Highlights of Spanish Astrophysics XII, Proceedings of the XVI Scientific Meeting of the Spanish Astronomical Society held on July 15 - 19, 2024, in Granada, Spain. M. Manteiga, F. González Galindo, A. Labiano Ortega, M. Martínez González, N. Rea, M. Romero Gómez, A. Ulla Miguel, G. Yepes, C. Rodríguez López, A. Gómez García and C. Dafonte (eds.), 2025

Estimating stellar parameters from photometry using ML techniques

Aguilar, J. F.[1,2], **Cruz, P.**[3], **and Solano, E.**[3]

[1] Departamento de Matemáticas, Universidad Militar Nueva Granada, kilómetro 2 vía Cajicá - Zipaquirá, Colombia, código postal 110111.
e-mail: john.aguilar@unimilitar.edu.co
[2] PhD Programme in Astrophysics, Doctoral School, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049, Madrid, Spain.
[3] Centro de Astrobiología (CAB), CSIC-INTA, Camino Bajo del Castillo s/n, E-28692, Villanueva de la Cañada, Madrid, Spain

Abstract

This work is focused on the estimate of stellar parameters (effective temperature, surface gravity and metallicity) using photometric data. We used a machine learning method known as the K-means to calculate stellar parameters from 2MASS and JPLUS DR3 photometry. The calculation of these parameters was done with 105 colors constructed from the bands under study. For that, we adopted as training data the synthetic photometry from MESA Isochrones & Stellar Tracks (MIST), and F- to M-type stars with known stellar parameters collected from Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) that also have photometric information in the above-mentioned surveys However, when obtaining these values of the stellar parameters under this routine, it was found that the computational cost was considerable. We then used the principal component analysis (PCA) to find the minimum set of colors that could reproduce the analysis, obtaining a set of 11 colors. The methodology was applied to over 5 million J-PLUS DR3 stars, those with 2MASS photometry available. We present comparisons of our results with LAMOST and APOGEE, showing that the obtained parameters are in good agreement. Our results are also compared with previous results based on J-PLUS DR1 photometry. We further discuss the scope and limitations of estimating the parameters from diferentes quality photometry.

1 Introduction

In recent decades, the massive growth of astronomical data has generated a series of challenges in the management and analysis of such large databases. Among the large ground-based photometric surveys, the Two-micron All Sky Survey (2MASS, [9]) and the Javalambre

Photometric Local Universe Survey (J-PLUS, [3]) has been reporting millions of photometric data in the optical and at near-infrared wavelengths that show high potential for estimating stellar parameters.

In this research, we aim to derive stellar atmospheric parameters, such as effective temperature ($T_{\rm eff}$), surface gravity ($\log g$), and metallicity ([Fe/H]), using only the available photometry from J-PLUS and 2MASS. The current work uses a modified version of the algorithm adopted by [6] and [5], which is based on machine learning (ML) techniques.

2 Sample

We collected 6,246,452 stars with photometric data from both 2MASS and JPLUS surveys. From these data, we defined as high photometric quality (HQP) objects those that fulfilled the following criteria: a 2MASS `FLAG` of "AAA", J-PLUS `FLAG` and `MASKFLAG` equal to 0, r-band < 22 mag, and errors in all bands of less than or equal to 5%. Those objects that do not meet all these criteria are named as non-HQP.

The comparison data, the sample used as comparison for the estimation of stellar parameters, is composed by both observational and synthetic data. In the case of observational data, we performed a cross match between the sources collected from 2MASS and J-PLUS surveys and the fifth data release from the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST DR5, [1]). We selected 56,131 objects whose spectra presented a signal-to-noise ratio SNR > 10. In addition, we gathered 111,879 models with synthetic photometry available in the mentioned bands from the MESA Isochrones and Stellar Tracks (MIST) project[1].

To validate the results obtained with the implementation of the proposed methodology, we selected a separate sample, not included in the comparison sample, with stellar parameters already derived from spectroscopy to compare the results from our method. The contrast sample contains 432,319 objects with spectroscopic stellar parameter derived with an error of less than 5%.

