physical parameter predicitions with the smallest errors.

Other types of features defined as

$$F' = \int_{\lambda'_1}^{\lambda'_2} (1 - \frac{f(\lambda)}{F_{cont}} \cdot \mathrm{d}\lambda) \tag{2}$$

where considered, where F_{cont} is the average flux in a spectral band between $\lambda_{cont;1}$ and $\lambda_{cont;2}$. This definition produced somewhat poorer results at least when applied to the lower resolution IPAC spectra (see below). We interpret this as a result of the difficulty in defining a continuum in very late-type spectra.

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.

predictive of the phenotypic variables. Multivariate selection methods

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.

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 then a predefined value.

Stage 4: The population of chromosomes is replicated. Chromosomes with a higher fitness score will generate a 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.

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 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.

From the implementation point of view, we used a binary codification of the chromosomes and a parallel implementation of the GA in a farm of fifteen computers per physical parameter.

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.

3. Physical parameters of the IPAC collection of spectra.

3.1. Spectral bands selected

3.2. Regression models

4. Physical parameters of the IRTF collection of spectra.

4.1. Spectral bands selected

After applying this technique the recommended features for temperature can be found in Table 1.

The authors have estimated the suggested features when theoretical BT_Settl is noised with different SNR and following tables 2 and 3 sumarize the findings.

As in (Cesetti et al. (2013)) the authors provided their best estimation for suitable features, our interest is also to verify how good it becomes in our particular case, as it can be an inderect assessment for the quality of the GA based recommendation.

As a matter of reference the bandpass presented in Table 4 exploits the following bandpass:

In regards with the Gravity, the GA recommends the features presentend in Table 5 for the pure synthetic signal.

The authors have produced the estimations for different SNR again as depicted in the tables 6 and 7.

Finally, features suggested for metallicity can be found in Table 8.

And when different SNR are considered, the suggested features can be found in tables 9 and 10

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	8376.10	8433.91	9346.13	9403.92	-6693.10	319
Feature 2	8385.99	8443.94	9346.13	9403.92	-6700.15	6
Feature 3	8195.96	8254.03	9386.01	9444.05	-6887.55	44
Feature 4	8186.06	8243.98	9235.98	9294.01	-7056.23	19
Feature 5	8406.00	8464.07	9515.96	9574.13	-7068.86	34

Table 1: Recommended features and Continuum bandpass for predicting T_{eff} by using BT_Settl with SNR= ∞ . The Fitness and frequency of occurence are also included.

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	8385.99	8443.94	9395.94	9454.03	-6734.59	136
Feature 2	8186.06	8243.98	9536.15	9593.96	-6857.65	7
Feature 3	8186.06	8243.98	9376.07	9433.92	-6947.54	7
Feature 4	8286.01	8343.92	9206.05	9264.00	-7123.10	10
Feature 5	8355.96	8414.03	9066.05	9124.05	-7207.55	37

Table 2: Recommended features and Continuum bandpass for predicting T_{eff} by using BT_Settl with SNR= 10 . The Fitness and frequency of occurence are also included.

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	8376.10	8433.91	9346.13	9403.92	-6693.10	545
Feature 2	8286.01	8343.92	9186.03	9244.04	-7250.83	17
Feature 3	8476.01	8534.03	9525.89	9584.05	-7392.49	38
Feature 4	9276.00	9333.87	9425.95	9484.00	-7578.17	19
Feature 5	9555.93	9614.06	8886.00	8943.99	-7652.77	54

Table 3: Recommended features and Continuum bandpass for predicting T_{eff} by using BT_Settl with SNR= 50 . The Fitness and frequency of occurence are also included.

	Signal_from	Signal_To	Cont1_From	Cont1_To	Cont2_From	Cont2_To
Pa1	8461	8474	8474	8484	8563	8577
Ca1	8484	8513	8474	8484	8563	8577
Ca2	8522	8562	8474	8484	8563	8577
Pa2	8577	8619	8563	8577	8619	8642
Ca3	8642	8682	8619	8642	8700	8725
Pa3	8730	8772	8700	8725	8776	8792
Mg	8802	8811	8776	8792	8815	8850
Pa4	8850	8890	8815	8850	8890	8900
Pa5	9000	9030	8983	8998	9040	9050
FeClTi	9080	9100	9040	9050	9125	9135
Pa6	9217	9255	9152	9165	9265	9275

Table 4: Recommended features and Continuum bandpass recommended in Cesetti et al. (2013) as relevant for temperature inside Band I

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	8176.03	8234.13	8295.99	8353.99	-1244.00	278
Feature 2	8176.03	8234.13	8955.88	9013.95	-1506.62	8
Feature 3	8636.06	8694.06	8536.03	8594.06	-1515.42	14
Feature 4	8486.02	8544.05	8985.93	9043.98	-1519.48	50
Feature 5	8496.05	8554.06	9436.02	9493.86	-1520.16	6

Table 5: Recommended features and Continuum bandpass for predicting $log(g)$ by using BT_Settl with SNR= ∞ . The Fitness and frequency of occurence are also included.

