

Automated stellar abundance analysis

Alejandra Recio-Blanco¹

¹ Laboratoire Lagrange (UMR7293), Université de Nice Sophia Antipolis, CNRS, Observatoire de la Côte d'Azur, BP 4229, F-06304 Nice cedex 4, France
email: arecio@oca.eu

Abstract. The advent of Milky Way high-resolution spectroscopic surveys has brought our attention to the importance of precise chemical abundance measurements to disentangle the stellar population puzzle of the Galaxy. Moreover, automated stellar parameters are the bedrock of Galactic spectroscopic surveys science. They allow a rapid and homogeneous processing of extensive data sets, necessary for an efficient scientific return. In this review, I discuss the different parametrization techniques, focusing on the automated determination of individual element abundances. Each of them has its optimal application conditions that mainly depend on the computation time constraints, the spectral resolution, the wavelength domain, the data signal-to-noise ratio and parameter degeneracy problems. The main algorithms in the literature are also reviewed.

Keywords. methods: data analysis, surveys, stars: abundances, stars: fundamental parameters, Galaxy: abundances

1. Introduction

A suite of ground-based vast stellar surveys mapping the Milky Way and culminating in the ESA Gaia mission, is revolutionizing the empirical information about Galactic stellar populations. In particular, in the recent years, the number of stars analysed with high enough spectroscopic resolution to provide detailed chemical diagnostics has increased from a few hundreds to several tens of thousands. Until the end of 2003, most of the information about the Milky Way was confined to small local samples, for which high-resolution spectroscopic data was obtained. In 2004, the Geneva Copenhagen Survey (Nordström *et al.*, 2004) collected the first large spectro-photometric sample of around 16 000 stars, as part of a Hipparcos follow-up campaign (hence, also confined to 100 pc from the Sun). More recently, optical spectroscopic low-resolution surveys, such as SEGUE (Yanny *et al.* 2009) and RAVE (Steinmetz *et al.*, 2006), have extended the studied volume to distances of a few kpc from the Sun (mainly in the range 0.5-3 kpc), and increased the numbers of stars with chemo-kinematical information by more than an order of magnitude (> 200 000 spectra for SEGUE and > 500 000 spectra for RAVE).

This effort is now complemented by new vast high-resolution spectroscopic surveys: the Gaia-ESO Survey (GES, 300 nights with the ESO/VLT), the Gaia/Radial Velocity Spectrograph (RVS) survey (part of the Gaia cornerstone mission), the Australian HERMES/GALAH survey, the LAMOST/LEGUE survey and APOGEE (part of the Sloan Digital Sky Survey III and After-SDSSIII). The general context of Galactic spectroscopic surveys is presented in Figure 1, by locating each project in a space formed by three axis: targeted galactic populations, spectral resolution and probed magnitude range. As the spectral resolution is usually fixed for a particular survey, each project appears as a plane in the figure.

As explained in Gilmore *et al.* (2012), there are four basic observational thresholds that we need to pass to have a complete view of the fossil record of our Galaxy formation and

