Astronomy & Astrophysics manuscript no. ipac_m ©ESO 2014
May 15, 2014

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

I. The I band perspective

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 May 15, 2014; accepted

ABSTRACT

Key words. class M stars – dynamic feature selection – physical parameter identification

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

1. Introduction

2. 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.

2.1. Dataset Selection.

The M dwarf database includes spectra for over 500 of the nearest and brightest M dwarfs in the same wavelength regime (roughly 6300-9000Å)(Kirkpatrick (2002)) with spectroscopic observa-

tions obtained at the Multiple Mirror Telescope (MMT, efective aperture 4.5m) on Mount Hopkins AZ and the McDonald Observatory 2.7 telescope on Mount Locke, TX. The registered specrta from the IPAC dataset has no uncertainty defined.

At the MMT spectra were obtained with a Red Channel Spectrograph equipped with an 800x800 TI CCD. A 270 line mm^{-1} grating with an LP-945 order blocking filter was used to cover the range 6300 - 9000Å at a resolution of 18Å. At McDonald, spectra covered the range 6400 - 9200Å at a resolution of 12Å(Henry et al. (1994)).

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.

2.2. 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.

Spectra degradation from the original 0.1Å stepsize till the required according to the IPAC resolution (3.6Å) was accomplished by a gaussian convolution with 100 steps and a standard deviation of 3.6Å too. Then, the spectra are trimmed to produce valid segments between 6000 and 9150Å. Indeed, in order to become independent of the star's distance normalization the area under spectrum has been performed to value of one. Total size of avaialble spectra is 429 (50K of distance each). As the total number of spectra is quite limited, we have decided to interpollate between them, according to a finner parameter's mesh.

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).

Now several other datasets have been created, by defininh 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.

The synthetic spectra are theoretical and they are noise free but it does not happen with real spectra, the IPAC 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.

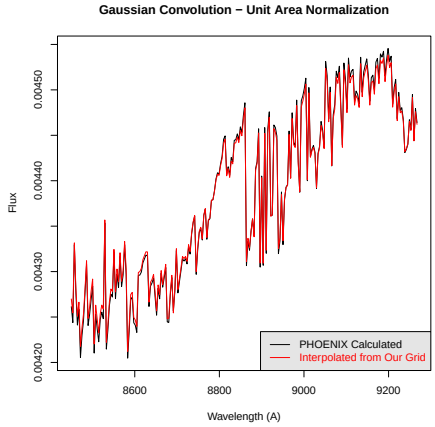

Fig. 1. Comparison between generated and interpollated spectrum

2.3. 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 are the continuum chunks, were no signal shall be expected.

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 and prodosed and even when the dataset is a different one, the main band I (0.80 - 0.90μm) is fully applicable in our research. Because of this their proposal for bands will be considered too in this research.

The underlying approach is to estimate the three star relevant physical parameters by establishing a set of features. 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).

$$F(i) = \int_{\lambda_1(i)}^{\lambda_2(i)} \left(1 - \frac{f(\lambda)}{F_{cont}(i)}\right) d\lambda \qquad \forall i \in [1 \dots N] \tag{1}$$

and

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

where N means the number of features to be selected and $f(x)$ denotes the normalized spectra of the star in the region of interest.

Now, the research question is how to identify $\{\lambda_1(i), \lambda_2(i), \lambda_{1c}(i), \lambda_{2c}(i)\} \quad \forall i \in [1 \dots N]$ in such a way they become useful to estimate the physical parameters. A few of minimal requirements need to be considered in the definition of those values like, for instance:

- $\diamond$ $\|\lambda_2(i) - \lambda_1(i)\| > 10\text{Å} \quad \forall i \in [1 \dots N]$.
- $\diamond$ $\|\lambda_{2c}(i) - \lambda_{1c}(i)\| > 20\text{Å} \quad \forall i \in [1 \dots N]$.

- $\overline{\lambda_2(i)\lambda_1(i)} \bigcap \overline{\lambda_{2c}(i)\lambda_{1c}(i)} = \emptyset \quad \forall i \in [1 \ldots N]$.

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

2.4. Determination of Features

To search for those features will depend on which specific phyisical 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 behaviour 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.