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	8176.03	8234.13	8295.99	8353.99	-1244.00	248
Feature 2	8176.03	8234.13	8305.94	8364.04	-1252.85	16
Feature 3	8176.03	8234.13	8266.11	8324.03	-1264.86	9
Feature 4	8556.06	8614.04	8716.00	8773.99	-1489.47	33
Feature 5	8536.03	8594.06	9096.06	9154.07	-1517.47	18

Table 6: Recommended features and Continuum bandpass for predicting $log(g)$ by using BT_Settl with SNR= 10 . The Fitness and frequency of occurence are also included.

4.2. Regression models

After producing the suitable set of features for each of the physical parameters we are interested in, the next step will be to produce the effective model becoming useful to predict those parameters. The researcher will decide how many features will be used in the multivarite model proposed to explain the physical parameter for the learning dataset (BT_Settl). The models produced by this way will be used to forecast the physical parameters for the IRTF library.

As a matter of analysis different cross-comparison tests were performed, like performance against the parameters infered from the closer BT_Settl by using the χ^2 distance with different

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	8176.03	8234.13	8205.98	8263.96	-1320.54	50
Feature 2	8415.91	8473.96	8166.02	8224.12	-1400.77	10
Feature 3	8645.93	8703.94	8665.99	8723.96	-1422.86	8
Feature 4	8515.98	8573.99	8205.98	8263.96	-1504.27	9
Feature 5	9425.95	9484.00	9146.00	9204.05	-1512.67	13

Table 7: Recommended features and Continuum bandpass for predicting $log(g)$ by using BT_Settl with SNR= 50 . The Fitness and frequency of occurence are also included.

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	9085.96	9144.03	9445.97	9504.01	-1146.72	348
Feature 2	9445.97	9504.01	9085.96	9144.03	-1150.63	24
Feature 3	8556.06	8614.04	9135.89	9193.92	-1209.47	11
Feature 4	9096.06	9154.07	8466.08	8523.98	-1271.45	8
Feature 5	9045.91	9103.99	8525.91	8583.93	-1276.04	5

Table 8: Recommended features and Continuum bandpass for predicting *Metallicity* by using BT_Settl with SNR= ∞ . The Fitness and frequency of occurence are also included.

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	9476.15	9534.00	9576.03	9634.04	-1199.37	17
Feature 2	8466.08	8523.98	9146.00	9204.05	-1207.84	189
Feature 3	8466.08	8523.98	9156.03	9214.07	-1208.98	10
Feature 4	9466.08	9523.82	9416.04	9474.02	-1220.23	12
Feature 5	9335.79	9393.93	9186.03	9244.04	-1225.37	26

Table 9: Recommended features and Continuum bandpass for predicting *Metallicity* by using BT_Settl with SNR= 10 . The Fitness and frequency of occurence are also included.

	Signal From	Signal To	Continuum From	Continuum To	Fitness	Freq
Feature 1	9085.96	9144.03	9445.97	9504.01	-1146.72	177
Feature 2	9445.97	9504.01	9085.96	9144.03	-1150.63	5
Feature 3	8556.06	8614.04	9135.89	9193.92	-1209.47	6
Feature 4	9096.06	9154.07	8466.08	8523.98	-1271.45	6
Feature 5	9045.91	9103.99	8525.91	8583.93	-1276.04	5

Table 10: Recommended features and Continuum bandpass for predicting *Metallicity* by using BT_Settl with SNR= 50 . The Fitness and frequency of occurence are also included.

SNR. Comparisons were indeed performed against other inductive Machine Learning strategies, like project the spectra in a smaller feature dimensional space by using Independent Component Analysis (ICA) and then, developping a regression model based on such features (see 4.4). For the special case of temperature a comparison between the temperature and the known spectral subclase makes possible to analyze the quality of the forecasted estimations (see 4.5).

4.3. 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:

SNR	Features	SVM	RF	GAM	MARS
50	Cesetti et al.	81.6	83.3	163.5	91.9
	GA	91.4	82.2	161.1	91.9
10	Cesetti et al.	135.8	138.5	268.8	166.8
	GA	123.2	122.6	212.6	130.9

Table 11: RSME for different models predicting T_{eff} [K].

- 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 Machine Models with Gaussian Kernel *(SVM)*.
- MLP Nerural Networks *(NNET)*.

Comparison of performance between sets of features for temperature derived from the GA based strategy can be analyzed, over the same testing dataset of BT_Settl and it was depicted in Table 11.

After calculating the bartlett test for both cases of SNR it was seen that variances are homogeneous since p > 0.05, and the Flinger-Killen shows that homkedascity is verified, then F-ANOVA test makes clear that there is no significative difference between models. Then, it is possible to conclude that quality of features from both sources are equivalent regarding modeling capability, even when GA only has proposed five features and Cesetti et al. requires seven features.

4.4. Full Spectra Oriented Models

As an alternative to build modles based on bandpasses, a similiar methodology to the one depicted in (Sarro et al. (2013)) was implemented.

For the projection an Independend Component Analysis (ICA) with ten dimensions was used and for Temperature regression an optimized SVM with parameters of C=10 and ϵ=0.001.

Considering the Gravity case, the most suitable ICA had twentysix dimensions and the best SVM parameters were C=1000 and ϵ = 0.001. This was the same case for Metallicity.

In terms of interpretation, this methodology looks to predict the physical paramenters by considering the whole star spectrum instead of information provided by specific bands. Thus it can be interesting to analyze suitability for prediction against the other approach.

