Astronomy & Astrophysics manuscript no. irtf_m
January 2, 2015

Effective temperatures, gravities and metallicities for M type stars in the IRTF database.

I. An approach from the Machine Learning Arena

L. Sarro-Baro[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 January 2, 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. Literature Review

3. Experimental work

This research, in accordance with the goals depicted in the previous section, started by considering the collection of M stars provided by DwarfArchives.org, a compendium of L, M and T dwarfs, maintained by Chris Gelino, Davy Kirkpatrick and Adam Burgasser.

3.1. Dataset Selection.

The M dwarf database includes around 68 spectra with spectroscopic observations obtained at Infrared Telescope Facility (IRTF) on Mauna Kea Hawaii, with the medium-resolution spectrograph, SpeX, at the NASA. The IRTF Spectral Library is a collection of $0.8 \sim 5.0\,\mu$m mostly stellar spectra observed at a resolving power of $R \equiv \lambda/\Delta\lambda \sim 2000$. Rayner et al. (2009). [1]

Features of the library include:

- A spectral range of 0.8 to 2.5 μm.
- S/N $\sim$ 100 at $\lambda < 4\mu$m.
- Spectral continuum shape is preserved.
- Absolutely flux calibrated using Two Micron All Sky Survey (2MASS) photometry.

As the goal was to develop a procedure making possible to identify suitable reference bands for signal and continuum in this particular class of stars, it was decided to use synthetic spectra. We have choosed the library BT-Settl (Allard et al. (2013)) where several operations have been performed.

3.2. Reshape of the theoretical library.

Firstly a filterage looking for spectra between 2000 and 4200K was performed over the whole BT-Settl dataset, considering log(gravitiy) in range (-4, -6) dex with a step of 0.5 dex. Metallicity observed was between 0, -0.5 and -1 dex. Total size of avaialble spectra is 535 (100K of distance each).

Spectra degradation from the original 0.1Å stepsize till the required according to the IRTF resolution factor (R 2000) was accomplished by a gaussian convolution considering datapoints closer than $\pm 3\sigma$, defined because of the $FWHM$ estimation depending on the position and R as $FWHM = 2 * ln(2) * R$. Then, the spectra were trimmed to produce valid segments between 8145.924 and 24106.846Å. Indeed, in order to become independent of the star's distance normalization the area under spectrum has been performed to value of one.

The authors have conducted also an interpolation procedure to increase the number of available learning patterns. To be sure that interpolation was a valid solution to inferre new synthetic spectra, a formal destillation of some spectra by using the PHOENIX code (Fuhrmeister et al. (2005)) was performed and then, compared to the one obtained by interpollating in accordance to the inverse square of the distance among the closest neighborgs available (see Fig. 1).

Then several other datasets can be created, by defining a mesh of 0.25 dex for both, log(gravity) and Metallicity. Temperature step was selected to be 50K, which produced 1329 spectra. Then, another reinterpolation is produced, with a new step for temperature of 25K and 0.125 dex for log(gravity), keeping the metallicity step in 0.25 dex and producing a dataset with 25912 spectra. They can be used as needed for specific research pourposes. In spite of these, and in order to keep

[1] Hay una mejor descripción de la librería IRTF en Cesetti et al., pero no se si merece la pena copiar su texto aquí

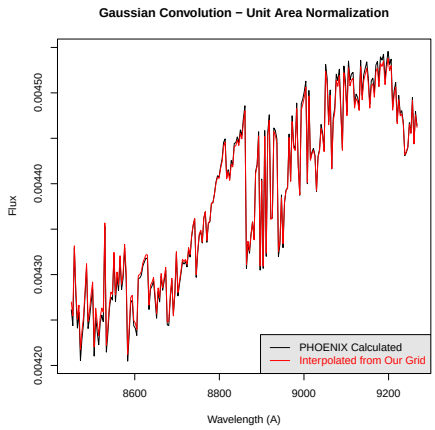


Fig. 1: Comparison between generated and interpollated spectrum

theirknowledge closer to the original BT-Settl source, most of the analyses have been performed with the original 535 spectra dataset.

The synthetic spectra are theoretical and they are noise free but it does not happen with real spectra, the IRTF dataset in our case. To increase similarities gaussian noise have been added with two different Signal to Noise Ratio(SNR). The selected SNR were 10 and 50.

3.3. Feature definition

It is well known that due to the special characteristics of physical parameters for these stars it is not easy to define suitable signal bands but it is even more complicated to identify the continuum chunks, were no signal shall be expected, because of the relative low temperatures they have and its impact on the emission rates at different wavelengths.

There are technical methodologies to identify suitable bands, as the one presented in (Cesetti et al. (2013)). In that case, bands I and K have been considered. His methodology is mainly based on the sensitivity exhibited by different spectra through different regions of the spectra. These observations were carried out by visual inspection.

The approach adopted in this paper was to estimate the three star relevant physical parameters (T_{eff}, Gravity and Metallicity) establishing a set of features, based on which regression models could be useful for each of the physical parameters.

As the temperature is considered an strong parameter, it was used as known feature for the regression of the Gravity and Metalicity physical parameters.

The central contribution of this paper is the way to identify the most suitable regions of spectrum (features) to be considered as signal providers but mainly the regions to be considered as candidates for being continuum for those signal regions. It will be performed by means of artificial intelligence techniques.