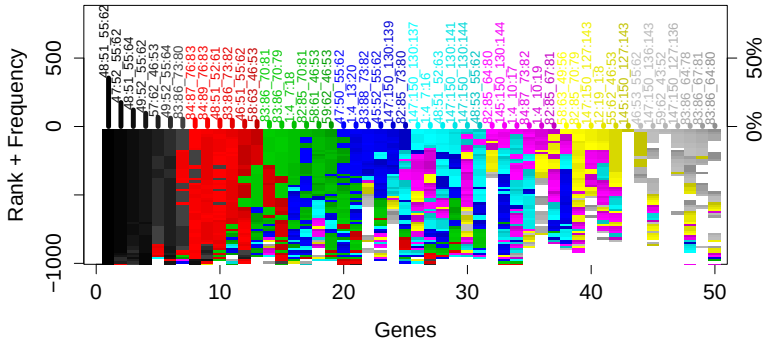

Fig. 2. Gene rank stability to predit T_{eff} with a NC fitness fucntion

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 four values for λ required to define a gene were coded for working in *band I* and they named according to the ordinal of the wavelength step. A fixed number of genes where agreed as memeber per chromosome and we have used five for this pourpose.

The same strategy was used for all the three physical parameters ($T_{eff}, log(g), met$). As fitness function two of them were evalauted. The first one was the Nearest Center (*NC*), defined as the point were the euclidean distance is minimum and it is a non parametric method, thus it is not required the data follows a normal distribution. The second one was the emsemble method known as Random Forest (Breiman (2001)) (*RF*), where fitness becomes the accuracy to predict the right classes for validation chromosomes. The RF fitness function is smarter but the required computational effort is much higher than NC.

The fitness function uses the classification accuracy of the model on the training samples to assign a score to each chromosome and make possible the selection of better predictors through the law of natural selection. There are different ways to estimate the error during training. The most obvious one is to further split the training set into training and validation sets.

The selected chromosomes have different persistence in population, but also their rank is relevant in order to establish the notion of stability. The Figure 2 summarizes that concept as provided by package Galgo in library R.

The most frequent fifty chromosomes were presented along the orizontal axis in eight colors, with six or seven genes per color. Upper vertical axis presents the gene frequency and the lower part of the vertical axis describes the colour code for that gene in previous epochs. This representation reflects both rank and persistence, wich can be seen as a metric for gene stability.

	Signal_from	Signal_To	Cont_From	Cont_To
Feature 1	8451.60	8458.80	8473.20	8509.20
Feature 2	8620.80	8628.00	8646.00	8667.60
Feature 3	8653.20	8667.60	8613.60	8635.20
Feature 4	8746.80	8754.00	8710.80	8739.60
Feature 5	8977.20	8984.40	8916.00	8937.60

Table 1. Recommended features and Continuum bandpass for predicting T_{eff} by using NC fitness function

	Signal_from	Signal_To	Cont1_From	Cont1_To	Cont2_From	Cont2_To
Pa1	8462.40	8473.20	8476.80	8484.00	8563.20	8574.00
Ca1	8484.00	8512.80	8476.80	8484.00	8563.20	8574.00
Ca2	8523.60	8559.60	8476.80	8484.00	8563.20	8574.00
Pa2	8577.60	8617.20	8563.20	8574.00	8620.80	8638.80
Ca3	8642.40	8682.00	8620.80	8638.80	8700.00	8721.60
Pa3	8732.40	8772.00	8700.00	8721.60	8779.20	8790.00
Mg	8804.40	8808.00	8779.20	8790.00	8815.20	8847.60
Pa4	8851.20	8887.20	8815.20	8847.60	8890.80	8898.00

Table 2. Recommended features and Continuum bandpass recommended in Cesetti et al. (2013) as relevant for temperature

	Signal_from	Signal_To	Cont_From	Cont_To
Feature 1	8462.40	8476.80	8484.00	8520.00
Feature 2	8620.80	8628.00	8667.60	8703.60
Feature 3	8646.00	8653.20	8689.20	8710.80
Feature 4	8656.80	8664.00	8689.20	8718.00
Feature 5	8782.80	8790.00	8797.20	8818.80

Table 3. Recommended features and Continuum bandpass for predicting $log(g)$ by using NC fitness function

The GA procedure provides us with a large collection of chromosomes. Although these are all good 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 that is, to some extent, representative of the population. The simpler strategy to follow is to use the frequency of genes in the population of chromosomes as criteria for inclusion in a forward selection strategy. The model of choice will be the one with the highest classification accuracy and the lower number of genes.