In the same sense it was decided to consider direct selection, which is also a technique based on the whole espectrum but, instead of regressing specific parameters, the closest labeled spectrum to the one under analysis is identified by a χ^2 distance. This becomes possible as interpolation between labeled spectra can be easily performed.

4.5. Temperature model based on Spectral Subtype.

4.6. Temperature Model for T_{eff}.

After training the set of models by using labelled BT_Settl dataset, those models were used to predict the IRTF temperature. The authors were interested in understand how relevant the SNR factor becomes in terms of model training and in terms of forecasting. Thus, performance analysis between direct spectra comparison by means of χ^2 and models using bandpass features was carried out. Only the most five relevant features based of the exhibited fitness were considered. Comparisons between models trained with different SNR and tested against features from other SNR were performed. Notation FTab will mean Forecasted temperature when a accounts for the SNR of the feature set considered for the forecast and b accounts for the SNR used for training the model. Both a and b have the meaning of 0 for SNR of ∞, 1 for SNR=10 and 5 for SNR=50. Training was performed by 10 fold cross valdiation technique, making possible to select de convenient model.

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 4.5). Both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) where calculated and it is presented in the table 12.

	RMSE	MAE
chi2d_10	428.83	261.53
chi2d_50	426.31	267.77
FT00	427.33	307.38
FT01	366.82	248.38
FT05	429.02	299.58
FT10	438.61	327.93
FT11	**410.11**	**291.79**
FT15	427.34	316.98
FT50	420.70	300.69
FT51	375.82	259.02
FT55	430.10	303.54

Table 12: RSME & MAE for different Random Forest models predicting T_{eff} [K].

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.

However a comparison between performance of different type of models with the same set of features has been performed. In table 13 the RMSE is presented for those different models and in table 14 the MAE is presented.

	rf	gbm	boosting	svm	gam	nnet	mars
FT00	427.33	431.65	419.67	376.60	1116.91	751.23	552.21
FT01	366.82	364.80	356.12	302.88	385.65	371.68	444.40
FT05	429.02	430.20	420.40	393.73	433.09	2514.35	492.70
FT10	438.61	443.49	446.94	321.55	2201.21	1269.82	590.28
FT11	410.11	407.80	400.37	359.50	430.10	419.32	487.69
FT15	427.34	439.24	428.89	317.02	371.22	2738.34	509.44
FT50	420.70	427.82	420.78	326.10	3742.74	1243.80	551.57
FT51	375.82	370.44	388.54	312.66	390.91	437.61	448.20
FT55	430.10	431.68	424.52	361.94	378.87	432.15	490.68
chi2d_10	428.83	428.83	428.83	428.83	428.83	428.83	428.83
chi2d_50	426.31	426.31	426.31	426.31	426.31	426.31	426.31

Table 13: RSME for different models predicting T_{eff} [K].

	rf	gbm	boosting	svm	gam	nnet	mars
YTn00	307.38	311.42	303.22	282.06	836.62	531.86	349.10
FT01	248.38	250.10	246.03	221.95	258.83	255.74	267.70
FT05	299.58	305.64	301.72	285.80	314.94	2378.76	335.36
FT10	327.93	331.29	334.81	254.95	1514.75	1192.57	389.98
FT11	291.79	291.86	291.25	272.35	305.62	304.18	319.47
FT15	316.98	327.49	314.83	250.50	271.28	2636.53	353.85
FT50	300.69	307.18	308.42	252.72	2559.17	1049.08	361.82
FT51	259.02	255.30	272.60	226.32	263.27	318.01	274.19
FT55	303.54	309.75	307.33	269.07	274.98	308.48	333.29
chi2d_10	261.53	261.53	261.53	261.53	261.53	261.53	261.53
chi2d_50	267.77	267.77	267.77	267.77	267.77	267.77	267.77

Table 14: MAE for different models predicting T_{eff} [K].

In Figure **??** the relationship between Temperature estimated from the GA model proposed features with SNR=50 and features from SNR=∞ and the Temperature estimation from spectral subtype in comparison with the chi^2 with SNR=50 can be seen.

The comparison against processing the whole spectrum by ICA projection has been performed and the results for SNR={10,50} can be seen in Figure 2b and Figure **??**.

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 Cesetti et al. (2013).

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

In Tables 15 and 16 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.

In Figure 2c and Figure 2d relationships between $log(g)$ predicted by global espectrum estimation and GA feature based estimation can be observed.

Finally, the same analysis is performed for the Metalicty parameter, again by considering Temperature as a fixed feature. In Tables 17 and 18 we can see the analysis of performance of different classes of models and cosidering a variety in features.

