

# PREPARING ESA Gaia MISSION: STELLAR SPECTRA PARAMETERIZATION BY NEURAL NETWORKS IN THE NEAR INFRARED

DIEGO ORDÓÑEZ BLANCO<sup>1</sup>, JOSÉ CARLOS DAFONTE VÁZQUEZ<sup>1</sup>,  
BERNARDINO ARCAJ VARELA<sup>1</sup>

*1 Departamento de Tecnologías de la Información y de las Comunicaciones,  
Universidade da Coruña, E-15071, A Coruña, SPAIN*

MINIA MANTEIGA<sup>2</sup>

*2 Departamento de CC. de la Navegación y de la Tierra, Universidade de A  
Coruña, E-15011 A Coruña, SPAIN*

ANA ULLA MIGUEL<sup>3</sup>

*3 Departamento de Física Aplicada, Área de Astronomía y Astrofísica, Universidade  
de Vigo, E-36310, Vigo, SPAIN*

**Abstract:** The Gaia mission of the European Space Agency, foreseen to be operative at the beginning of 2012, will extend the Hipparcos legacy by carrying out a census of the Galaxy and providing accurate information about the composition and motion of its main components. Data handling and analysis of information regarding the complete sky down to magnitude 17-18 will be, with no doubt, a challenge for both Astrophysics and Computational Sciences. We present here our preliminary results on the on-going study about the automated derivation of stellar atmospheric parameters in the spectral region of the Gaia RVS (Radial velocity spectrometer) instrument. The use of artificial neural networks (ANN) trained with synthetic model spectra was the method selected for such automated derivation. Both direct stellar fluxes and Fourier transform moduli of them have been considered as inputs to train and test the ANN performance. It is shown that ANN represent a good approach to analyze and parameterize such a large dataset as the one expected from Gaia. Preliminary results achieved are comparable to those obtained by direct spectroscopic or spectrophotometric analysis with synthetic model atmospheres, being their accuracy highly dependent on the spectral signal to noise ratio.

**Keywords:** Stellar spectroscopy – Stellar parameterization – Gaia mission – Artificial Neural Networks.

## 1 Gaia RVS instrument

The Gaia satellite, which is foreseen to be launched near the end of 2011, is one of the present key scientific missions of the European Space Agency (ESA). Gaia mission

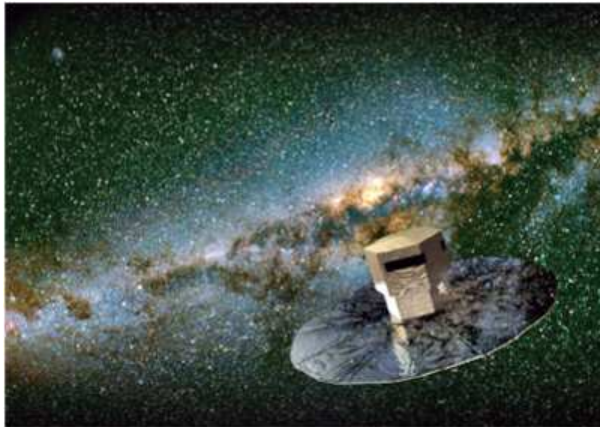


Figure 1: Artist's impression depicting the Gaia spacecraft against a background suggestive of the Milky Way. Figure courtesy of ESA. Illustration by Medialab.

will carry out the most accurate study of the Galaxy components, by compiling exact information on their nature and motion. In the course of its 5 operative years, it will perform precise astrometry (up to 10 microarcsecs at  $V=15$  mag.) and measure the motions of all the sources in the sky plane. This will allow to calculate distances via parallaxes with a precision of 1% for a total of 2.5 million stars down to 2.5 kpc. An artistic view of Gaia spacecraft orbiting the Milky Way can be seen in Figure 1.

Gaia will be equipped with a wide-band spectrophotometer and a radial velocity spectrograph, RVS, that will contribute to the study of the sources' nature and will allow to determine radial motions with precisions between 1-10 km/s for  $V=16-17$  mag. [1]. The RVS domain is the Ca II infrared region, 847-874 nm, a region which is rich in diagnostic lines for the determination of stellar atmospheric parameters, in particular, effective temperatures, surface gravities, overall metallicities and non canonical alpha-elements abundances. This wavelength range has been selected to coincide with the energy-distribution peaks of G and K-type stars which are the most abundant RVS targets. For these late type stars the wavelength interval displays three strong ionized Calcium lines and numerous weak lines, mainly due to Fe, Si and Mg. In early type stars, RVS spectra will be dominated by Hydrogen Paschen lines and may contain weak lines due to carriers as CaII, HeI, HeII and NI. The instrument

operates in time-delayed integration mode, observing each source about 40 times during the 5 years of the mission. Figure 2 shows Gaia RVS spectral domain for a F type star and two values of the signal to noise ratio. Comprehensive information about this project can be found in the Gaia web area at <http://www.rssd.esa.int/gaia>