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

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. They have provided, inside the range of the IPAC dataset, the bandpass presented in Table 2

In regards with the Gravity, the GA recommends the features presentend in Table 3

Finally, features suggested for metallicity can be found in Table 4.

	Signal_from	Signal_To	Cont_From	Cont_To
Feature 1	8516.40	8538.00	8548.80	8577.60
Feature 2	8620.80	8628.00	8635.20	8671.20
Feature 3	8782.80	8790.00	8797.20	8818.80
Feature 4	8970.00	8977.20	8916.00	8952.00
Feature 5	9009.60	9016.80	8980.80	9002.40

Table 4. Recommended features and Continuum bandpass for predicting *Metallicity* by using NC fitness function

3. Regression models for parameter estimation

After selecting the convenient 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. In order to do that, the noised, theoretical BT-Settl library was used again by splitting-out 33% as for testing and using the remaining 66% for training. In order to produce reslient models, a ten folders crossvalidation without repetition was adopted, where the best model in the training folders was retained as the most convenient.

3.1. Models considered.

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 Linear Models (GLM).
- Generalized Additive Models (GAM).
- Project Pursuit Regression Method (PPR).
- Ridge Regression Model (RRM).
- Kernel Partial Least Square Method (KPLS).
- Multiadaptative Spline Regression Models (MARS).
- Random Forest Regression Models (RF).
- Gradient Boosting with Regression Trees (GBR).
- Generalized Boosted Regression Models (GBM).
- Support Vector Machine Models with Gaussian Kernel (SVM).
- MLP Nerural Networks (NNET).

We have considered, in addition to the previous models, two ensembles techniques among the already built models, one linear and another one Greedy; all these models were considered in order to produce the most suitable model to be brought as predictor.

Comparison of performance between sets of features for temperature can be analyzed, over the same testing subdataset and it was depicted in Table 5.

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.

SNR	Features	SVM	RF	GLM	MARS	Greedy
50	Cesetti et al.	81.6	83.3	163.5	91.9	71.4
	GA	91.4	82.2	161.1	91.9	77.3
10	Cesetti et al.	135.8	138.5	268.8	166.8	125.9
	GA	123.2	122.6	212.6	130.9	116.5

Table 5. RSME for different models predicting T_{eff} [K].

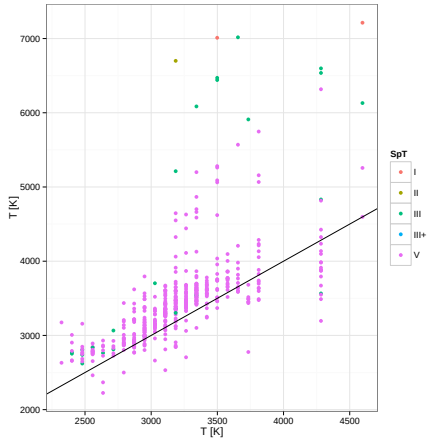

Fig. 3. Comparison between Temperature estimations from Spectral Subtype in x axis and the Random Forest for Ga based features trained with BT-Settl at SNR=50 on y-axis

3.2. 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 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.

3.3. Temperature model based on Spectral Subtyoe.

3.4. Temperature Model for T_{eff}.

After training the set of models by using labelled BT-Settl dataset, those models were used to predict the IPAC temperature. In Figure 3 the relationship between Temperature estimated from GA proposed features with SNR=50 and the Temperature estimation from spectral subtype.

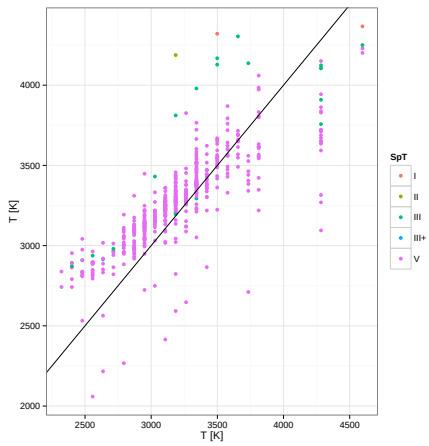

Fig. 4. Comparison between Temperature estimations from Spectral Subtype in x axis and the Random Forest for Ga based features trained with BT-Settl at SNR=10 on y-axis