These features consist of a central bandpass covering the interesting lines and another bandpass referring to the local continuum. Then, the feature can be written like Eq. (1).

$$F(i) = \frac{\int_{\lambda_{1s}(i)}^{\lambda_{2s}(i)} (f(\lambda))\, d\lambda}{\int_{\lambda_{1c}(i)}^{\lambda_{2c}(i)} (f(\lambda))\, d\lambda} \qquad \forall i \in [1 \dots N] \tag{1}$$

where:

- N means the number of features to be considered, as decided by the researcher.
- $f(x)$ denotes the normalized flux spectra from the star.
- $(\lambda_{1s}, \lambda_{2s})$ Å accounts for the region of spectrum where the signal is considered.
- $(\lambda_{1c}, \lambda_{2c})$ Å accounts for the region of spectrum where the continuum is considered.

Now, the research question is how to identify $\{\lambda_{1s}(i), \lambda_{2s}(i), \lambda_{1c}(i), \lambda_{2c}(i)\} \quad \forall i \in [1 \dots N]$ in such a way they become useful to estimate the physical parameters.

Some constaints were estabished as designing of the selection process:

- $|\lambda_{2k(i)} - \lambda_{1k(i)}| = 30$ pixels of spectrum $\quad \forall i \in [1 \dots N]$ and $k \in \{s, c\}$.
- $min(\lambda_{1k(i)}, \lambda_{1k(j)}) = 5$ pixels of spectrum $\quad \forall i, j \in [1 \dots N]$; $i \neq j$ and $k \in \{s, c\}$.
- $\overline{\lambda_2(i)\lambda_1(i)} \bigcap \overline{\lambda_{2c}(i)\lambda_{1c}(i)} = \emptyset \quad \forall i \in [1 \dots N]$.

which become a guarantee avoiding any overlap and a minimum size for both signal and continuum bandpasses.

3.4. Determination of Features

Search for those features will depend on which specific physical parameter is under consideration but, the proposed methodology will look for those values trying to solve an optimization problem, which shall be the forecast capabilities of one specific set of features, to be retained when it becomes bigger than a threshold.

To accomplish such optimization problem involving the selection of variable subsets, the use of the Genetic Algorithms technique was accepted.

It was proposed to use the software tools R(R Core Team (2013)). There are different statistics to identify features that are differentially expressed between two or more groups of samples and then uses the most differentially expressed to construct a statistical model.

These methods have demonstrated to perform well, 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 may seem to be more suitable for the analysis of data since variables 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.

For the sake of simplicity let us define Genetic Algorithms (GAs) are variable search procedures that are based on the principle of evolution by natural selection. The procedure works by evolving sets of variables (chromosomes) that fit certain criteria from an initial random population via cycles of differential replication, recombination and mutation of the fittest chromosomes. The concept of using in-silico evolution for the solution of optimization problems has been introduced by John Holland in 1975 (Holland (1975)). Although their application has been reasonably widespread (see Goldberg's book (Goldberg et al. (1989)), they became very popular only when sufficiently powerful computers became available.

The implementation of the GA follows the next steps:

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

Stage 2: Each chromosome in the population is evaluated for its ability to predict the group membership of each sample in the dataset (fitness function).

Stage 3: Chromosome preselection, 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 that new genes are used in chromosomes. Steps 5 and 6 are applied over the chromosomes establised at Step 4.

Stage 7: This process is repeated from Stage 2 until enough accuracy is obtained or the maximum of iterations is attained.

The features were built in our case as indicated in 1 and they were named according to the ordinal of the wavelength step both for signal and continuum. The name includes also the number of ofset induced because of the constraint 3.3. Polulation size was choosed as one thousand individuals and accepted iterations were four thousand. Three randomly started different repetions where produced bringing the opportunity for enough variety and probabilities were established as 0.85 to corossover and 0.35 to mutation. The elitism was fixed to be 0.15. Fitness for features were established as related to the Akaike Criterion (-AIC) for linearity between the potential feature against the physical parameter. The most frequent and efficient features were suggeswted as candidates to describe the behavior of specific physical parameters.

From the implementation point of view a binarized codification was selected in accordance to the naming convention and in order to speed up the computation, a parallel implementation from a farm of fifteen connected computers were used for each of the physical parameters.

The GA procedure provides us with a large collection of chromosomes. Although these are all potential solutions of the problem, it is not clear which one should be chosen for developing a model becoming for interpretation. For this reason there is a need to develop a single model

	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
Feature 6	9326.07	9384.15	8406.00	8464.07	-7349.99	18
Feature 7	8496.05	8554.06	9576.03	9634.04	-7505.21	38
Feature 8	9036.07	9094.04	9075.93	9133.98	-7535.71	15
Feature 9	9135.89	9193.92	9085.96	9144.03	-7622.16	27
Feature 10	9515.96	9574.13	8876.08	8934.03	-7655.59	6
Feature 11	8716.00	8773.99	9025.93	9084.07	-7703.40	77
Feature 12	9156.03	9214.07	8255.97	8314.06	-7708.62	15
Feature 13	8266.11	8324.03	8235.96	8294.04	-7856.20	30
Feature 14	8235.96	8294.04	8255.97	8314.06	-7860.73	69
Feature 15	8705.93	8763.97	8886.00	8943.99	-7919.62	27
Feature 16	8536.03	8594.06	8336.02	8394.08	-7940.67	24
Feature 17	8605.97	8663.96	8346.02	8404.06	-7953.63	59
Feature 18	8946.07	9004.01	8756.09	8814.06	-8117.90	46
Feature 19	9135.89	9193.92	9485.83	9544.11	-8211.98	36
Feature 20	8536.03	8594.06	8496.05	8554.06	-8337.36	31

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.

that is, to some extent, representative of the population. The simpler strategy to follow is to use the frequency of the chromosme in the population of chromosomes as criteria for inclusion in a forward selection strategy, however for this particular applciation, the choice was to include features based on their highest fitness.