Over the 5 years mission, RVS will observe around 5 billion transit spectra of the brightest 100-150 million stars on the sky. The on-ground analysis of these spectroscopic data set will be a complex and challenging task, not only because of the volume but because the interdependence of different instruments and modes of observation. As a consequence, data extraction and parameterization should be performed completely in an automatic fashion. These are the reasons why the use of Artificial Intelligence techniques and, in particular, artificial neural networks (ANN), is a good approach to be tested for the case of Gaia-RVS dataset.

## 2 Spectralib: a library of synthetic spectra for Gaia RVS.

Our initial approach consisted of performing simulations on stellar parameter extraction by means of synthetic spectra. For these first tests it was used the Gaia RVS *Spectralib*, a library of 9285 stellar spectra compiled by A. Recio-Blanco and P. de Laverny from Niza Observatory, and B. Plez from Montpellier University. The spectra are based on the new generation of MARC models from The Uppsala Observatory in the RVS region. A technical note describing the models used for the atmospheres from which the synthetic spectra were calculated and what parameters were used is available ([2]). This set of spectra will be named RVS1. More recently, tests were performed with an updated new set of 7168 spectra based in the same library of MARC models specifically computed for RVS Gaia simulations (from now on RVS2 set of spectra). Both grids are essentially equivalent and consist on spectra corresponding to effective temperatures between 4000 and 8000 K (step 250K),  $\log g$  between -1.0 to 5.0 (step 0.5 dex, range -0.5 to 5.0 dex for RVS2), and overall metallicities between -5.0 and 1.0 (with variable step from 1.0 to 0.25 dex). For each model atmosphere, alpha-elements abundance variations of at least the following values: +0.4, +0.2, +0.0, -0.2 dex, were considered with respect to the original abundances in the models. The grid closely covers HR diagram positions for A5, F, G and K stars. Access to RVS-Spectralib is open via the ESA Gaia web pages (<http://gaia.esa.int/spectralib/>).

## 3 The use of ANN for spectral parameterization

Astrophysics is moving towards a more rational use of the costly observational material by means of the intelligent exploitation of large terrestrial and spatial astro-

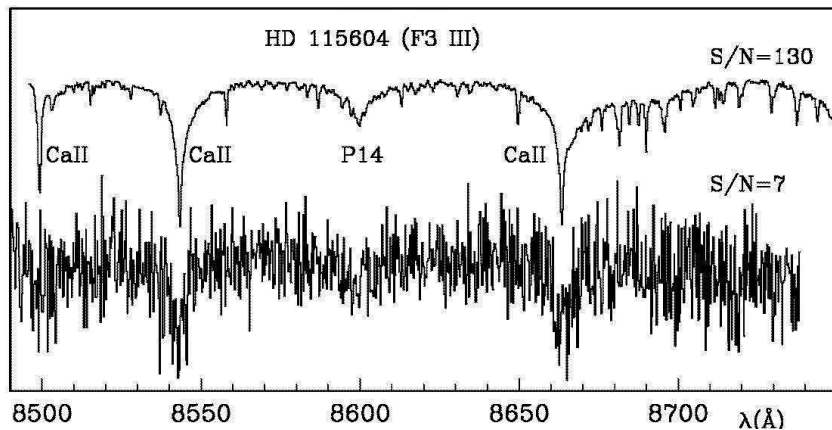


Figure 2: RVS spectral region showing the principal features for a F3III star and two SN levels.

nomical databases. Nowadays, every single project on astronomic instrumentation includes the creation of data archives and their future exploitation, with automatic or pseudo-automatic analysis tools. Several international projects are currently working on management systems for astronomical information by means of the common exploitation of all the available spatial archives and astronomical databases (international virtual observatories initiatives). Our research group is a member of the Spanish Virtual Observatory ([3]), which consists of various networked Spanish groups and aims, among other things, at developing applications for the automatic analysis of astronomical data. AI techniques are among the techniques that are being used for those developments.