3 Machine learning algorithms

To distribute a data set into subgroups, one of the most relevant parameters is the definition of the number of clusters. This number is often taken according to the nature of the data, taking into account the previous knowledge in the related area. However, this number is not always known and a method that suggests the optimal number of groups into which the sample should be divided is necessary ([2]; [4]).

The Hartigan test ([10]) is defined as the logarithm of the ratio between the sum of squares between cluster (SSW) and the sum of squares within a cluster (SSB), where:

[1] `https://waps.cfa.harvard.edu/MIST/`.

$$H = -log\left(\frac{SSW}{SSB}\right), \tag{1}$$

$$SSW = \frac{1}{n}\sum_{i=1}^{m}\sum_{j\in C_i}\left\|x_j - C_{P(j)}\right\| \tag{2}$$

$$SSB = \frac{1}{n}\sum_{i=1}^{m} n_i \left\|C_i - \bar{x}\right\|, \tag{3}$$

with C being the number of test partitions in the index, m indicating the cluster number, n_i the number of elements in each cluster, x_j representing an element of the dataset, and $\bar{x}$ the mean of these data. The use of this algorithm allows to find the minimum difference in the determination of an optimal number of clusters, in which $H \leq 10$.

K-means is a clustering method that allows grouping the data set $X = \{x_1, x_2, ..., x_n\}$ within k groups. To achieve this, the algorithm performs three main steps:

1. Randomly assign k centroids, $V = \{v_1, v_2, ..., v_k\}$ ([7]).

2. Once these initial centroids have been assigned, each sample point is assigned to a cluster using the closest k centroid as the criterion.

3. New centroids are defined from the mean value of the coordinates of all the points that are part of each cluster, using the expression:

$$V_i = \left(\frac{1}{C_i}[i = 1]C_i\sum x_i\right), \tag{4}$$

   with x_i representing the points in the ith cluster.

Once the new centroids are defined, the procedure described in steps 2 and 3 is repeated until the centroids converge (which means that the distance between the step centroid m and $m-1$ tends to zero). These iterations are a good indicator of the efficiency of the method, since, if the convergence occurs with a low number of iterations, it is a good indicator of the reliability of the suggested groupings.

To implement this algorithm it is important to define aspects such as the metric used and the number of clusters (K-means is a non-hierarchical method cluster analysis). Regarding the first aspect, the metric to be used is the Euclidean distance, which is defined as:

$$\|AB\| = \left((a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2\right)^{\frac{1}{2}}, \tag{5}$$

with $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ as coordinates of A and B in the space $\mathbb{R}^n$.

4 Methodology

To estimate the stellar parameters, we have built an algorithm that uses the information of 105 colors constructed from the combination of all 12 J-PLUS filters (u, $J0378$, $J0395$, $J0410$, $J0430$, g, $J0515$, r, $J0660$, i, $J0861$, and z) plus 3 2MASS bands (J, H, and K_s).

In the algorithm, to define clusters of objects with similar stellar parameters, we use the Hartigan test to construct an optimal number of groups in terms of the data distribution. Once this number is obtained, we use the K-means algorithm to determine the stars that will be part of each of these groups. With these groups formed, an analysis of the distances of these objects to the cluster's pseudo-center is performed. These distances are called radius, although they are not spherical and are measured in the $\mathbb{R}^{105}$ space.

From the minimum, 5th, 25th, 75th, 95th and 99th percentiles, mean, median, mode, and maximum statistics, 10 different neighborhood radii are estimated, with 1 being the smallest radius and 10 the largest radius, that is, the one farthest from the cluster pseudo-center. As a last step, each of the studied objects is taken and the closest comparison data are checked, testing the smallest comparison neighborhood. From them, we estimated each stellar parameter as the average of the stellar parameters of the comparison objects contained in the neighborhood and the error of this parameter as their standard deviation.