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	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
Feature 6	9276.00	9333.87	8415.91	8473.96	-7436.13	32
Feature 7	8455.96	8513.93	9055.94	9114.07	-7476.12	32
Feature 8	8616.00	8673.98	9576.03	9634.04	-7491.24	6
Feature 9	8536.03	8594.06	9135.89	9193.92	-7573.72	58
Feature 10	9395.94	9454.03	9356.05	9414.08	-7586.78	45
Feature 11	9576.03	9634.04	9045.91	9103.99	-7641.39	19
Feature 12	9066.05	9124.05	8166.02	8224.12	-7642.28	15
Feature 13	9386.01	9444.05	9536.15	9593.96	-7684.66	7
Feature 14	8616.00	8673.98	8756.09	8814.06	-7686.46	17
Feature 15	9576.03	9634.04	8145.92	8204.03	-7767.19	13
Feature 16	9466.08	9523.82	9036.07	9094.04	-7772.45	37
Feature 17	9445.97	9504.01	9545.87	9604.02	-7830.60	35
Feature 18	8286.01	8343.92	8486.02	8544.05	-7863.15	49
Feature 19	9186.03	9244.04	9135.89	9193.92	-7884.30	13
Feature 20	9306.03	9363.93	8745.93	8803.93	-8020.27	56
Feature 21	8305.94	8364.04	8215.93	8273.93	-8068.94	27
Feature 22	8186.06	8243.98	8326.00	8383.94	-8288.51	6
Feature 23	8786.02	8844.10	8886.00	8943.99	-8305.69	12
Feature 24	8855.96	8913.97	8366.04	8424.04	-8309.97	7
Feature 25	8235.96	8294.04	8795.98	8853.95	-8312.84	17
Feature 26	8786.02	8844.10	8385.99	8443.94	-8318.31	59
Feature 27	8186.06	8243.98	8795.98	8853.95	-8324.63	6
Feature 28	8855.96	8913.97	8286.01	8343.92	-8334.47	34

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
Feature 6	9536.15	9593.96	8195.96	8254.03	-7665.45	37
Feature 7	8515.98	8573.99	8936.05	8994.03	-7687.69	19
Feature 8	8605.97	8663.96	8846.03	8904.03	-7819.75	41
Feature 9	9285.87	9344.05	9096.06	9154.07	-7841.07	12
Feature 10	8775.95	8833.94	9036.07	9094.04	-7846.05	47
Feature 11	9346.13	9403.92	8826.01	8883.94	-7873.89	16
Feature 12	9566.01	9623.96	9105.87	9163.91	-7967.65	7
Feature 13	9566.01	9623.96	9536.15	9593.96	-7998.01	42
Feature 14	8536.03	8594.06	8385.99	8443.94	-8055.85	39
Feature 15	8846.03	8904.03	8395.98	8453.99	-8118.20	16
Feature 16	9135.89	9193.92	9455.86	9514.14	-8123.39	38
Feature 17	8515.98	8573.99	8415.91	8473.96	-8233.55	42
Feature 18	8235.96	8294.04	8886.00	8943.99	-8316.98	6
Feature 19	8366.04	8424.04	8195.96	8254.03	-8320.41	17
Feature 20	8305.94	8364.04	8846.03	8904.03	-8327.38	37
Feature 21	8795.98	8853.95	8835.93	8893.97	-8327.43	35
Feature 22	8835.93	8893.97	8795.98	8853.95	-8336.98	20

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
Feature 6	9085.96	9144.03	9276.00	9333.87	-1522.24	15
Feature 7	9315.88	9374.02	9566.01	9623.96	-1524.35	11
Feature 8	8985.93	9043.98	9285.87	9344.05	-1527.06	9
Feature 9	8245.98	8304.08	9006.05	9064.02	-1529.28	29
Feature 10	9156.03	9214.07	8376.10	8433.91	-1532.13	17
Feature 11	8215.93	8273.93	8596.11	8654.10	-1534.68	21
Feature 12	9276.00	9333.87	9395.94	9454.03	-1537.74	6
Feature 13	8235.96	8294.04	8205.98	8263.96	-1540.04	33
Feature 14	9576.03	9634.04	8145.92	8204.03	-1541.89	61

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
Feature 6	9135.89	9193.92	9536.15	9593.96	-1517.92	38
Feature 7	8446.03	8503.94	9135.89	9193.92	-1524.31	7
Feature 8	8446.03	8503.94	9306.03	9363.93	-1525.64	8
Feature 9	9506.13	9563.85	9306.03	9363.93	-1530.90	31
Feature 10	8925.98	8983.96	9485.83	9544.11	-1534.00	7
Feature 11	9455.86	9514.14	9265.98	9323.99	-1534.03	13
Feature 12	8355.96	8414.03	8406.00	8464.07	-1534.04	8
Feature 13	9216.01	9274.05	8336.02	8394.08	-1534.62	9
Feature 14	8925.98	8983.96	9436.02	9493.86	-1535.91	27
Feature 15	9295.98	9354.08	8596.11	8654.10	-1537.61	37
Feature 16	9255.86	9314.01	9365.95	9424.02	-1538.41	8
Feature 17	8955.88	9013.95	9265.98	9323.99	-1540.02	36
Feature 18	8566.08	8624.07	8585.96	8643.95	-1540.09	49
Feature 19	9576.03	9634.04	9085.96	9144.03	-1540.34	13
Feature 20	8446.03	8503.94	8665.99	8723.96	-1541.80	7
Feature 21	8446.03	8503.94	8556.06	8614.04	-1542.23	38
Feature 22	8726.06	8784.07	8315.97	8374.00	-1542.31	20

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.