In Figure **??** and Figure 3b relationships between metalicity predicted by global espectrum estimation and GA feature based estimation against the real values provided by Cesetti et al. (2013) can be observed.

	rf	gbm	boosting	svm	gam	nnet	mars
G_chi2_10	1.68	1.68	1.68	1.68	1.68	1.68	1.68
G_chi2_50	1.79	1.79	1.79	1.79	1.79	1.79	1.79
FG00	2.01	1.62	2.32	1.78	0.98	3.39	2.13
FG01	2.52	2.56	2.62	1.87	2.52	2.32	2.45
FG05	2.49	2.42	2.40	1.78	2.29	2.89	2.16
FG10	2.34	2.59	2.75	1.78	32.52	3.39	3.04
FG11	2.49	2.48	2.69	1.90	2.67	2.50	2.57
FG15	2.75	2.72	2.51	1.78	2.95	3.94	2.52
FG50	2.61	2.11	2.58	1.78	17.95	3.39	6.07
FG51	2.78	2.82	2.77	1.92	2.73	2.43	2.65
FG55	2.58	2.57	2.71	1.78	2.80	2.34	2.63

Table 15: RMSE for different models predicting $Log(G)$ [dex].

	rf	gbm	boosting	svm	gam	nnet	mars
G_chi2_10	1.46	1.46	1.46	1.46	1.46	1.46	1.46
G_chi2_50	1.46	1.46	1.46	1.46	1.46	1.46	1.46
FG00	1.78	1.50	2.05	1.54	0.82	3.06	1.84
FG01	2.14	2.19	2.27	1.75	2.18	1.80	2.07
FG05	2.13	2.07	2.07	1.35	1.59	2.70	1.54
FG10	2.09	2.31	2.43	1.54	27.48	3.06	2.73
FG11	2.12	2.16	2.34	1.77	2.30	2.03	2.17
FG15	2.35	2.29	2.17	1.35	2.52	3.64	1.86
FG50	2.50	1.99	2.23	1.54	15.75	3.06	4.03
FG51	2.46	2.48	2.43	1.79	2.50	1.89	2.37
FG55	2.15	2.16	2.34	1.35	2.48	2.05	2.29

Table 16: RMSE for different models predicting $Log(G)$ [dex].

	rf	gbm	boosting	svm	gam	nnet	mars
M_Chi2_10	0.19	0.19	0.19	0.19	0.19	0.19	0.19
M_Chi2_50	0.35	0.35	0.35	0.35	0.35	0.35	0.35
FM00	0.30	0.43	0.28	1.04	2.19	0.51	1.03
FM01	0.51	0.46	0.44	0.74	0.76	0.99	50.64
FM05	2.05	2.85	1.31	1.89	3.46	7.56	6.46
FM10	1.09	1.02	0.94	1.04	10.75	1.65	13.49
FM11	0.47	0.39	0.49	0.74	0.31	0.45	43.43
FM15	1.91	2.73	1.08	1.89	4.65	13.27	16.95
FM50	0.82	0.87	0.43	1.04	6.10	2.25	11.87
FM51	1.02	1.10	0.56	0.74	2.29	3.44	119.29
FM55	1.70	3.14	1.15	1.89	7.64	7.00	12.04

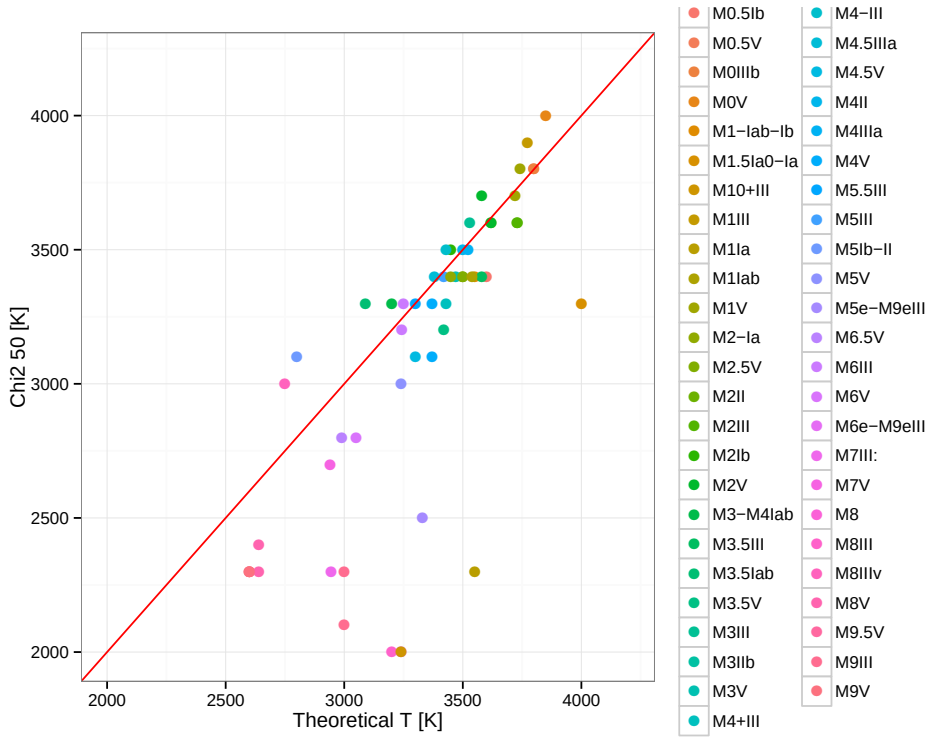
Table 17: RMSE for different models predicting Met [dex].

It is possible to present relationships between $log(g)$ and $log(T_{eff})$ as a matter of congruence analysis between predictions. In the Figure 4 such relationship is presented for models based on artificial intelligence selected features. In Figure 5 the values for $log(T_{eff})$ and $log(G)$ are inferred from the closest BT_Settl spectra.

