Machine learning for stellar parametrization

Presenter: Alessio Turchi

Alessio Turchi[1] Elena Pancino[1,2] Fabio Rossi[1] Maria Tsantaki[1] Aleksandra Avdeeva[3] Paola Marrese[2,4] Silvia Marinoni[2,4] Nicoletta Sanna[1] Giorgio Fanari[2]

[1]INAF-Osservatorio Astrofisico di Arcetri, L.go Enrico Fermi 5, Firenze, Italy

[2]Space Science Data Center, Via del Politecnico SNC, I-00133 Rome, Italy

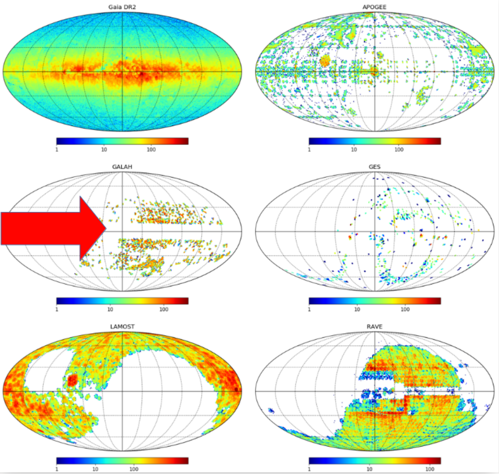
[3]Institute of Astronomy, Russian Academy of Sciences, 48 Pyatnitskaya St., Moscow 119017, Russia

[4]INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00040, Monte Porzio Cato

Monday 8th July, 2024

The context

There are billions of stars out there, distributed over lots of different catalogues.



The context

Studying the formation and evolution of our Galaxy is hampered by the heterogeneity of instruments, selection functions, analysis methods, and measured quantities.

The Survey of Surveys (SoS) is an effort to homogenize star parameters from different surveys (APOGEE, GALAH, Gaia-ESO, RAVE and LAMOST)

- SoS started the first release with radial velocities, however a preliminary sample with other stellar parameters such as effective temperature (Teff), surface gravity ($\log g$) and iron metallicity ([Fe/H]) was prepared (SoS-spectro).
- SoS-spectro contains around 11 million stars with spectroscopic-level quality parameters.
- Focus is to remove trends and biases between the different catalogues.
- More info here: *Survey of Surveys. I. The largest compilation of radial velocities for the Galaxy, Tsantaki, M., et. al., A&A, 659, A95, 2022*

The context

- Deriving Teff, $\log g$ and [Fe/H] from spectra is relatively accurate (giving some reasonable assuption such as LTE).
- Typical uncertainties of SoS-Spectro are of the order of $\lesssim$100 K in $\mathrm{T_{eff}}$ and of $\simeq$0.1 dex in $\log g$ and [Fe/H].

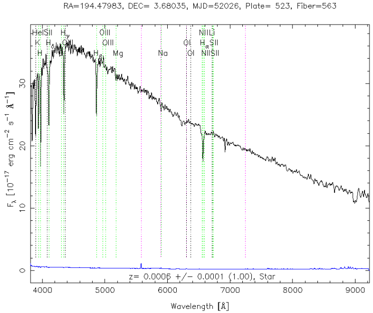
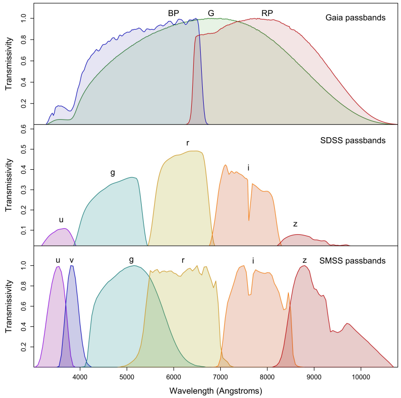


Figure: From Sloan Digital Sky Survey

The context

However most stars in the galaxy have only photometric measurements, often in just few bands

- Obtaining reliable estimates from photometric measurements is harder and involves fitting theoretical curves.
- Typical uncertainties are typically double (or more) with respect to spectroscopic measurements, especially for $\log g$ and [Fe/H].
  For more info see: *Directly Deriving Parameters from SDSS Photometric Images, Wu, Fan, et. al., AJ, 166, 3, 2023*



The Problem

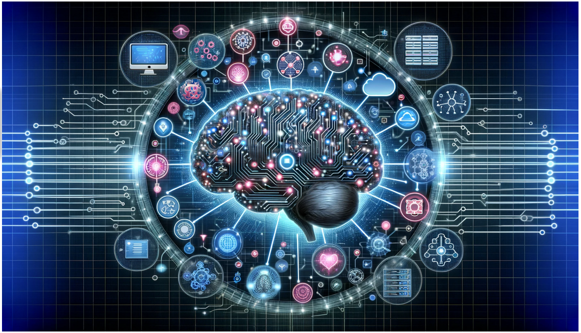
Question

Using multiple catalogs and information, can we use some method to enhance photometric estimates for Teff, $\log g$ and [Fe/H] using additional data?