	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
Feature 6	8486.02	8544.05	8936.05	8994.03	-1517.66	10
Feature 7	8476.01	8534.03	8976.09	9034.00	-1522.11	7
Feature 8	8446.03	8503.94	9455.86	9514.14	-1526.30	15
Feature 9	9156.03	9214.07	8346.02	8404.06	-1529.63	9
Feature 10	9636.16	9693.91	8215.93	8273.93	-1531.01	9
Feature 11	9135.89	9193.92	8385.99	8443.94	-1531.25	9
Feature 12	8865.98	8923.94	9536.15	9593.96	-1532.35	8
Feature 13	8665.99	8723.96	8685.95	8744.09	-1532.53	21
Feature 14	9126.00	9184.09	8215.93	8273.93	-1533.15	10
Feature 15	9525.89	9584.05	8355.96	8414.03	-1534.59	17
Feature 16	8616.00	8673.98	9406.09	9463.96	-1534.81	9
Feature 17	8626.02	8683.99	9335.79	9393.93	-1534.98	8
Feature 18	8305.94	8364.04	8286.01	8343.92	-1535.35	13
Feature 19	8685.95	8744.09	9096.06	9154.07	-1535.57	19
Feature 20	8636.06	8694.06	8235.96	8294.04	-1536.95	12
Feature 21	9395.94	9454.03	8826.01	8883.94	-1539.98	6
Feature 22	8195.96	8254.03	9115.99	9174.04	-1540.19	15
Feature 23	8786.02	8844.10	9235.98	9294.01	-1542.00	10

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
Feature 6	8466.08	8523.98	8936.05	8994.03	-1230.60	8
Feature 7	9255.86	9314.01	9545.87	9604.02	-1239.94	39
Feature 8	9285.87	9344.05	9495.98	9553.95	-1249.38	7
Feature 9	9576.03	9634.04	8855.96	8913.97	-1254.27	6
Feature 10	9285.87	9344.05	8505.89	8563.93	-1255.30	7
Feature 11	9285.87	9344.05	9455.86	9514.14	-1256.35	84
Feature 12	9636.16	9693.91	8245.98	8304.08	-1258.18	6
Feature 13	9525.89	9584.05	9285.87	9344.05	-1267.30	27
Feature 14	9545.87	9604.02	8826.01	8883.94	-1269.76	34
Feature 15	9096.06	9154.07	8255.97	8314.06	-1270.05	6
Feature 16	8286.01	8343.92	9235.98	9294.01	-1270.11	6
Feature 17	8685.95	8744.09	8286.01	8343.92	-1270.29	16
Feature 18	8775.95	8833.94	8556.06	8614.04	-1271.08	6
Feature 19	9096.06	9154.07	9436.02	9493.86	-1272.52	40
Feature 20	9485.83	9544.11	8336.02	8394.08	-1272.86	31
Feature 21	8406.00	8464.07	9306.03	9363.93	-1280.13	46
Feature 22	9045.91	9103.99	8665.99	8723.96	-1294.98	6
Feature 23	8305.94	8364.04	8835.93	8893.97	-1347.20	113
Feature 24	8385.99	8443.94	8205.98	8263.96	-1348.16	9
Feature 25	8765.97	8823.95	8395.98	8453.99	-1350.18	17

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.

4. Regression models for parameter estimation

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 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 4.2). 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 4.3).

4.1. 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:

- 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

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].

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.

4.2. 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 $\epsilon = 0.001$. 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.

4.3. Temperature model based on Spectral Subtype.

4.4. 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 4.3). Both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) where calculated and it is presented in the table 12.

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.

	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].

- 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 3b 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.

	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].

In Figure 4a and Figure 4b 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.

	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].