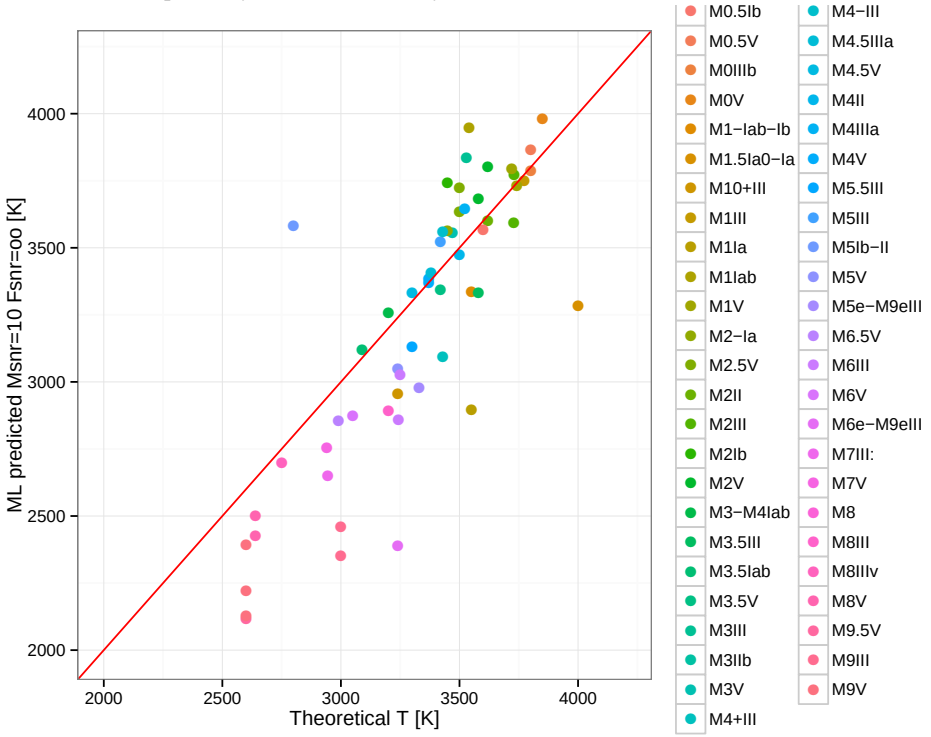
And, for sure, it is possible to do it for estimations based on parameters from nearest labeled BT-Settl spectra. In this particular case, it is possible to see how considering the global spectrum is positive for stronger physical parameters like T_{eff} but the approach reduces drastically its likelihood when other softer parameters are involved.

	rf	gbm	boosting	svm	gam	nnet	mars
M_Chi2_10	0.17	0.17	0.17	0.17	0.17	0.17	0.17
M_Chi2_50	0.26	0.26	0.26	0.26	0.26	0.26	0.26
FM00	0.24	0.38	0.25	1.02	2.01	0.33	0.92
FM01	0.47	0.40	0.40	0.72	0.66	0.90	33.30
FM05	2.04	2.85	1.29	1.88	3.41	7.38	6.28
FM10	0.80	0.71	0.62	1.02	8.88	0.99	10.82
FM11	0.43	0.37	0.46	0.72	0.24	0.37	25.59
FM15	1.90	2.68	1.05	1.88	3.68	11.77	13.21
FM50	0.78	0.79	0.40	1.02	5.13	1.83	9.90
FM51	1.01	1.08	0.54	0.72	2.22	3.36	77.52
FM55	1.67	3.10	1.13	1.88	7.11	6.33	11.22

Table 18: MAE for different models predicting *Met* [dex].

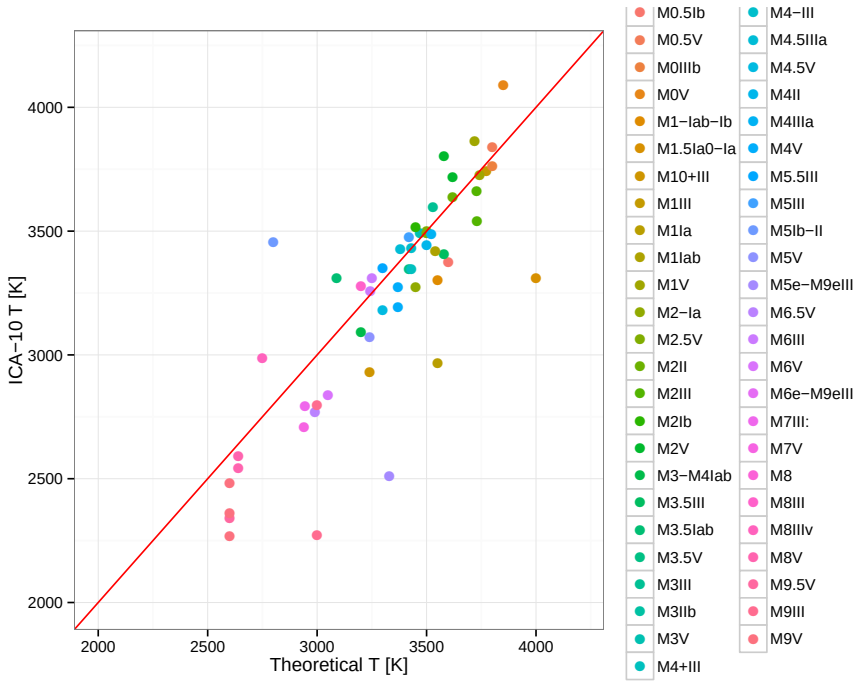


(a) Comparison between Temperature estimations from Spectral Subtype in x axis and the closest BT_Settl spectra by χ^2 at SNR=50 on y-axis