Answer

Enters Machine Learning!

- Forgive me, but 80% of the work is in the data, not in the ML code... definitely not magic!



The Problem

Starting point

The idea is to start from Gaia DR3 estimates of Teff, $\log g$ and [Fe/H] , (one of the largest collection available), which are computed from low resolution BP-RP spectra and fitted on isocrones. More info here: Gaia Reference Documentation

- We join other photometric catalogues (e.g. SKYMAPPER or SDSS) by cross-matching and add additional parameters and magnitudes in different bands.
- We use the SoS spectroscopic estimates for the above parameters as a reference (true) value.
- We let the Machine Learning algorithm learn how to enhance the parameters by using the additional catalogues.

The Data

This is the comparison between SoS parameters and PASTEL, which is a compilation of high resolution R>25000, high signal-to-noise (SN>50) spectra (Soubiran, C., et. al., 2016).

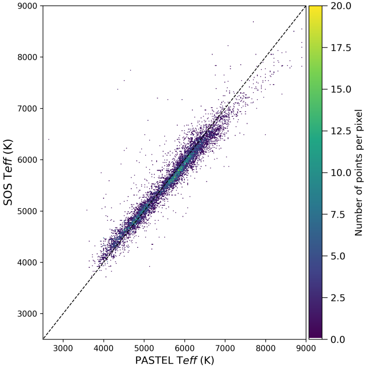


Figure: PASTEL Teff vs SoS

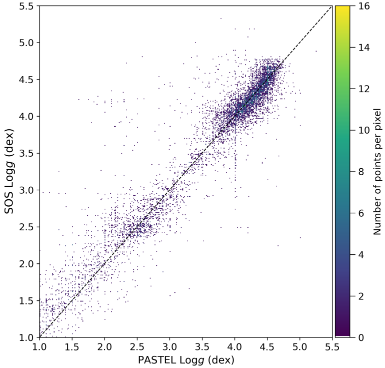


Figure: PASTEL Log g vs SoS

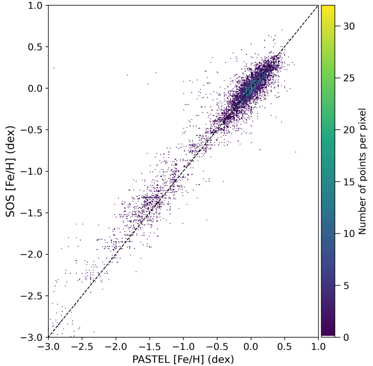


Figure: PASTEL Fe/H vs SoS

The Data

Comparison between Gaia photometry-derived parameters and SoS-spectro:

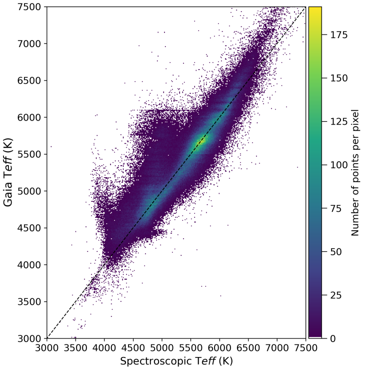


Figure: Gaia Teff vs SoS

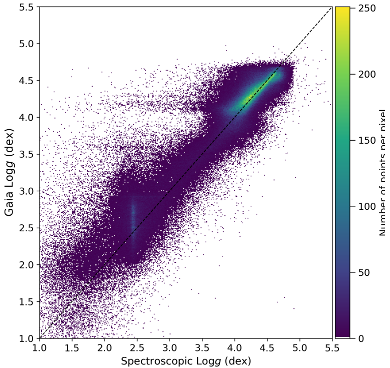


Figure: Gaia Log g vs SoS

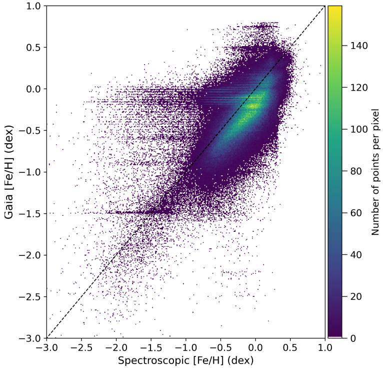
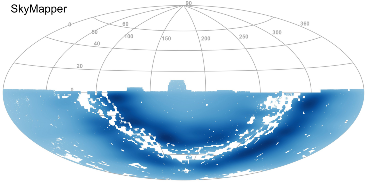


Figure: Gaia Fe/H vs SoS

The Data

We built our initial Dataset by cross-matching Gaia DR3 and SKYMAPPER, with distances taken from *Bailer-Jones et al. 2021*. We obtained a dataset of 11.4 million stars, spatially distributed in the southern hemisphere.