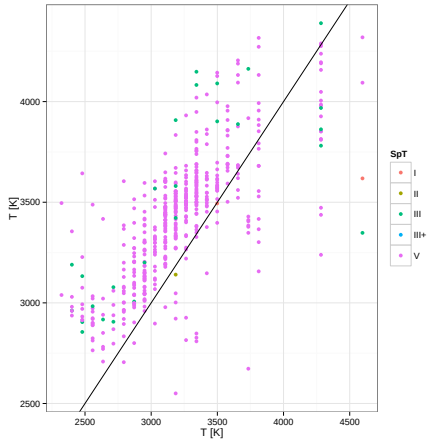

Fig. 5. Comparison between Temperature estimations from Spectral Subtype in x axis and the Random Forest for Cesetti et al. features trained with BT-Settl at SNR=50 on y-axis

Similarly Figure 4 shows the relationship against GA features and Random Forest model trained by BT-Settl at SNR=10.

The same was made with features forposed by Cesetti et al. (see Figure 5 and Figure 6).

The analysis was done for Global spectrum based approach. Thus in Figure 9 and Figure **??** presents the relationship for the diemensional reduction and SVM approach and Figure **??** and Figure **??** accounts for similarity based estimation of pysical parameters accoding to χ^2 metric.

Te same approach can become useful to produce $log(gg)$ estimations. Here comparisons can only be possible between GA based features and global spectra based approach.

In Figure 11 and Figure 12 relationships between $log(g)$ predicted by global espectrum estimation and GA feature based estimation can be observed. Additionally Figure 13 and Figure 14 present the relationship between the GA based estimation and the χ^2 nearest BT-Settl spectrum.

Finally, the same analysis is performed for the Metallicty parameter

It is possible to present relationships between $log(g)$ and $log(T_{eff})$ as a matter of congruence analysis between predictions.

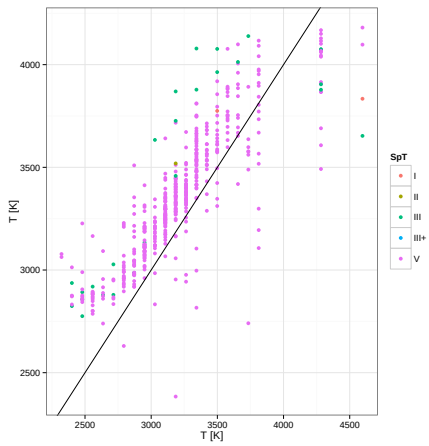

Fig. 6. Comparison between Temperature estimations from Spectral Subtype in x axis and the Random Forest for Cesetti et al. features trained with BT-Settl at SNR=10 on y-axis

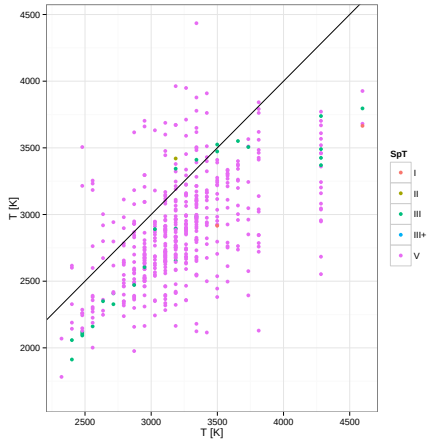

Fig. 7. Comparison between Temperature estimations from Spectral Subtype in x axis and the Random Forest for full length spectra trained with BT-Settl at SNR=50 on y-axis

The same can be performed when the estimations arise from models based on features provided by the GA technique.

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.