Among the different techniques of Artificial Intelligence, ANN have already proved their success in classification problems: they are generally capable of learning the intrinsic relations that reside in the patterns with which they were trained. Some well-known previous works have applied this technique to the problem of stellar spectral parameterization, obtaining different grades of resolution in the extraction of parameters  $T_{\text{eff}}$ ,  $\log g$ ,  $[\text{Fe}/\text{H}]$  and  $[\alpha/\text{Fe}]$ . Most authors have used multilayer neural networks (MLP) with supervised learning algorithms based on the backpropagation of the training error. As examples we can mention the already classical works by [4], [5], [6], and [7], as well as the more recent works by [8] and [9]. A summary of the current status of automated stellar classification techniques and achievable accuracies can be found in the reviews by [10] and [11]. The range of accuracies ( $1\sigma$ ) obtained in the derivation of the atmospheric parameters are of the order of 150-200 K, 0.3-0.4

dex and 0.15-0.2 dex for Teff, log g and [Fe/H], respectively.

In order to probe the adequacy of RVS-spectral region for the derivation of stellar atmospheric parameters, we choose to train ANN with the ad-hoc calculated synthetic spectra already introduced in the previous section. Tests with ANN were performed in two stages. In an initial phase we wanted to study how signal to noise ratio influences the ANN training phase and parameter determination errors for the case of Teff, log g and [Fe/H] extraction. In this phase the RVS1 set of spectra was considered for the tests. In the second stage, we were aimed to consider the performance of the ANN when a different transformed domain for the input spectra is taken into account. Instead of feeding the ANN with the continuum subtracted spectra, their Fourier transformed absolute values (showing the energy distribution of frequencies among the spectra) were considered as the input information to train the networks and validate the parameters values. In this second phase the RVS2 set of spectra was considered for the tests.

### 3.1 RVS spectral parameterization. Signal to noise dependence.

A simple network architecture was chosen for the tests: a feedforward ANN with 1004 input nodes (the number of pixels in the spectra), 1 hidden layer with 150 nodes, and 4 output layers providing the atmospheric parameters, Teff, logg, [Fe/H] and  $[\alpha/\text{Fe}]$ .

A total of 1764 spectra well distributed in the space of parameters were considered in the training set, while tests were performed on subsets of 465 spectra. Typically, a good performance in the network convergence and low parameter errors were achieved after about 10000 training cycles, which translates to about 24 hours of computational time on an AMD 64 computer (AMDx64 Dual Core 3200GHz 2GB RAM).

Gaia-RVS spectra will be of very different quality depending mostly on the stellar brightness. It has been proposed that the end-of-mission SN ratio for a typical star in the Galaxy, a G5V with  $V=15.5$  or a F2II with  $V=14.5$  will be about 10. In order to have into account the effect of the SN values in the ANN performance, we have delivered tests taking into account six values of the SN ratio: 10, 25, 50, 75, 100 and 500. The model of noise considered was a simple gaussian white noise.

As expected, it was found that the mean errors in the extraction of the parameters depend on the SN ratio. It was also found that better results were achieved when the training set contains spectra with a similar distribution of noise as the validation set. Validation on spectra displaying different SN provides different degrees of accuracy for the atmospheric parameters. We chose to train the networks with a training set composed by spectra showing a distribution of SN ratios among 10 to infinite, representative of the whole sample to be validated. Mean errors as low as 50.0 K, 0.10, 0.09 and 0.06 dex for effective temperatures, logg, metal abundance and  $[\alpha/\text{Fe}]$ , respectively, were found for ANN tested on synthetic spectra with no noise added.

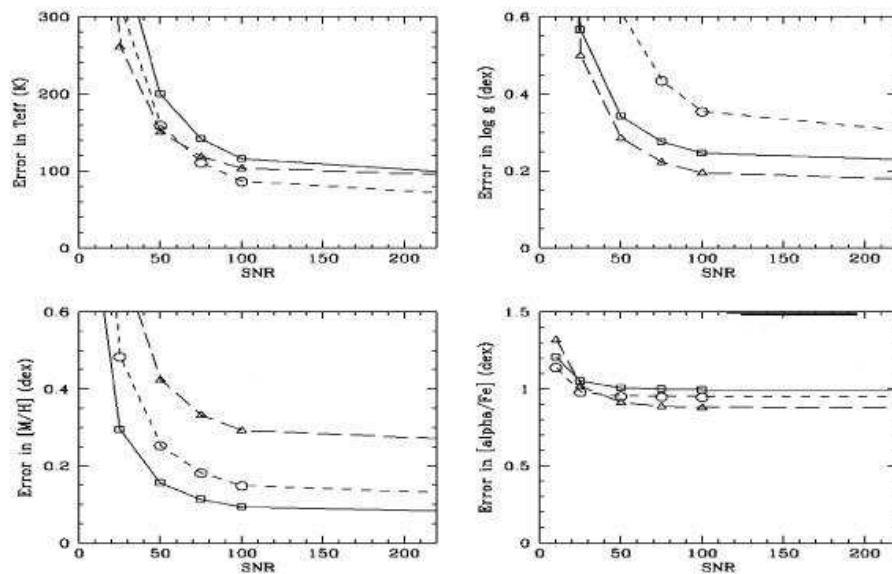


Figure 3: Results on the parameters extraction as a function of S/N, and for different stellar populations as follows: squares (cool dwarfs, metal rich), circles (cool giants, intermediate metallicity) and triangles (hot subgiants, very metal poor).