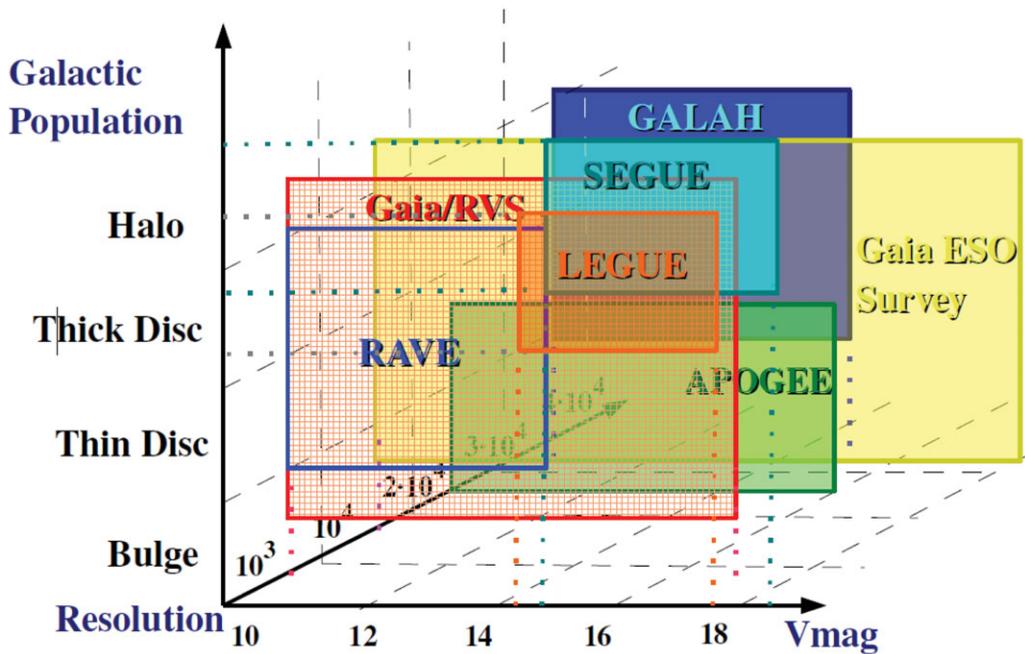


Figure 1. Galactic spectroscopic surveys presented in a space formed by three axis: targeted galactic populations, spectral resolution and probed magnitude range.

evolution: the first step consists in source identification (or discovery) through its position and photometric data. This is provided today by photometric surveys of the Galaxy like VISTA and VST. Secondly, the temporal domain is added, allowing us to explore a 5-dimensional space formed by parallaxes (and therefore distances) and proper motions (or transverse velocities). The third step is radial velocity, that allows to determine stellar orbits. Finally, if the spectral domain and resolution and the data quality are high enough, the stellar chemical abundances can be derived.

The possibility of determining individual abundances from the above mentioned intermediate and high-resolution surveys of the Milky Way has brought our attention to the importance of precise chemical abundance measurements to disentangle the stellar population puzzle of the Galaxy. The advent of Milky Way spectroscopic surveys and of automatized chemical analysis techniques have improved both the statistical robustness and the homogeneity of the data as a consequence. From the theoretical side, new approaches based, at least partially, on the chemical identification of disc sub-populations, as the chemical-tagging (Freeman & Bland-Hawthorn, 2002) or the mono-abundance populations (Rix & Bovy, 2013) methods, are opening promising pathways for constraining the Milky Way's evolutionary processes. In addition, as discussed in Bovy *et al.* (2012), defining stellar populations by abundances patterns is a better approach than the traditional kinematical criteria, as chemical abundances can, for instance, correlate with disc structure, but are formally independent of it. The stellar chemical patterns can guide the definition of useful stellar sub-samples to unveil the evolutionary paths of the Milky Way components formation history and to give clearer constraints to the models.

All the above mentioned Galactic archaeology approaches rely on the success of automated techniques of spectral analysis and classification, capable to perform a rapid and homogeneous processing of the data and to allow an efficient scientific return. Figure 2 shows the increase of stars with available spectra in the next five years, as expected