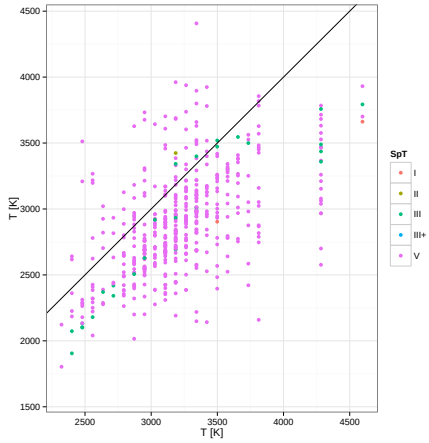

Fig. 8. Comparison between Temperature estimations from Spectral Subtype in x axis and the Random Forest for full length spectra trained with BT-Settl at SNR=10 on y-axis

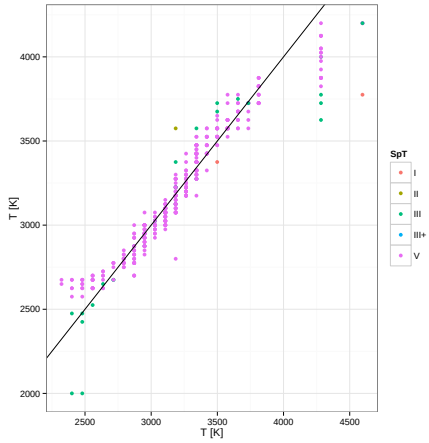

Fig. 9. Comparison between Temperature estimations from Spectral Subtype in x axis and the closest BT-Settl spectrum with SNR=50 on y-axis

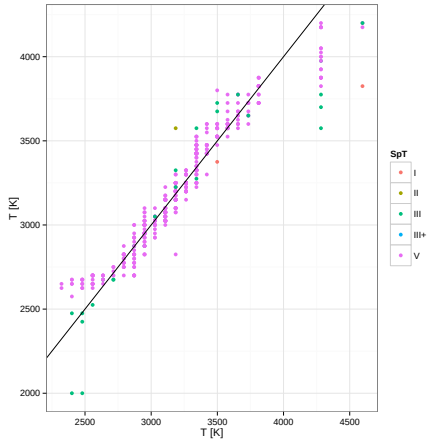

Fig. 10. Comparison between Temperature estimations from Spectral Subtype in x axis and the closest BT-Settl spectrum with SNR=10 on y-axis

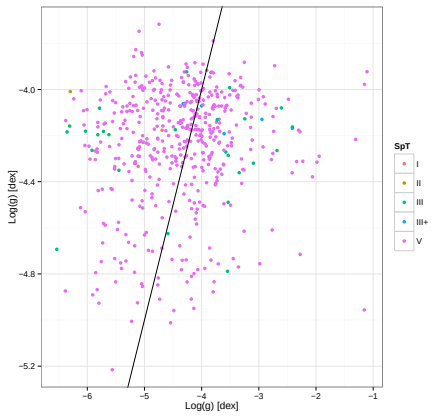

Fig. 11. Comparison between $log(g)$ estimations from Random Forest model using GA features in x axis and the Global projected spectrum (ICA+SVM) with SNR=50 on y-axis

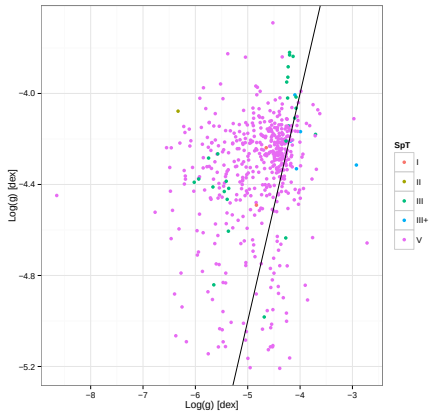

Fig. 12. Comparison between $log(g)$ estimations from Random Forest model using GA features in x axis and the Global projected spectrum (ICA+SVM) with SNR=10 on y-axis

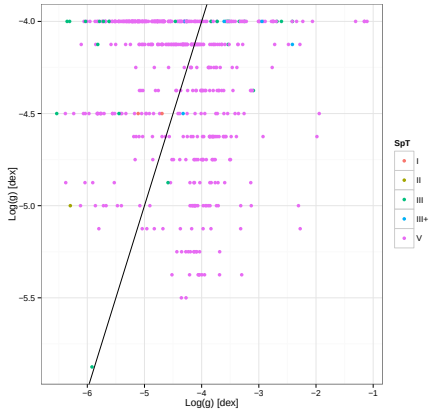

Fig. 13. Comparison between $log(g)$ estimations from Random Forest model using GA features in x axis and the nearest χ^2 BT-Settl spectrum with SNR=10 on y-axis