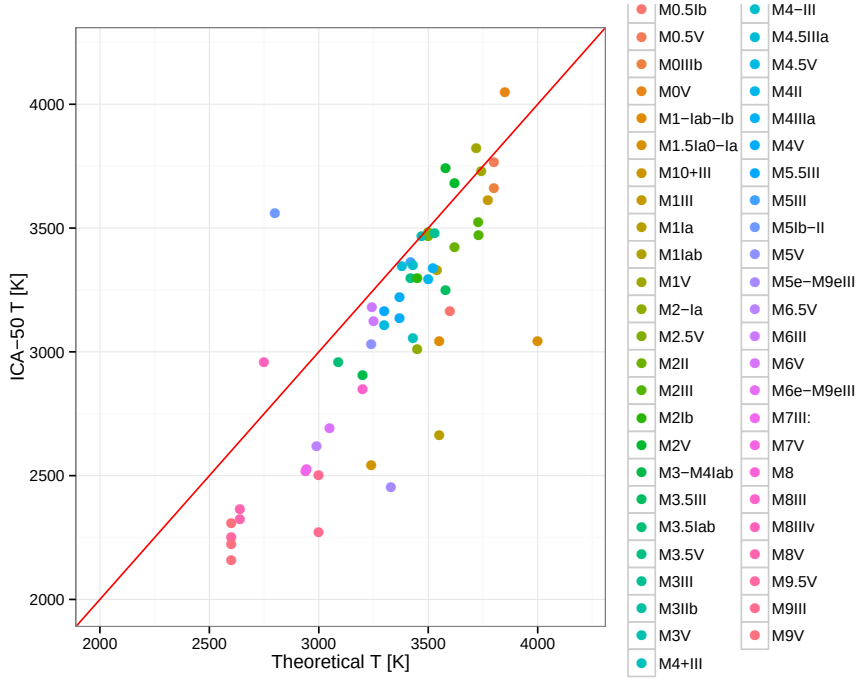


(b) Comparison between Temperature estimations from Spectral Subtype in x axis and the Support Vector Machines for Ga based features trained with BT_Settl at SNR=∞ and features for forecasting at SNR=10 on y-axis

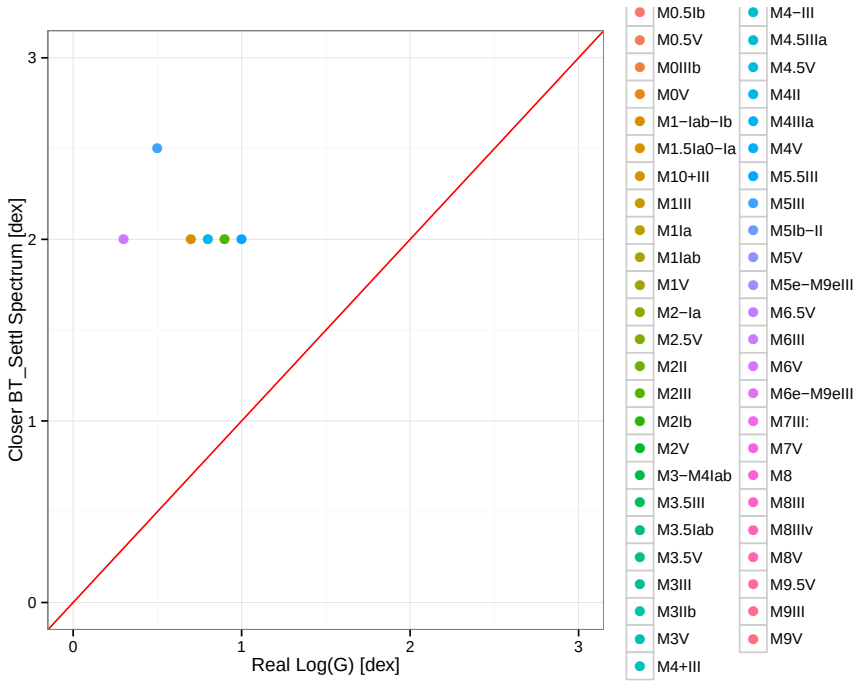
Fig. 1: Performance comparison between the *chi*^{2} based selection and the band oriented features