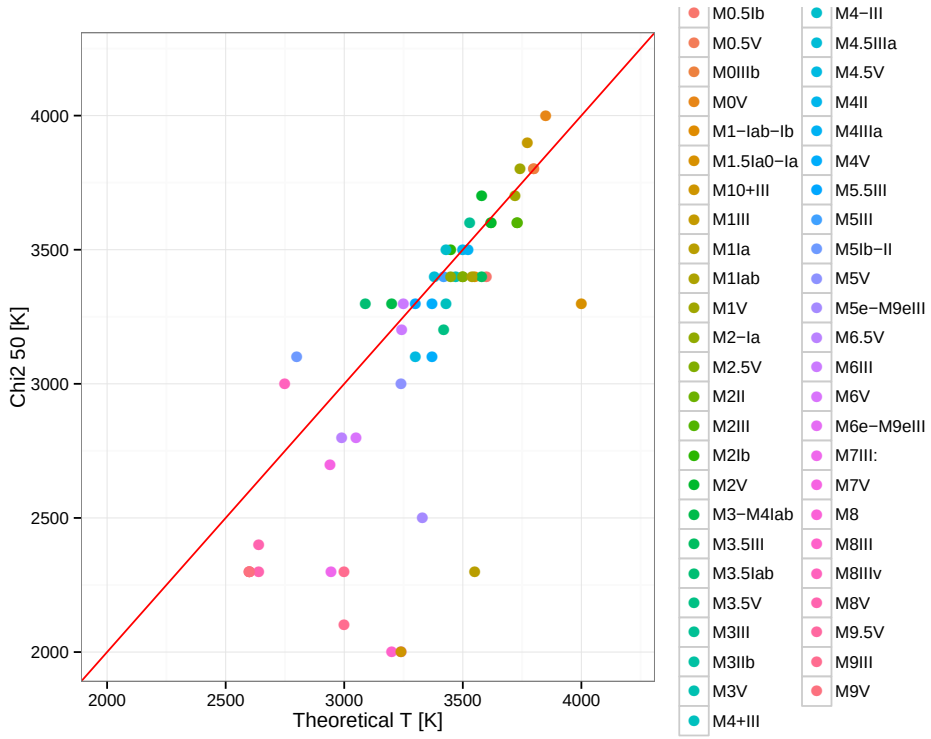
	rf	gbm	boosting	svm	gam	nnet	mars
M_Chi2_10	0.17	0.17	0.17	0.17	0.17	0.17	0.17
M_Chi2_50	0.26	0.26	0.26	0.26	0.26	0.26	0.26
FM00	0.24	0.38	0.25	1.02	2.01	0.33	0.92
FM01	0.47	0.40	0.40	0.72	0.66	0.90	33.30
FM05	2.04	2.85	1.29	1.88	3.41	7.38	6.28
FM10	0.80	0.71	0.62	1.02	8.88	0.99	10.82
FM11	0.43	0.37	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].

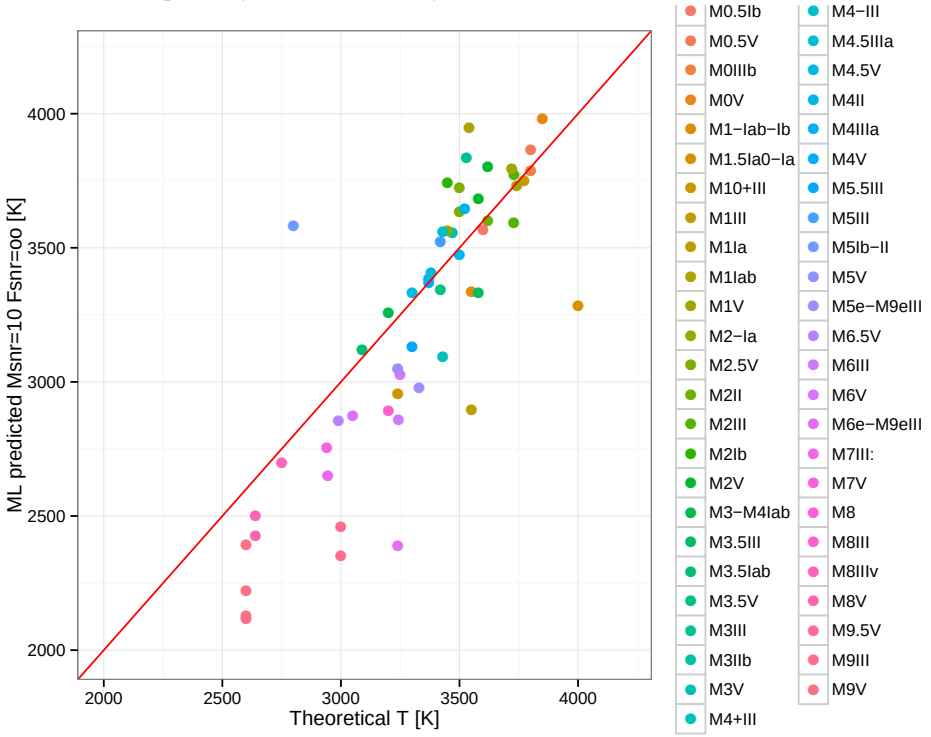
In Figure **??** and Figure 5b relationships between metalicity predicted by global espectrum estimation and GA feature based estimation against the real values provided by Cesetti et al. (2013) can be observed.

It is possible to present relationships between $log(g)$ and $log(T_{eff})$ as a matter of congruence analysis between predictions. In the Figure 6 such relationship is presented for models based on artificial intelligence selected features. In Figure 7 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.