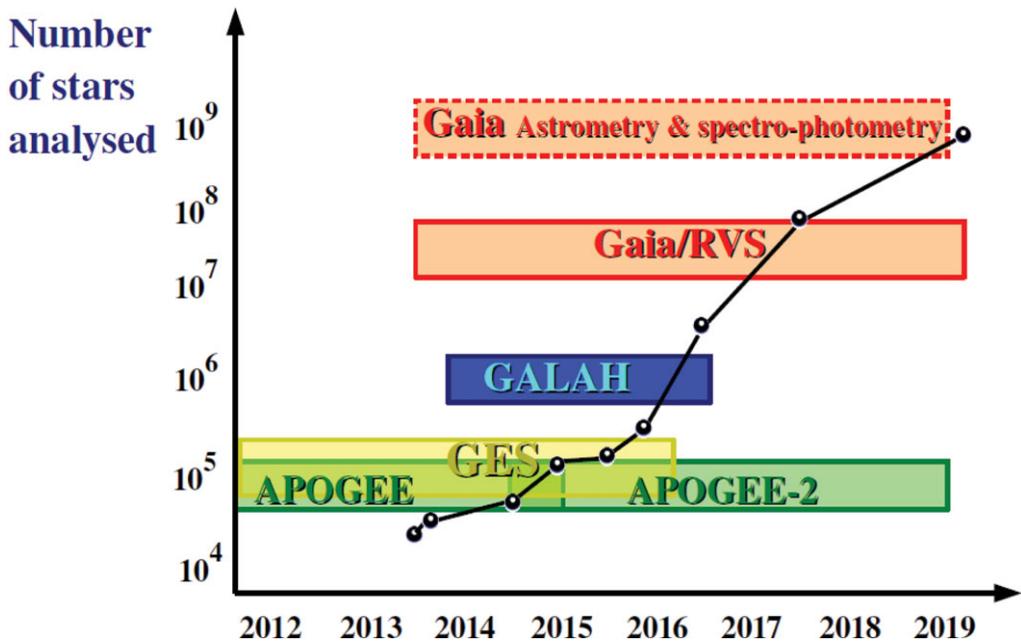


Figure 2. Number of stars with available spectra in the next years, as expected from the announced data releases of the different Milky Way surveys.

from the announced data releases of the different Milky Way surveys. One of the main challenges of Galactic Archaeology will be the correct and rigorous treatment of all those data sets, including their chemical characterization.

2. Parametrization

Automated stellar abundance analysis is based on data mining methodologies that use different parametrization approaches. The used data mining techniques depend on the degree of knowledge of the studied objects. First, when the observed data are badly constrained or when no *a priori* exists, un-supervised classification algorithms are used. Secondly, when the different types or classes of objects are known in advance, supervised classification methods are employed. In this case, reference data are needed to perform the mapping between the observations and the corresponding classes. The Morgan-Keenan classification of stellar spectra is an example of this kind of approach. Finally, automated parametrization is used when the physics of the studied objects is enough well known, and modelled through continuous variables. For instance, the stellar effective temperature, the surface gravity, the global metallicity and the individual element abundances are more appropriate to describe a stellar spectra than spectral types and luminosity classes.

2.1. Mathematical approaches of parametrization

As for supervised classification approaches, parametrization algorithms use reference data to define the mapping between the observed targets and the models. Those models, usually synthetic spectra, constitute a N-dimensional grid, where N is the number of parameters to determine.

The main goal of a parametrization algorithm is to minimize the distance function defined by:

$$D = \sum_{j=1}^J |O(j) - M(j)|^2 \quad (2.1)$$

where $O(j)$ and $M(j)$ are the observed spectrum and the reference (usually synthetic) one, respectively. They are defined through J variables (the spectrum pixels, for instance). The distance function can be a non-convex function depending on the considered reference model of the grid, the measured parameter value and the noise threshold. Each reference model with a distance value smaller than the adopted noise threshold can be considered as a potential solution.

There are three main mathematical parametrization approaches: optimization methods, projection methods and classification. All of them try to find the absolute minimum of the distance function, with different techniques.