The errors grew to values of 74 K, 0.17, 0.13 and 0.08 dex in the case of SN 100; 115 K, 0.24, 0.19 and 0.11 dex for SN 50; 191 K, 0.37, 0.30 and 0.16 dex for SN 25 and finally to 391 K, 0.77, 0.55 and 0.25 dex for SN 10.

Figure 3 shows the performance of the ANN as a function of SN when considering validation on 3 set of spectra representative as tracers of three galactic stellar populations: cool dwarfs, metal rich stars ( $T_{\text{eff}} 5000\text{-}6000\text{K}$ ,  $\log g \geq 3.5$  dex and  $[\text{Fe}/\text{H}] \geq -0.5$  dex); cool giants, intermediate metallicity ( $T_{\text{eff}} \geq 5000$  K,  $\log g = 1.0\text{-}3.5$  dex,  $[\text{Fe}/\text{H}] \geq -1.0$  dex); and hot subgiants, very metal poor stars ( $T_{\text{eff}} \geq 6000\text{K}$ ,  $\log g = 2.0\text{-}4.0$  dex,  $[\text{Fe}/\text{H}] = -1.0\text{-}2.5$  dex). From the data in the figures, it is obvious that the SN heavily influence the learning process. In our experience, essentially poor results are encounter when training and testing spectral samples with different SN values.

### 3.2 RVS spectral parameterization in the Fourier transformed domain.

RVS2 set of spectra needed to be preprocessed before transformed to the Fourier domain. The procedure consisted in the following steps: continuum fit by linear regression and continuum subtraction, subtraction of the mean value and multiplication by a hamming window. Then, calculation the Fast Fourier Transform (FFT) algorithm was applied and the absolute values of the FFT determined. As in the previous experiment, one network was trained for each parameter ( $T_{\text{eff}}$ ,  $\log g$ ,  $[\text{Fe}/\text{H}]$  and  $[\alpha/\text{Fe}]$ ). Again, feed-forward ANN fully connected were considered and for the training phase a standard on line backpropagation algorithm. The dimension of the input layer was the same as the number of points in the processed spectra (400 points after preprocessing and filtering), and the output layer had a unique process element to obtain each of the parameters.

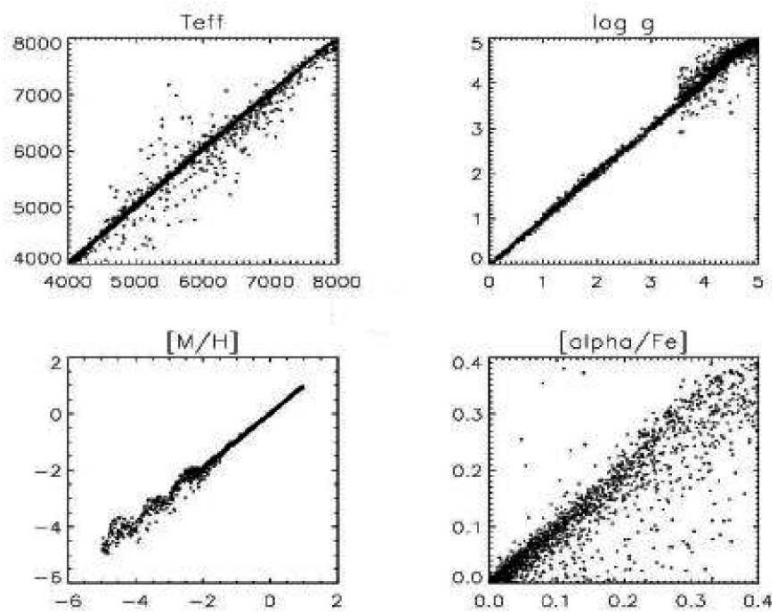


Figure 4: Astrophysical parameters recovering with ANN trained in the Fourier domain.