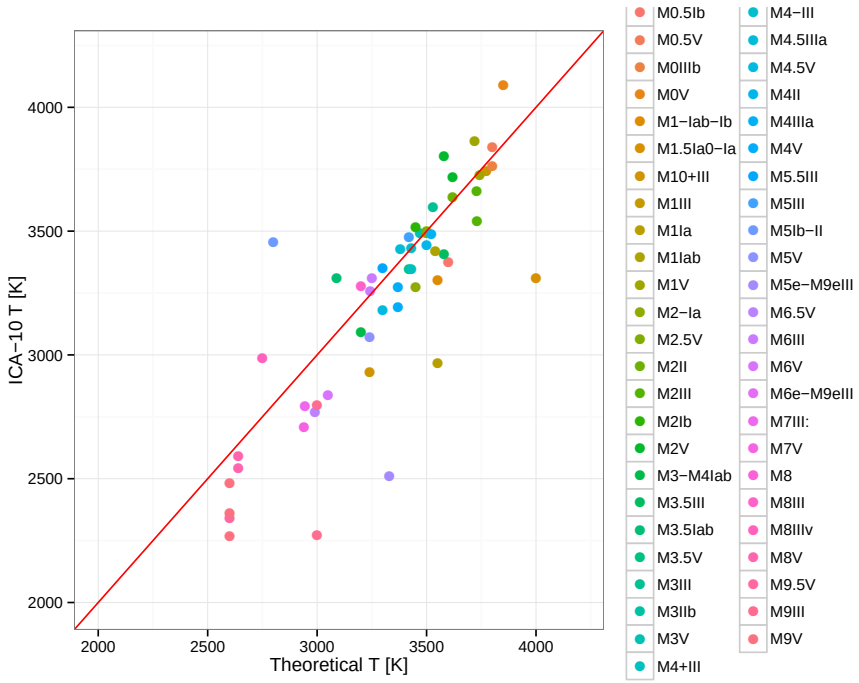


(a) Comparison between Temperature estimations from Spectral Subtype in x axis and the closest BT_Settl spectra by χ^2 at SNR=50 on y-axis

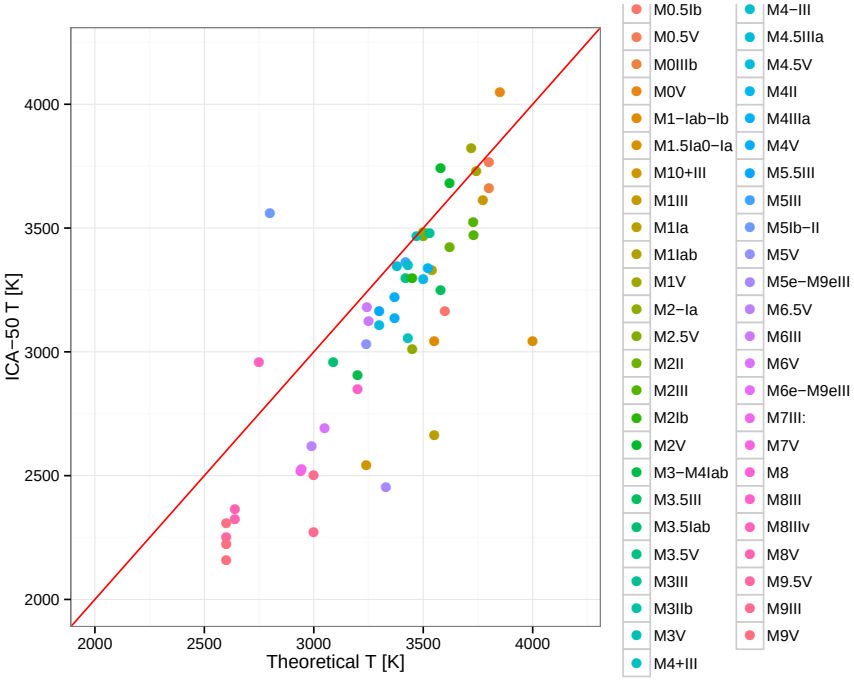


(b) Comparison between Temperature estimations from Spectral Subtype in x axis and the Support Vector Machines for Ga based features trained with BT_Settl at SNR=∞ and features for forecasting at SNR=10 on y-axis

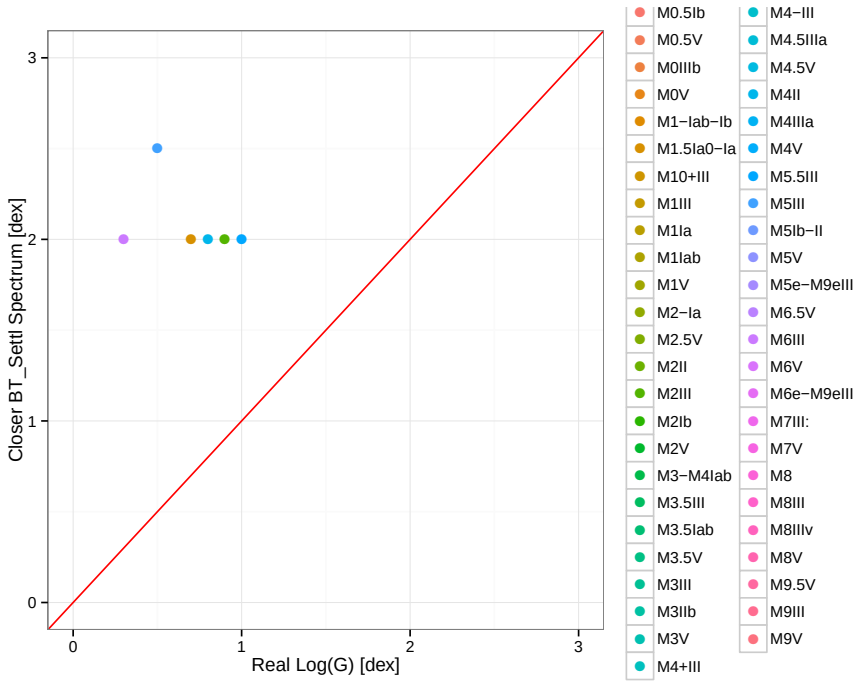
Fig. 2: Performance comparison between the *chi*^{2} based selection and the band oriented features