- *Optimization methods:* The exhaustive exploration of the grid of reference models in order to find the nearest neighbour is the simplest optimization method. It consists in calculating the distance between the observed spectrum and each of the spectra of the reference grid. This technique can be very time consuming, but it ensures the finding of the absolute minimum. Nevertheless, the precision of the result is limited by the parameter step of the reference grid. To tackle this problem, different methods of interpolation in the parameter space can be used. The nearest neighbour method can also be affected by the noise, as noise features can be confused with spectral signatures of a given physical parameter. For instance, the strength of a magnesium line can be modified by the noise in a way that mimics a higher or lower magnesium abundance than the real value. In those cases, the reference model to which the calculated distance is smaller, will not be the one with the nearest parameters (e.g. magnesium abundance) to the target one. Gradient descent methods, as the Gauss-Newton algorithm, have been developed in order to reduce the computing time of the minimum distance method (see, for instance, Bijaoui *et al.*, 2012). Their goal is to find the direction in the parameter space that has the highest negative gradient as a function of the distance. Once this direction is found, the methods proceed in an iterative way, by modifying the initial guess of the studied parameter and re-calculating the gradient again, until convergence of the parameter solution. This kind of algorithms does not guarantee the convergence to the absolute minimum, as they can be trapped in secondary minimum when the application conditions are those of bad quality data. Other methods, as the Nelder & Mead (1965) algorithm have been developed to avoid the local minima traps, although the computation time can significantly increase.

- *Projection methods:* Contrary to optimization methods, projection algorithms need a training phase during which a set of projection vectors is calculated. Those vectors contain the most important signatures of the flux that allow the derivation of a given physical parameter. Then, during the application phase, this parameter value (for instance, the abundance of α -elements with respect to iron) is determined by the projection of the target spectrum into the calculated vectors. The MATISSE method (Recio-Blanco *et al.*, 2006) and the penalized χ^2 algorithm are examples of this kind of approach. When the distance function is convex and linear, projection methods allow to project the target spectra close to their nearest neighbours. This considerably reduces the application time. In addition, the projection vectors can be adjusted in order to take into account the noise effects, given less weight to high order variations of the stellar flux. Nevertheless, this kind of methods can also be trapped in secondary minima, usually not very far from the

absolute one, as they have a local application that avoids a global view of the parameter space.

- *Classification methods:* In the limit of the sampling precision, the parameter estimation is a recognition problem. The grid of synthetic spectra can be treated as a known set of patterns among which we aim to identify the observed spectra. In the learning phase, the recognition rules are established using the grid of theoretical spectra. Decision trees (as DEGAS, Kordopatis *et al.*, 2011), neural networks and support vector machines are examples of classification algorithms used for stellar parametrization.

3. Important issues linked to element abundance determination

In this section, I would like to briefly discuss a certain number of questions, related to stellar element abundance determination, that will not be treated more extensively in the following.

First of all, the derivation of a given element abundance depends on the stellar atmospheric parameters of the target star: effective temperature, surface gravity, global metallicity and microturbulence. Those parameters can be known *a priori* (for instance, asteroseismic data can provide very precise surface gravity estimations), they can be derived as a first step of the spectral parametrization, or they can be determined simultaneously with the element abundances. The last approach faces the problem of deriving a high number of parameters at the same time. The exploration of a higher dimensional space increases the complexity of the algorithms and the risk of parameter degeneracies (that occurs when the variation of two different parameters produces similar changes in the normalised stellar flux of spectral features). Degeneracy problems are usually more severe when the available information about the parameters decreases: e.g. with a more narrow spectral range, lower spectral resolution, lack of spectral signatures, and so on. In addition, secondary minima can also be artificially generated by noise disturbing the distance function. Different projects (e.g. GES, AMBRE - see Worley *et al.*, 2012) allow the derivation of one individual element abundance or one element abundance ratio (usually $[\alpha/Fe]$) simultaneously with the atmospheric parameters. When more than one chemical element is studied, the abundance analysis is usually done as a second step. The total number of parameters, that the algorithms are presently able to simultaneously determine, is generally not higher than four.

On the other hand, continuum normalization is a crucial part of any abundance analysis. For this reason, automated procedures of element abundance determination are linked, in an iterative way, to automated continuum normalization algorithms (e.g. Kordopatis *et al.*, 2011; Zwitter *et al.*, 2008).