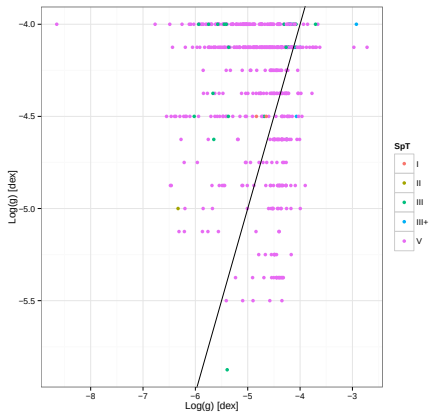

Fig. 14. Comparison between $log(g)$ estimations from Random Forest model using GA features in x axis and the nearest χ^2 BT-Settl spectrum with SNR=10 on y-axis

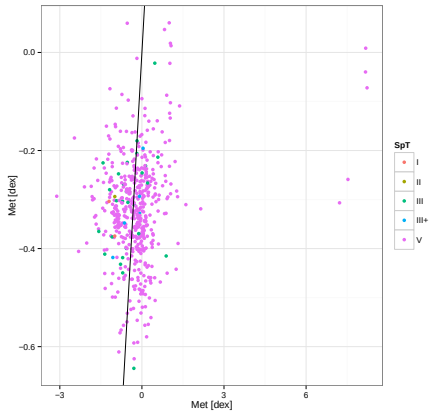

Fig. 15. Comparison between Metallicity estimations from Random Forest model using GA features in x axis and the Global projected spectrum (ICA+SVM) with SNR=50 on y-axis

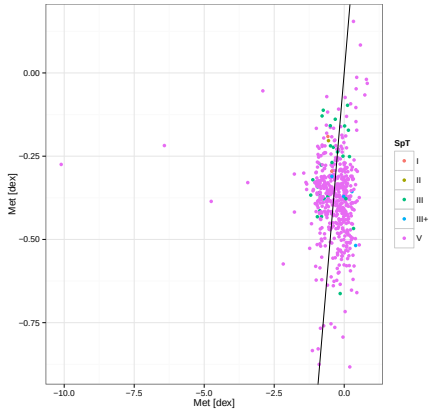

Fig. 16. Comparison between Metallicity estimations from Random Forest model using GA features in x axis and the Global projected spectrum (ICA+SVM) with SNR=10 on y-axis

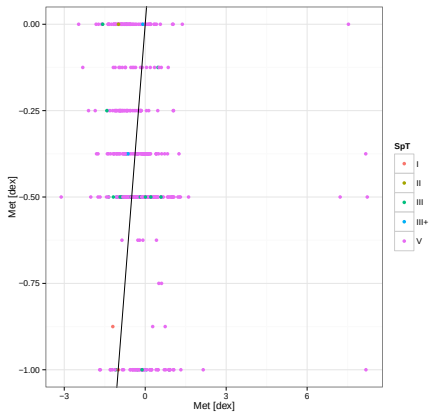

Fig. 17. Comparison between Metallicity estimations from Random Forest model using GA features in x axis and the nearest χ^2 BT-Settl spectrum with SNR=10 on y-axis

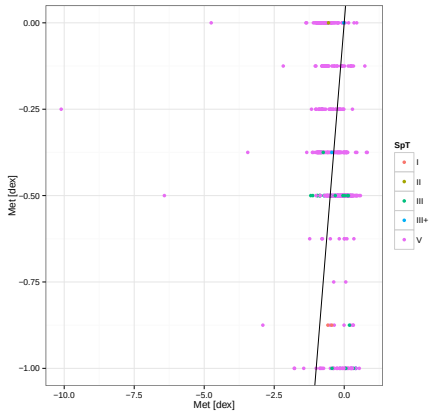

Fig. 18. Comparison between Metallicity estimations from Random Forest model using GA features in x axis and the nearest χ^2 BT-Settl spectrum with SNR=10 on y-axis

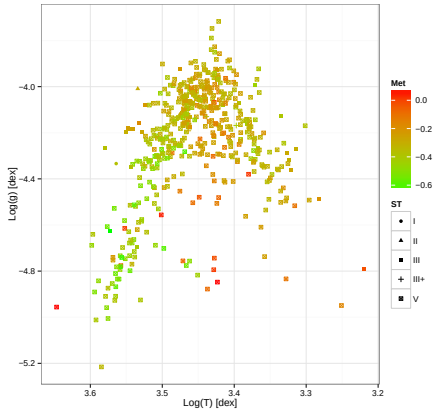

Fig. 19. Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=50 when Global ICA+SVM model is used