- We filtered out bad stars on various criteria (see SPIE paper). Basically we limit to stars with values in both Bp and Rp mags in Gaia and G mag <=18. We removed stars with more than two missing mags in the *uvgriz* SKYMAPPER bands. Lastly we removed stars with no distance measurement.
- We finally cross-matched with SoS-spectro leaving 624410 stars



The Model

- First let's introduce the restrictions: it has to run on this! (Yes, that's my desktop PC)

Figure: Core i7-10700 with 64Gb ram

Of course we bought a big server with GPUs, but for this preliminary work we were short on hardware ... :-)

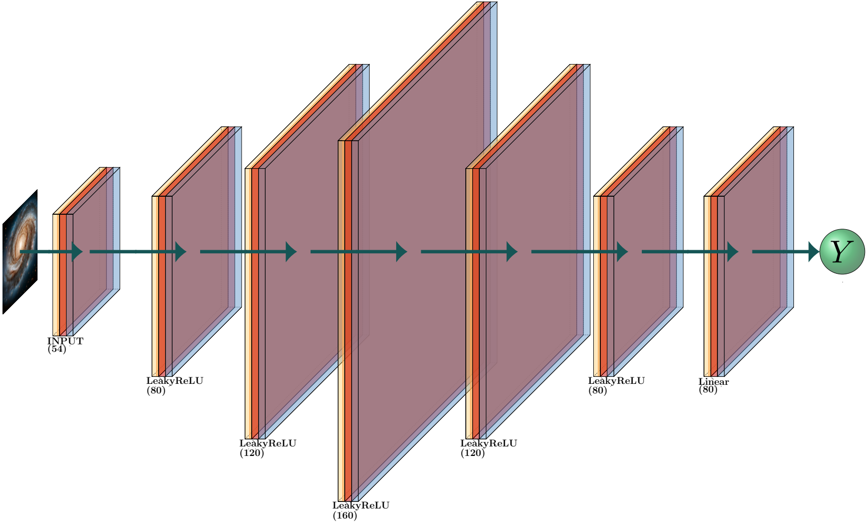
The Model

We train the model on 80% of our dataset (randomly selected), leaving 20% for validation and testing.

- We had some preliminary test with simple SVR on a small ($\sim$ 10K stars) with promising results (*Pancino E. et. al., A&A 664, A109, 2022)*
- We developed a Multilayer Perceptron model with Tensorflow library. Made it so that it can cope with NaNs in the columns (by flagging them) and trained a different model for each parameter (Teff, $\log g$ and [Fe/H])
- We repeatedly checked that we got no overfitting, and results were reasonably repeatable with different trainings (within a MAD below 5K for Teff and below 0.01 dex for $\log g$ and [Fe/H] over 100 realizations)

The Model

- Simplicity is the key to success!
- Each layer is a Dense layer with Batch-Normalization and Dropout at 0.02 rate
- One single scalar as an output



The INPUT Parameters

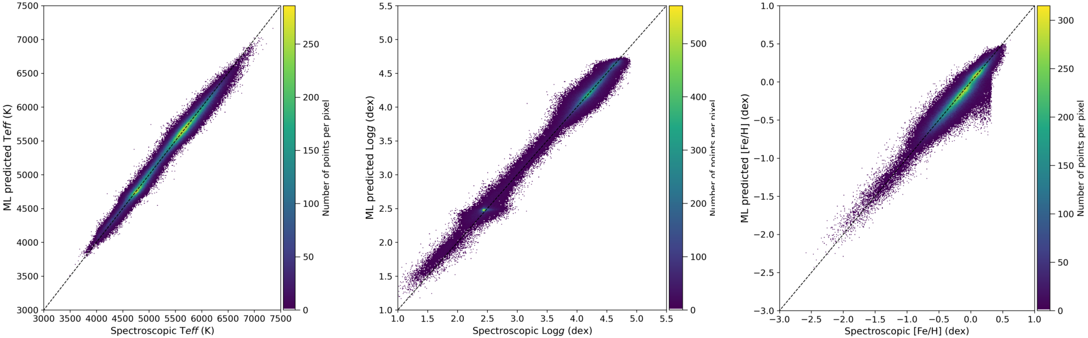
- Geometric distance from *Bailer-Jones et al. 2021*; Gaia parameter; SKYMAPPER parameter. Absolute magnitudes have both regular and de-reddened version. Errors are included. Real params are doubled due to NaN flags.