In addition, stellar rotation is also linked to element abundance derivation, as it modifies the spectral line profiles. It can be derived in a previous step, as it is the case for the Gaia-ESO Survey spectrum analysis, thanks to a first guess of the stellar atmospheric parameters. It can also be determined simultaneously with them, specially for fast rotators.

Finally, I would like to warn that the quality of the reference physics included in the used line lists, model atmospheres, spectral synthesis, etc... is as important for automated element abundance methods as for classical non-automated approaches. It can be the source of considerable external errors and it has also to be checked when several results, coming from different abundance analysis, are combined to determine the final recommended values.

	Reference spectra	Main applications	Mathematical approach
 Computation time ↑	- On the fly computations:		
	Spectral Synthesis	GES, GALAH	Optimization
	Equivalent widths	GES	Optimization
	- Pre-computed grid:		
Without training	GES, RAVE, SEGUE, LEGUE	Optimization	
With training	GES, RAVE, AMBRE, SEGUE, Gaia	Projection & Classification	

Figure 3. Different types of automated abundance analysis methods currently used by Milky Way spectroscopic surveys and galactic archaeology projects.

4. Automated abundance analysis methods

Figure 3 shows the different types of automated abundance analysis methods found in the literature, currently used by Milky Way spectroscopic surveys and galactic archaeology projects. The algorithms can be divided depending on the way in which the reference models are computed and used. This has an important influence on the computation time and, ultimately, it depends on the implemented mathematical approach. In addition, Figure 3 shows what approaches are used by the main spectroscopic surveys and projects, including the European Space Agency Gaia mission.

In the next subsections, the main algorithms in the literature are reviewed, following the scheme of Figure 3. The list of methods does not intend to be complete, but it will guide the reader through the general picture of approaches developed up to date.

4.1. Methods using on the fly computations of the reference synthetic spectra

These are the simplest parametrization algorithms. They implement the atomization of the classical procedures through optimization approaches. The methods in the literature can be divided between those using spectral synthesis and those using the equivalent widths of spectral lines.

- *Algorithms using spectral synthesis:*

Spectroscopy Made Easy (SME) was developed by Valenti & Piskunov (1996). The SME algorithm is a χ^2 optimization method. It is used for the determination of atmospheric parameters and individual element abundances. SME is part of the methods used in the Gaia-ESO Survey spectrum analysis of FGK-type stars. Other minimum of distances methods recently implemented are the one developed by Van der Swaelmen *et al.* (2013) and Mikolaitis *et al.* (2013, submitted to A&A). The last one is used for the derivation of individual abundances from the GIRAFFE data of FGK-type stars observed by GES. Finally, the Australian GALAH survey project is implementing an AUTOMOOG based method for its future abundance determination pipeline.

- *Algorithms using equivalent widths:*

Automatized classical manual procedures are implemented in this kind of approach. The automatic measurement of the spectral lines equivalent widths is coupled with on the fly fits of the observed equivalent widths with theoretical ones. The algorithm of the

Porto group of Sousa *et al.*, uses the automated equivalent width measurements with the code ARES (Sousa *et al.* 2007) and it is part of the GES methods used for the analysis of UVES data. It is also currently used for other studies of the solar neighbourhood like Adibekyan *et al.* (2012), and the characterization of stars hosting exoplanets. More recently, Mucciarelli *et al.* (2013) presented the GALA method, that uses an automated equivalent width measurement with DAOSPEC.

4.2. Methods using a pre-computed grid of reference synthetic spectra

The use of a pre-computed grid of synthetic spectra reduces the computing time of the algorithms application. The methods using this kind of approach are divided into those without and with a phase of algorithm training.

- *Without training:*

This category of methods is based on optimization approaches. It includes the Nelder-Mead method implemented by Allende-Prieto *et al.* (2006). This non-linear downhill simplex method was already used for the SDSS-SEGUE SSPP pipeline for the derivation of both the iron abundance and the $[\alpha/\text{Fe}]$ (Lee *et al.* 2011) from low resolution stellar spectra. It is also part of the methods used by the Gaia-ESO Survey and it is the core of the APOGEE ASPCAP pipeline.