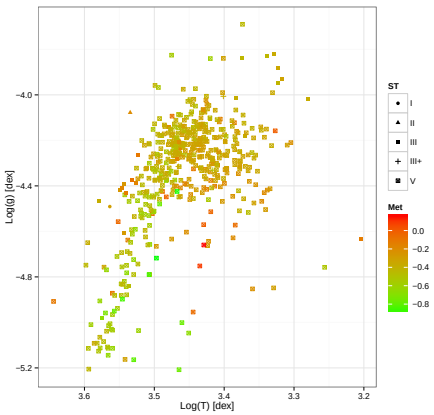

Fig. 20. Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=10 when Global ICA+SVM model is used

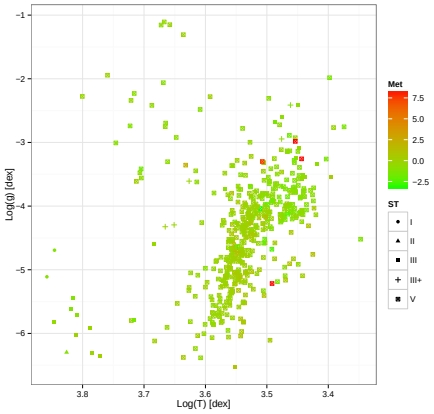

Fig. 21. Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=50 when the Random Forest model over the GA provided features is used

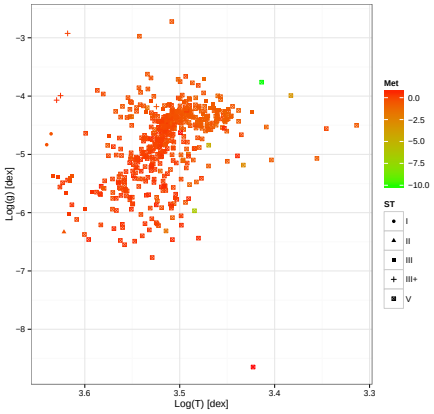

Fig. 22. Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=10 when the Random Forest model over the GA provided features is used

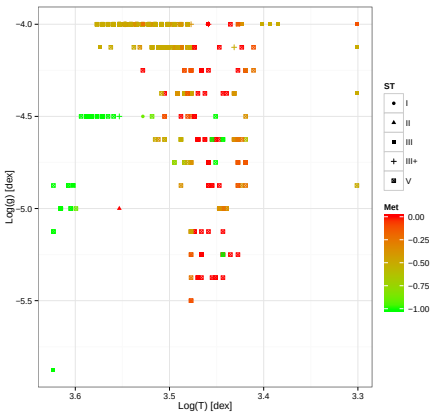

Fig. 23. Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=50 when the nearest BT-Settl spectrum is used

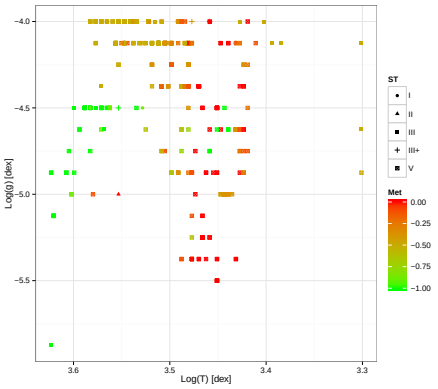

Fig. 24. Relationship between $log(T_{eff})$ in the x axis and $log(g)$ in the y axis for SNR=10 when the nearest BT-Settl spectrum is used

In the end deeper analysis needs to be carried out considering several factors like coherence between labeled referenced library spectrum, density of the labeled spectra, etc.

4. Physical parameters for the sample of IPAC M-type stars

5. Conclusions

Acknowledgements. This research has benefitted from the M, L, T, and Y dwarf compendium housed at DwarfArchives.org.

References

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

Breiman, L. 2001, Machine learning, 45, 5

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)

Henry, T. J., Kirkpatrick, J. D., & Simons, D. A. 1994, AJ, 108, 1437

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

Kirkpatrick, J. D. 2002

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