Parameter	Parameter	Parameter
r_med_geo	Rabs	e_r_psf
phot_bp_mean_mag	Rabs_nored	e_i_psf
phot_rp_mean_mag	Iabs	e_z_psf
Uabs	Iabs_nored	teff_gspphot
Uabs_nored	Zabs	mh_gspphot
Vabs	Zabs_nored	logg_gspphot
Vabs_nored	e_u_psf	teff_gspphot_err
Gabs	e_v_psf	logg_gspphot_err
Gabs_nored	e_g_psf	mh_gspphot_err

The Results

Scatter plots on test subsamle (ML vs SoS-spectro)

Teff | log g | [Fe/H]



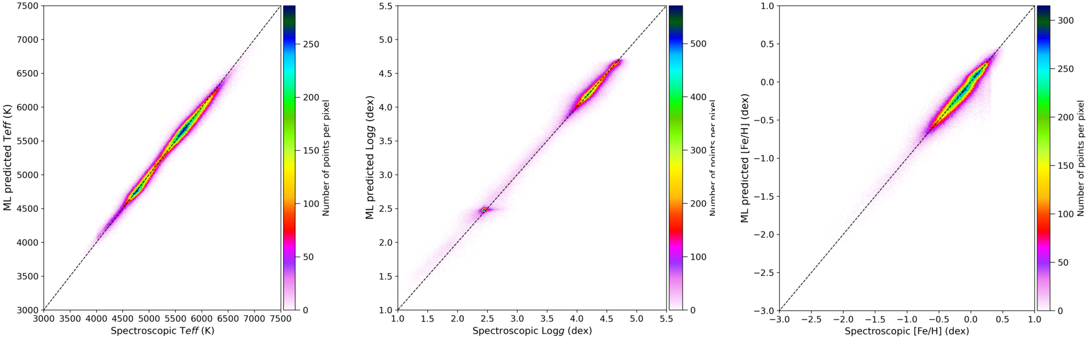
The Results

Just the same scatterplot with a different colorscale to appreciate the dense regions.

Teff

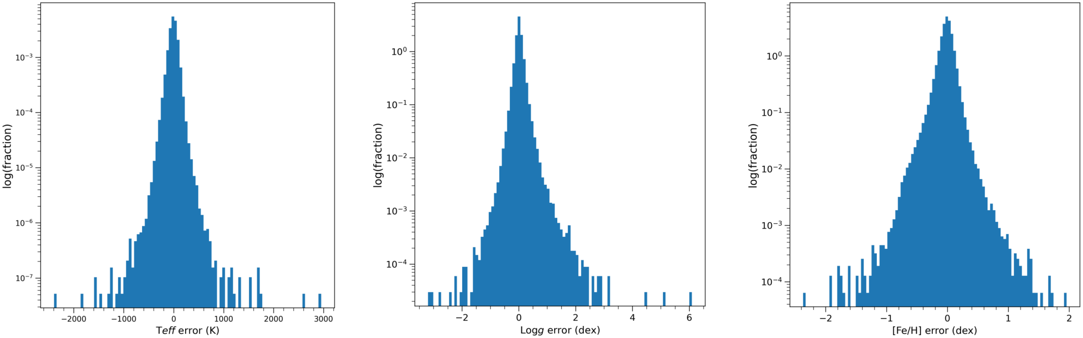
log g

[Fe/H]



The Results

Error distribution on test subsamle (Y axis in log scale!)



Teff

log g

[Fe/H]

The Results

Table of error metrics for the ML-predicted variables. One big caveat! We assume that SoS-spectro is the true value!

Metric	T_{eff} (K)	log g (dex)	[Fe/H] (dex)
Mean	–6.8	0.016	0.016
Median	-5.0	0.011	0.0097
σ	85.0	0.14	0.12
MAD	70.5	0.086	0.083

Comparison between Gaia photometry-derived parameters and other spectroscopic surveys:

Parameter	APOGEE, GALAH, LAMOST,RAVE		SoS	
	MAD	RMSE	MAD	RMSE
Teff (K)	150-418	228-1294	165	233
log g (dex)	0.102-0.406	0.163-0.626	0.213	0.304
Fe/H (dex)	0.238-0.303	0.295-0.450	0.170	0.277

Future Work

- We plan to replicate the work with SDSS to add nothern hemisphere coverage.
- Future plans include also PANSTARRS, 2MASS, APASS. We aim to include most of the surveys to make the future SoS releases as complete as we can.
- We bought a big server with GPUs, so we aim to implement more complex models in order to achieve potentially better performance!
- Another line of work will probably be trying to predict different key parameters such as distance.

Future Work

Thanks for your attention!