The penalized χ^2 method of Zwitter *et al.* (2008) is also in this category of algorithms. It has been used for the RAVE survey first and second data releases for the derivation of the iron abundance. The RAVE survey also developed the Boeche *et al.* (2011) method for the individual abundance analysis of the high signal-to-noise data (third data release). It is a minimum of distances method using of a grid of pre-computed equivalent widths. The UlySS method of Koleva *et al.* (2009) implements a full spectrum fitting and a parametric minimization using χ^2 maps. It was part of the SEGUE SSPP pipeline and it is actually integrated in the LASP LAMOS pipeline for the analysis of the LEGUE survey data.

The GAUGUIN method (Bijaoui *et al.* 2012) uses the Gauss-Newton algorithm for the determination of the global metallicity simultaneously with stellar atmospheric parameters. It can also be used for the derivation of individual abundances in a second step of the spectrum analysis. The GAUGUIN algorithm is already applied to GES data and it is been prepared for its integration in the Apsis pipeline for the individual abundance analysis of the Radial Velocity Spectrograph (RVS) data, collected by the Gaia mission of the European Space Agency.

Finally, the Abbo and MyGIsFOS approach (Bonifacio & Caffau, 2003) and the SPADES method (Posbic *et al.* 2012) are also part of this type of algorithms.

- *With training:*

The methods with a faster application are those relying on a training phase. They are based on projection and classification approaches. The neural network algorithms of Re Fiorentin *et al.* (2007) is an example of this kind of method. It is part of the SEGUE SSPP pipeline for the derivation of the iron abundance, simultaneously with the effective temperature and the surface gravity. It implements a global and non-linear regression mapping.

The MATISSE and DEGAS methods are part of the algorithms developed by the Nice group. The MATISSE algorithm is a local multi-linear regression method. The stellar parameters are determined through the projection of the input spectra on a set of vectors, calculated during a training phase. The DEGAS method is based on an oblique k-d decision tree and uses the pattern recognition approach for stellar parameterization. The

MATISSE and DEGAS methods have been used in Kordopatis *et al.* (2011) for a study of the Thick Disc outside the solar neighbourhood (700 stars analysed) and for the last data release (DR4) of the RAVE Galactic Survey (Kordopatis *et al.* 2013, submitted, 228 060 spectra). These two applications share the same wavelength domain and resolution of the RVS one. In addition, MATISSE is the core method of the AMBRE project. AMBRE (see de Laverny *et al.* 2012), under agreement between the European Southern Observatory (ESO) and the Observatoire de la Côte d'Azur, aims at determining the parameters (T_{eff} , $\log g$, $[M/H]$ and $[\alpha/Fe]$) of the high resolution stellar spectra contained in the ESO archive. This concerns the FEROS, HARPS, UVES and FLAMES spectrographs. The results of the AMBRE project are presented in Worley *et al.* 2012 (FEROS data analysis), De Pascale *et al.* (2013, in preparation; HARPS data analysis) and Worley *et al.* (2013, in preparation, UVES data analysis). MATISSE has also been used for the characterization of several disc fields observed by the CoRoT mission (Gazzano *et al.* 2010 and Gazzano *et al.* 2013). In addition, the MATISSE algorithm is part of the methods used for the stellar parametrization of FGK type targets of the Gaia-ESO Large Public Survey. In particular, the first data release of GES parameters for the FGK-type stars observed with the GIRAFFE spectrograph includes the MATISSE results for those data.

Finally, the spectrum analysis of the Gaia mission is based on this type of algorithm with a training phase. The computational time is a crucial constraint for the Gaia pipeline, as several tens of millions of RVS spectra will be analysed in cycles of 6 months. The MATISSE and DEGAS algorithms are already included in the Gaia Apsis parametrization pipeline. They are the core of the Generalized Stellar Parametrizer-spectroscopy (GSPspec) module.