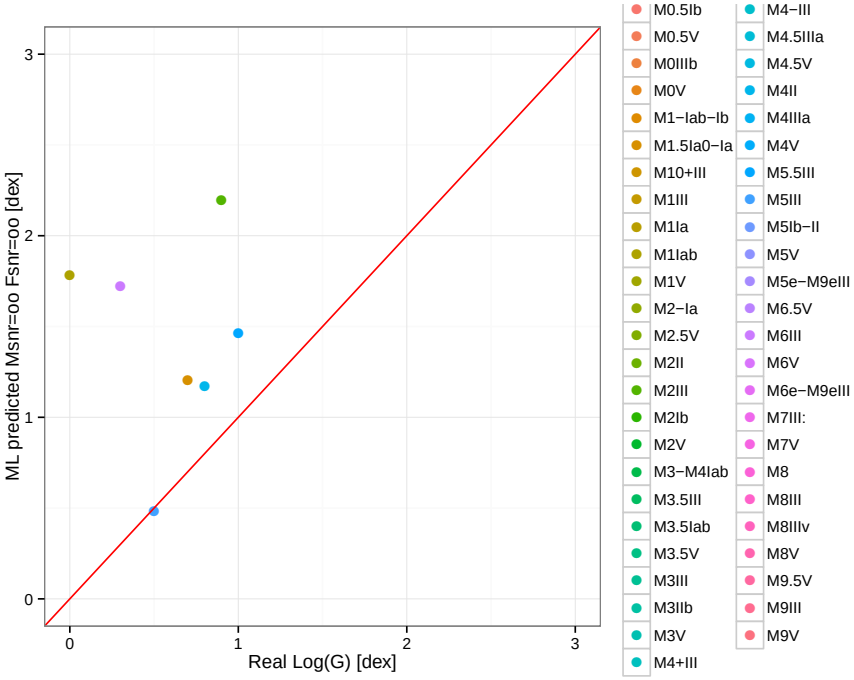
(a) Comparison between Temperature estimations from Spectral Subtype in x axis and the prediction based on SVM models over the ICA projection with 10 components at SNR=10 on y-axis



(b) Comparison between Temperature estimations from Spectral Subtype in x axis and the prediction based on SVM models over the ICA projection with 10 components at SNR=50 on y-axis

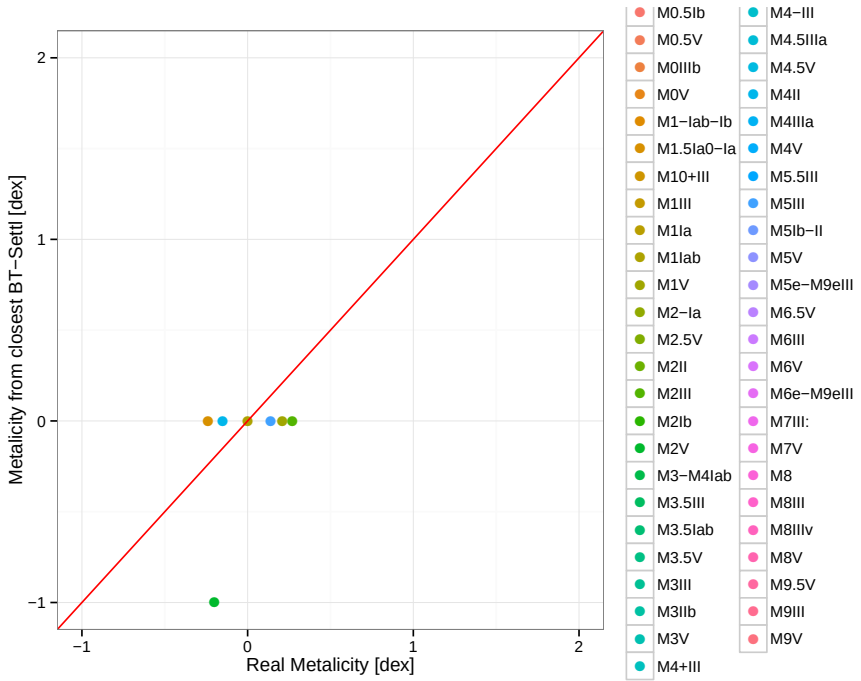


(c) Comparison between Gravity estimations from Spectral Subtype in x axis and the closest BT_Settl spectra by χ^2 at SNR=50 on y-axis

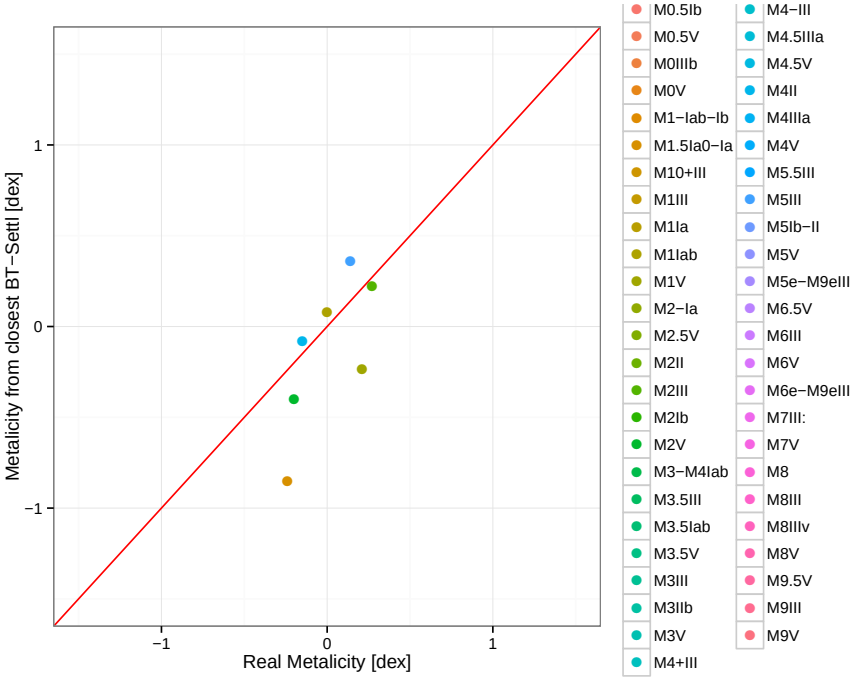


(d) Comparison between Gravity estimations from Spectral Subtype in x axis and the Support Vector Machines for Ga based features trained with BT_Settl at SNR=∞ and features for forecasting at SNR=∞ on y-axis