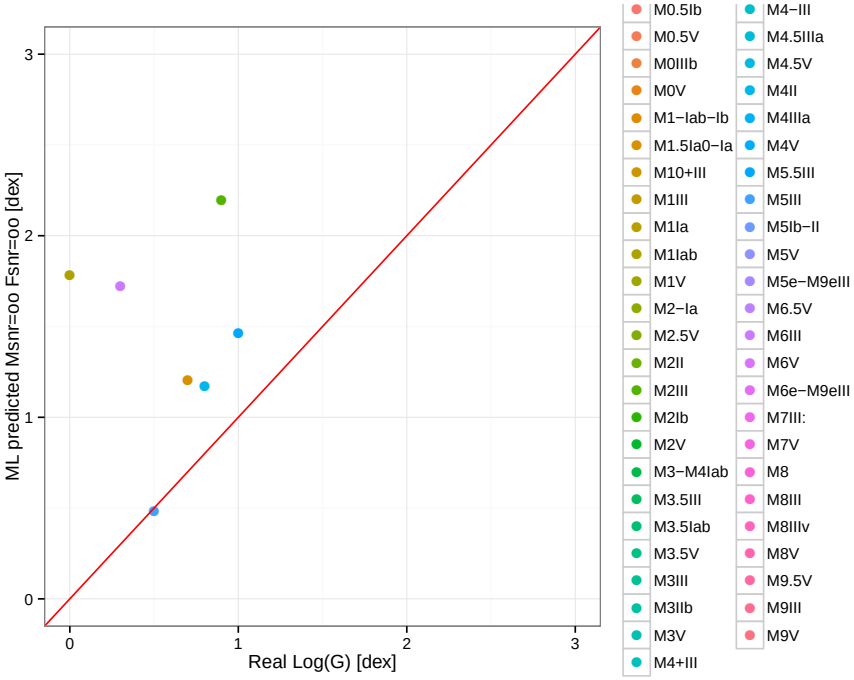
(a) Comparison between Temperature estimations from Spectral Subtype in x axis and the prediction based on SVM models over the ICA projection with 10 components at SNR=10 on y-axis



(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

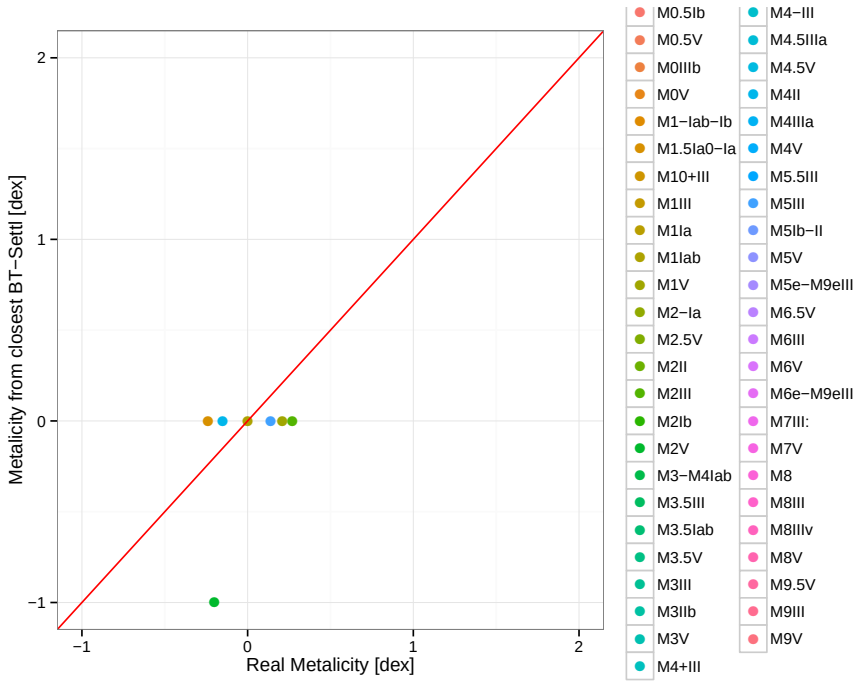


(a) Comparison between Gravity estimations from Spectral Subtype in x axis and the closest BT_Settl spectra by χ^2 at SNR=50 on y-axis