5 Results

By implementing this method, stellar parameters for 5,689,987 stars were obtained. The analyzed sample is composed by F, G, and K-type stars with $T_{\rm eff}$ between 3,184 and 8,390 K, $\log g$ between 0.29 and 4.77 and [Fe/H] between -2.62 and 0.60 dex. These objects were represented in a Kiel diagram (Fig. 1), showing the HQP and non-HQP samples separately, in order to validate whether the results obtained showed a behavior with similar characteristics to those of the comparison data (both observational and synthetic data).

Figure 1, constructed with TOPCAT[2] ([8, 11]), shows a greater dispersion in the parameter values obtained for larger neighborhoods (greater than 3). These parameters estimated from larger cluster radii, although they make a first estimate of what the stellar parameter value could be, they carry larger errors compared to those obtained within smaller neighborhoods.

We compared the stellar parameters obtained with our method and those reported by the LAMOST survey for those stars composing the contrast sample. Table 1 shows the median absolute deviation (MAD) of these differences. There, the data are reported separated into two groups, data with high quality photometry (HQP) and those with non-high quality (non-HQP).

Stellar parameters calculated with our method showed a high level of coincidence with those obtained by spectroscopic measurements. On the other hand, the generation of different quality samples provided an important opportunity to calculate stellar parameters for most of the data under study. We propose as advances in our research the implementation of this

[2]Tool for OPerations on Catalogues And Tables, [8, 11], available at `http://www.star.bris.ac.uk/mbt/topcat/`.