Fig. 2: Performance comparison between the *chi*^{2} based selection and the band oriented features to forecast Log(g)



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



(b) Comparison between Metalicity estimations from Spectral Subtype in x axis and the Support Vector Machines for Ga based features trained with BT_Settl at SNR=∞ and features for forecasting at SNR=∞ on y-axis

Fig. 3: Performance comparison between the *chi*^{2} based selection and the band oriented features to forecast Log(g)

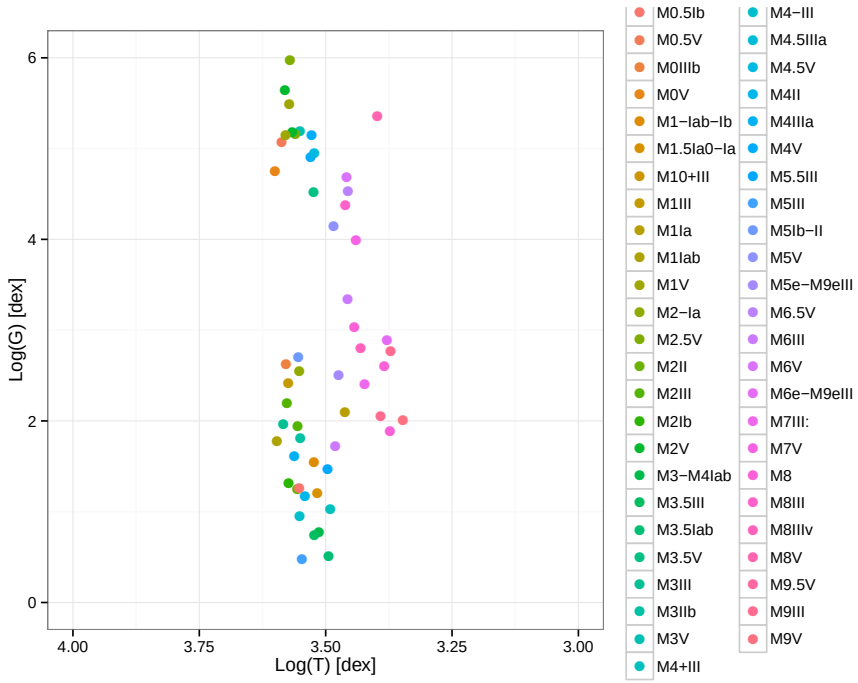


Fig. 4: Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for models based on bandpass features

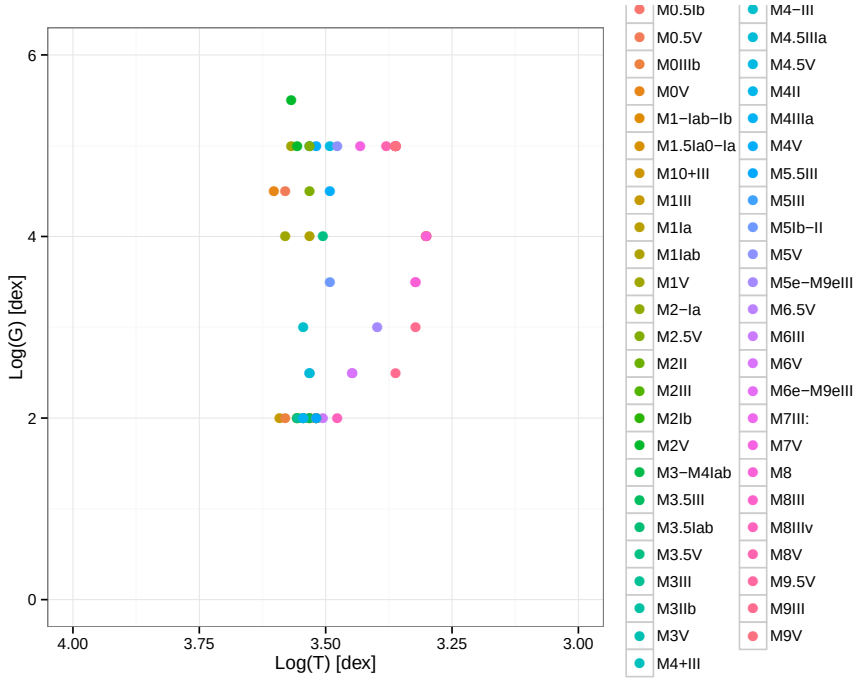


Fig. 5: Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=50 when the nearest BT-Settl spectrum is used.

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)

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

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

Sarro, L. M., Debosscher, J., Neiner, C., et al. 2013, A&A, 550, A120

List of Objects