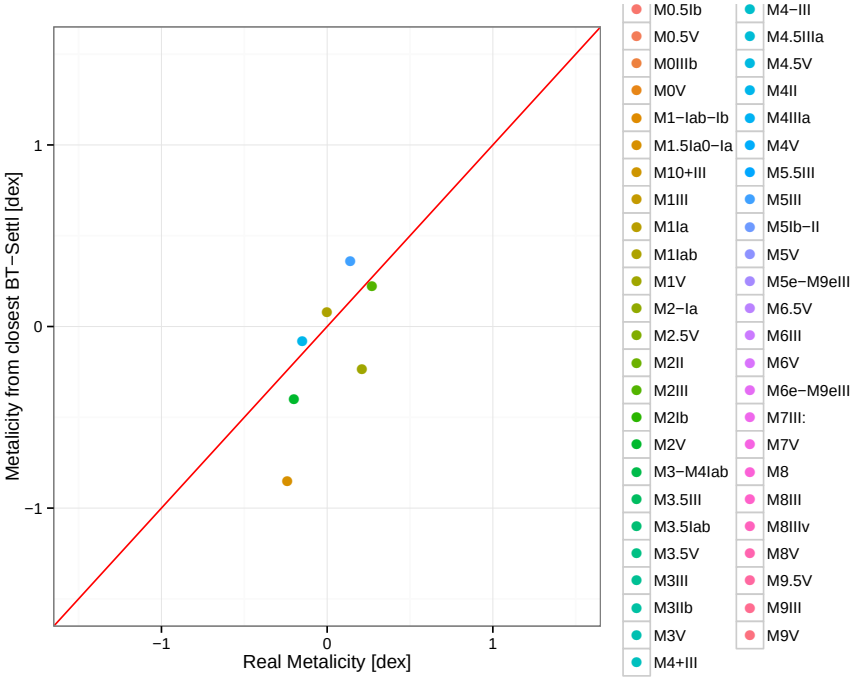


(b) Comparison between Gravity estimations from Spectral Subtype in x axis and the Support Vector Machines for Ga based features trained with BT_Settl at SNR=∞ and features for forecasting at SNR=∞ on y-axis

Fig. 4: 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. 5: Performance comparison between the *chi*^{2} based selection and the band oriented features to forecast Log(g)

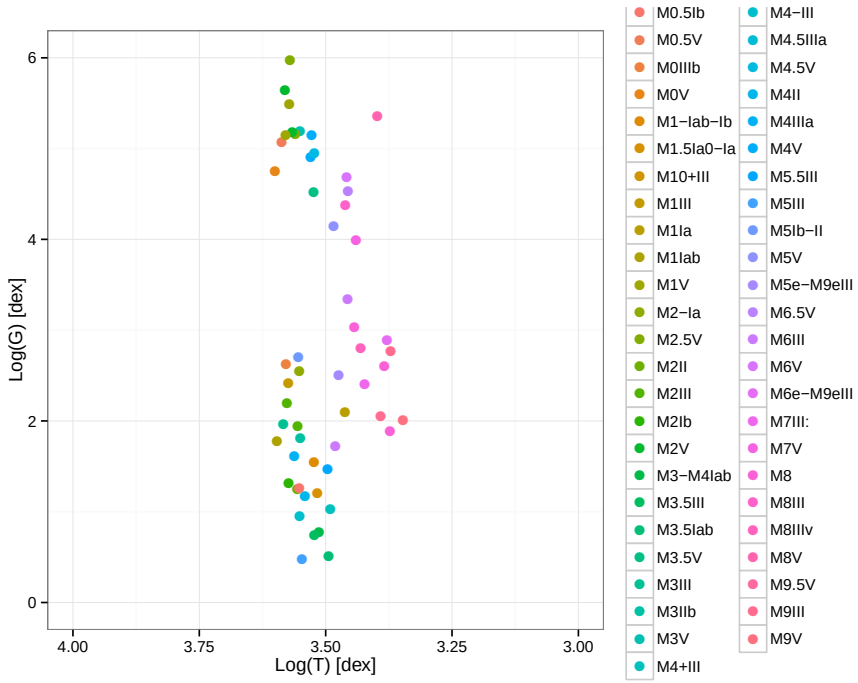


Fig. 6: Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for models based on bandpass features

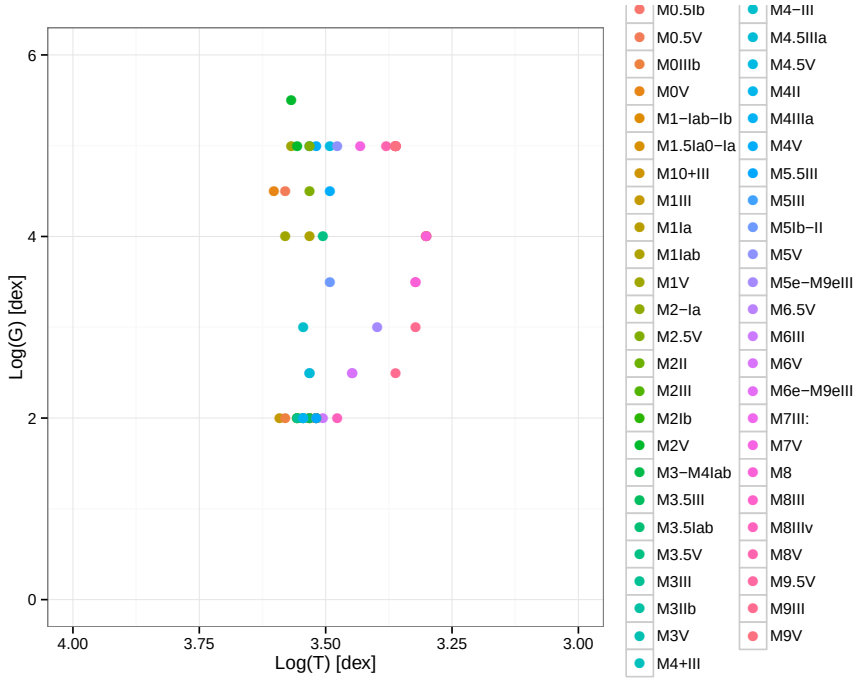


Fig. 7: Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=50 when the nearest BT-Settl spectrum is used.

5. Physical parameters foreseen for IRTF M-type stars

6. 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

Rayner, J. T., Cushing, M. C., & Vacca, W. D. 2009, ApJS, 185, 289

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

List of Objects