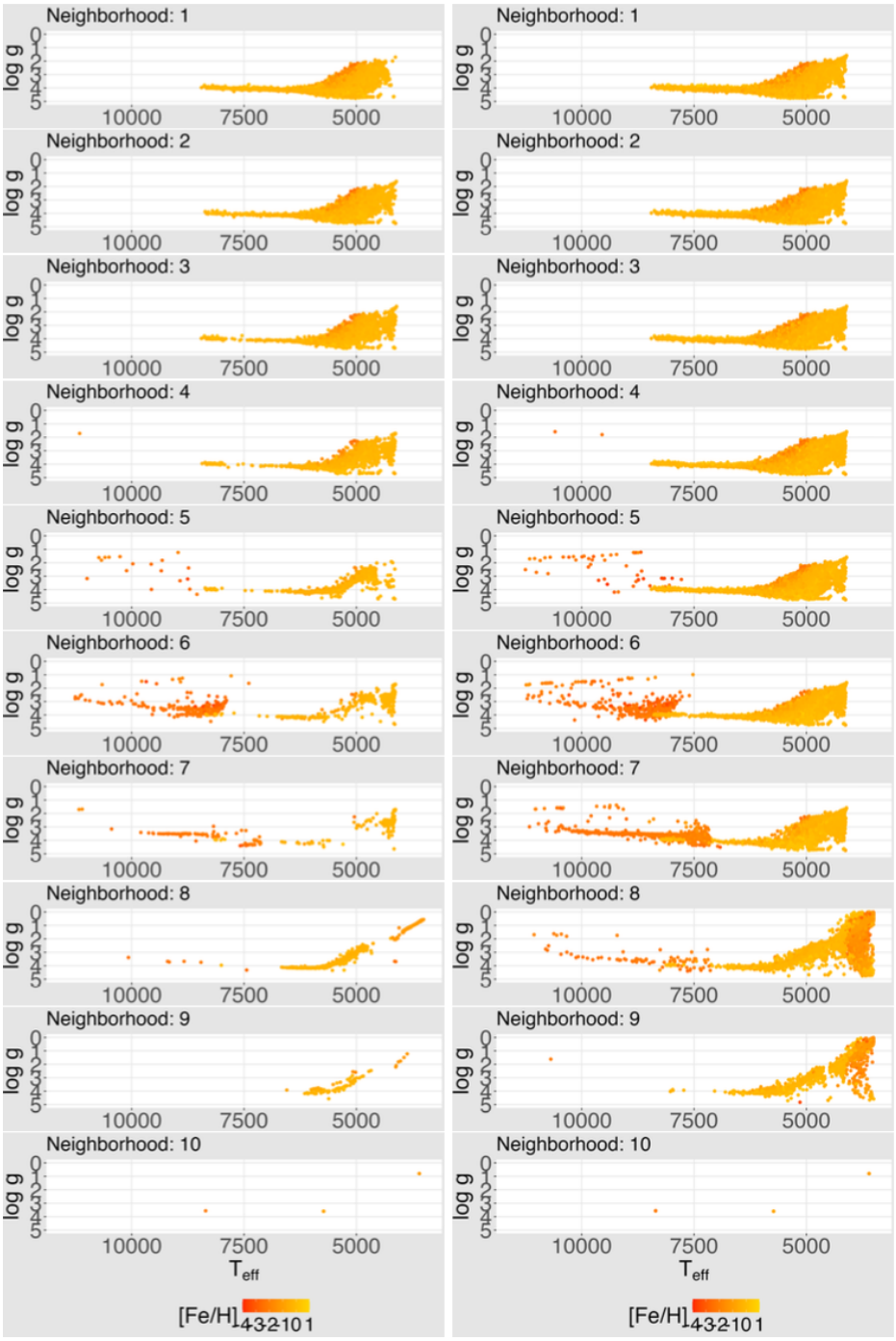

Figure 1: Kiel diagram. Each row represents the estimated stellar parameter for the analyzed sample present in each one of the 10 proposed neighborhoods. The left column presents the HQP objects and the right column presents the non-HQP ones.

method with a subset of the colors used and other possible modifications to the algorithm generated in this work.

	Median Absolute Deviation (MAD)					
	Δ Teff		Δ log g		Δ [Fe/H]	
	Best estimates	Other estimates	Best estimates	Other estimates	Best estimates	Other estimates
HQP	95	370	0.131	0.262	0.164	0.435
Non-HQP	89	284	0.143	0.392	0.148	0.261

Table 1: Median absolute deviation between the stellar parameters calculated with our method and the values obtained from LAMOST for the stars in the contrast sample. The best estimates correspond to objects for which stellar parameters were calculated within neighborhood radius smaller than 4. The other estimates column corresponds to larger neighborhoods.

Acknowledgments

J.A.'s contribution to this work is a product of his academic exercise as a professor at the Universidad Militar Nueva Granada, Bogotá, Colombia. P.C. and E.S. acknowledge financial support from the Spanish Virtual Observatory project funded by the Spanish Ministry of Science and Innovation/State Agency of Research MCIN/AEI/10.13039/501100011033 through grant PID2020-112949GB-I00.

References

[1] Bai, Z.-R., Zhang, H.-T., Yuan, H.-L., et al. 2021, *Research in Astronomy and Astrophysics*, 21, 249.

[2] Bishop, C. M. 2016, "Pattern Recognition and Machine Learning", *Springer*.

[3] Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2019, *A&A*, 622, A176.

[4] Chattopadhyay, A., and Chattopadhyay, T. 2014, "Statistical Methods for Astronomical Data Analysis".

[5] Cruz, P., Aguilar, J. F., Garrido, H. E., et al. 2022, *MNRAS*, 515, 1416.

[6] Garrido, H. E., Cruz, P., Diaz, M. P., and Aguilar, J. F. 2019, *MNRAS*, 482, 5379.

[7] Le Cam, L. M., and Neyman, J. 1967, "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Statistics", *Univ of California Press*.

[8] Shopbell, P., et al. 2005, *Astronomical Society of the Pacific Conference Series*, 735.

[9] Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, 131, 1163.

[10] Strauss, D. J., and Hartigan, J. A. 1975, "Clustering Algorithms", *Biometrics*, 31, 793.

[11] Taylor, M. 2011, "TOPCAT: Tool for OPerations on Catalogues And Tables", *Astrophysics Source Code Library* (ascl:1101.010).