The method was found to be very robust. Astrophysical parameters can be determined with the following mean errors :  $T_{\text{eff}}$  54 K,  $\log g$  0.08 dex,  $[\text{Fe}/\text{H}]$  0.06 dex

and  $[\alpha/\text{Fe}]$  0.03. Ruling out the 20% worse cases, errors in the following range were found:  $T_{\text{eff}}$  -35 to 62 K,  $\log g$  -0.05 to 0.16 dex,  $[\text{Fe}/\text{H}]$  -0.01 to 0.12 dex, and  $[\alpha/\text{Fe}]$  -0.07 to 0.1 dex. When compared with the previous experiment (RVS1 set with no noise added) it can be seen that an improvement in the parameterization of  $\log g$ ,  $[\text{Fe}/\text{H}]$ , and  $[\alpha/\text{Fe}]$  has been achieved, while essentially the same errors are found in the recovery of  $T_{\text{eff}}$ .

The performance of the method is shown in Figure 4 where the error distribution for the recovery of each four parameters is shown, while in Figure 5 it is shown a gaussian fit for the central, best behaved population.

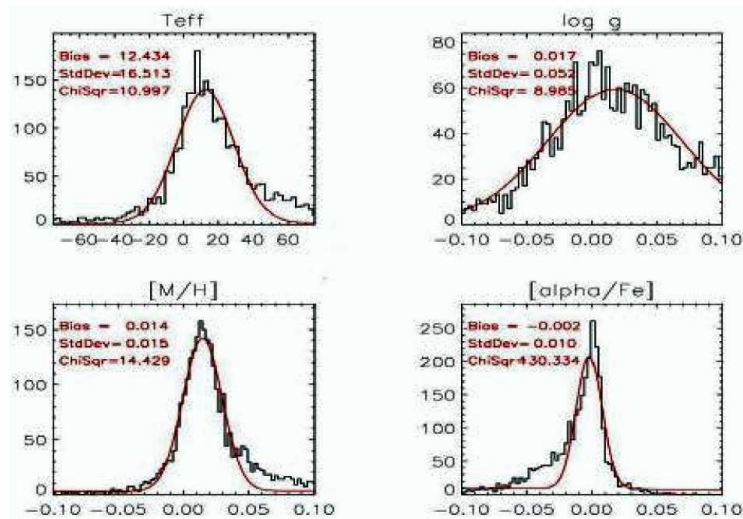


Figure 5: Distribution of errors for derivation of astrophysical parameters with ANN trained in the Fourier domain.

## 4 Conclusions and future work

We presented our first results on the automatic derivation of stellar parameters in the RVS spectral region, by the use of artificial neural networks trained with synthetic model spectra. The results achieved are comparable to those obtained by the use of spectrophotometry, being the accuracy highly dependent on the signal to noise ratio of the spectra. The training of ANN in the FFT domain provides a novelty factor, and the good results found in the performed experiment open a new way to be explored in the future with noise-added spectra.

Our results show that ANN can be a good approach to extract atmospheric parameters from Gaia-RVS spectra, providing that the SN ratio of the training and testing spectral set be well characterized. Mean errors as low as 115 K, 0.24, 0.19 and 0.11 dex for effective temperatures, logg, metal abundance, and  $[\alpha/\text{Fe}]$ , respectively, were reached for synthetic spectra with SN 50.

*Acknowledgements:* This work is supported by the Spanish MEC ESP2006-13855-C02-02 and FEDER funds.

## References

- [1] Katz, D., Munari, U., Cropper, M. et al. 2004, MNRAS 354, 1223
- [2] Recio Blanco, A., de Laverny, P., Plez, B. 2005, European Space Agency technical note RVS-ARB-001
- [3] Solano, E. 2006, Lecture Notes and Essays in Astrophysics, vol. 2, 71
- [4] von Hippel, T., Storrie-Lombardi, L.J. et al. 1994, MNRAS 269, 97
- [5] Singh, H.P., Gulati, R.K. and Gupta, R. 1998, MNRAS 295, 312
- [6] Bailer-Jones, C.A.L., Irwin, M., Gilmore, G. et al. 1997, MNRAS 292, 157
- [7] Vieira, E.F., Ponz, J.D. 1995, A&AS 111, 393
- [8] Gupta, R., Singh, H.P., Volk, K. et al. 2004, ApJS 152, 201
- [9] Willemsen, P.G., Hilker, M., Kayser, A. et al. 2005, A&A 436, 379
- [10] Bailer-Jones, C.A.L. 2002, in “Automated Data Analysis in Astronomy”, R. Gupta, H.P. Singh, C.A.L. Bailer-Jones (eds.), Narosa Publishing House, New Delhi, India. Preprint: arXiv:astro-ph/0102223v1
- [11] Allende Prieto, C. 2004, Astronomische Nachrichten, Vol.325, Issue 6, 604