5. Conclusions

Automated stellar parameters are the bedrock of Galactic spectroscopic surveys science. If they are wrong, with hidden anomalous behaviours, the surveys scientific conclusions can be wrong. In the literature, several types of automated methods exist for the derivation of individual abundances. Each of them has its optimal application conditions that depend on the computation time constraints, the spectral resolution and wavelength domain, the data signal-to-noise ratio, the parameter degeneracy problems, etc... Present Galactic Surveys, including the Gaia mission, have developed a blossom of mathematical approaches that are currently used today. In this new era of Milky Way exploration, data analysis already has a crucial position.

References

- Allende-Prieto, C., Beers, T. C., Wilhelm 2006, *ApJ*, 636, 804
- Bijaoui, A., Recio-Blanco, A., de Laverny, P. *et al.* 2012, *Statistical Methodology*, 9, 55
- Boeche, C., Siebert, A., Williams, M. *et al.* 2011, *AJ*, 142, 193
- Bonifacio, P., Caffau, E. 2003, *A&A*, 399, 1183
- Bovy, J., Rix, H.-W., Liu, C. *et al.* 2012, *ApJ*, 753, 148
- de Laverny, P., Recio-Blanco, A., Worley, C. C. *et al.* 2012, *A&A*, 544, A126
- Freeman, K., Bland-Hawthorn, J. 2002, *ARA&A*, 40, 487
- Gazzano, J.-C., de Laverny, P., Deleuil, M. *et al.* 2010, *A&A*, 523, A91
- Gazzano, J.-C., Kordopatis, G., Deleuil, M. *et al.* 2013, *A&A*, 550, A125
- Gilmore, G., Randich, S., Asplund, M. *et al.* 2012, *The Messenger*, 147, 25
- Koleva, M., Prugniel, P., Bouchard, A. *et al.* 2009, *A&A*, 501, 1269
- Kordopatis, G., Recio-Blanco, A., de Laverny, P. *et al.* 2011, *A&A*, 535, 106
- Kordopatis, G., Recio-Blanco, A., de Laverny, P. *et al.* 2011b, *A&A*, 535, A107

- Lee, Y. S., Beers, T. C., An, D. *et al.* 2011, *ApJ*, 738, 187
- Mucciarelli, A., Pancino, A., Lovisi, L. *et al.* 2013, *ApJ*, 766, 78
- Nordström, B., Mayor, M., Andersen, J. *et al.* 2004, *A&A*, 418, 989
- Nelder, J. A. & Mead, R. 1965, *Computer Journal*, 7, 308
- Posbic, H., Katz, D., Caffau, E. *et al.* 2012, *A&A*, 544, 154
- Re Fiorentin, P., Bailer-Jones, C. A. L., Lee, Y. S. *et al.* 2007, *A&A*, 467, 1373
- Recio-Blanco, A., Bijaoui, A., de Laverny, P. 2006, *MNRAS*, 370, 141
- Rix, H.-W., Bovy, J. 2013, *A&A Rev.*, 21, 61
- Sousa, S. G., Santos, N. C., Israelian, G. *et al.* 2007, *A&A*, 469, 783
- Steinmetz, M., Zwitter, T., Siebert, A. *et al.* 2006, *AJ*, 132, 1645
- Valenti, J. A., Piskunov, N. 1996, *A&A Supp. Ser.*, 118, 595
- Van der Swaelmen, M., Hill, V., Primas, F. *et al.* 2013, *ArXiv*, 1306.4224
- Worley C., de Laverny, P., Recio-Blanco, A. *et al.* 2012, *A&A*, 542, A48
- Yanni, B., Rockosi, C., Newberg, H. J. *et al.* 2009, *AJ*, 137, 4377
- Zwitter, T., Siebert, A., Munari, U. *et al.* 2008, *AJ*, 136, 421