Minutes of the CU8 Calibration meeting

held in Bordeaux, 15–16 November 2011

prepared by: Coryn Bailer-Jones, Caroline Soubiran
reference: GAIA-C8-MN-MPIA-CBJ-062
issue: 1
revision: 0
date: 2011-11-25
status: Issued

Abstract

The meeting Wiki page is http://www.rssd.esa.int/wikiSI/index.php?title=CU8:_Calibration_meeting_Nov_2011&instance=Gaia. The presentations are available from SVN in http://gaia.esac.esa.int/dpacsvn/DPAC/CU8/docs/meetings/Calibration_Bordeaux_2011Nov/.

Document History

Issue	Revision	Date	Author	Comment
1	0	2011-11-25	CBJ	Final version following comments from meeting participants
D	0	2011-11-15	CBJ, CS	First draft

Contents

The particpants at the meeting were the set of all presenters, plus Sergi Blanco Cuaresma. Coryn Bailer-Jones was only present on the first day, and Andreas Korn only on the second.

1 The calibration issues and some thoughts on resolving them (Coryn Bailer-Jones)

The overall problem is that algorithms are trained on synthetic data, but for given astrophysical parameters (APs) these may differ from real data. This has been referred to as the *synthetic spectra mismatch problem*, and it arises from limitations of both the stellar models and the Gaia data simulators. Real (calibration) data somehow need to be used to minimize this difference. In an ideal world we would avoid this problem entirely by observing a sufficiently large number of objects with Gaia which have known APs, and simply train the AP estimation algorithms on these, then apply them to the rest. In practice we do not have enough such stars (although the proposed Gaia reference stars go some way towards fulfilling this).

Main points:

- physical parameters are unavoidably tied to physical models: control what physical system the resulting APs are tied to
- avoid intractable library homogenization: do AP estimation library-by-library (see below)

- the supervised learning paradigm requires us to model all effects we expect to see
- prioritize what needs to be done: dont try to solve the general problem; focus on getting good results on the sources/stars that matter most scientifically; quantify the level of AP accuracy we want to achieve
- there needs to be consistency between the libraries used in the Gaia processing and those used to parametrize the AP stars (or else we end up with a hybrid physical system)

Important open issues:

- how do we make our algorithms robust to real noise (e.g. outliers, artefacts, missing data)?
- how are our parametrization results affected by biases in the parameter distributions in the training data?

Editorial note: One issue raised in the presentation was how we can correct systematic errors we already see in our results based on synthetic only training and test data sets. To some extent these are an unavoidable consequence of the lack of signature of an AP in the data for some ranges of that AP (e.g. [Fe/H] at low [Fe/H]). This is explained and demonstrated in, for example, CBJ-046, section 2.

1.1 Proposal to do AP estimation library-by-library

Currently GSP-Phot is trained on a combination of different libraries (Phoenix, MARCS etc.), which are not consistent (i.e. predict different BP/RP spectra for given APs one is trying to determine). This can leads to inconsistencies and discontinuities in the resulting APs, plus these APs are tied to some hybrid and non-specific physical system. One solution is to try to homogenize the libraries. The alternative solution put forward was not to attempt this, but rather to perform **library-by-library** AP estimation. That is, a separate GSP-Phot model is trained on each library. Each star is parametrized by every model, so we end up with multiple AP estimates for each star. We ca report all, but as also need to make a sensible (i.e. weighted) combination of these estimates. Some stars may not be appropriate at all for a given model (e.g. classifying a hot star with a cool star model/library), so some kind of goodness-of-fit will need to be taken into account in the weighting (with some weights probably being zero). If DSC were trained on the same libraries, then it would actually provide a class (=library) probability for each model/library, which might be sufficient to use as the weights. For example, if we only had one white dwarf library, then for a real shite dwarf the hope is that the DSC probability for that library would be almost 1 (and therefore almost 0 for all others).

This concept is actually exactly what we normally do in astronomy: the literature is full of different estimates of the parameters or properties of objects based on different assumptions, models, data etc. The SDSS SSPP also used multiple models (trained on different libraries) to estimate APs and then combined these (latest approach described in Smolinski 2011). GSP-Phot already provides three AP estimates (from each module SVM, ILIUM and q-method), so it is not inconceivable to extend this to multiple libraries too (not necessarily all libraries with all modules). The physical system arising from combining APs is arguably easier to understand than the results of a nonlinear machine learning method trained on a mixture of partially overlapping libraries (i.e. *AP-side combination* as opposed to *data-side combination*). Machine learning algorithms are also likely to produce better results from a single, homogeneous training set than from a heterogeneous one. Furthermore, the library-by-library approach entirely separates the (time consuming and complex) process of model fitting from the procedure of library combination. So it we later wanted to reconsider our combination procedure (e.g. reweight the results, remove entirely libraries found to be problematic), this is much simpler in the library-by-library approach.

There are two ways we could apply the calibration information from the benchmark and reference stars. Either we apply these also library-by-library, or we produce a homogenous calibration set across the whole AP space. (Actually we would never homogenize the whole AP space, as hot/emission line stars, white dwarfs, brown dwarfs etc., are likely to remain separate from F,G,K stars.)

The combination algorithm need not be limited to GSP-Phot outputs, but could also include the AP estimates from ESP and GSP-Spec (although circularity must be avoided if the latter are using GSP-Phot AP estimates as priors).

2 Calibration with reference stars (Rene Andrae)

The performance of an algorithm using APs of reference stars just tells you about the difference between the two. It doesn't tell you how good the algorithm is in an absolute sense because the AP estimates of the AP reference stars themselves may have biases. Once he has got the calibration code and forward modelling code working, Rene will investigate variations of calibration scheme of CBJ-044, for example assuming that the calibration is an additive rather than multiplicative effect. **Action 1**

3 How other surveys have managed to calibrate/validate their APs (Caroline Soubiran)

The Bordeaux group on WP-811-20000 has now 5 poeple for 1.2 FTE. GBOG observations still on-going on NARVAL. Analyses of spectra with SME code started. Participation to LUMBA in Gaia-ESO survey will help to define the good inputs for SME.

RAVE and SEGUE have already faced the problem of calibrating APs. Several other surveys coming soon with same issue. RAVE 3rd release (Siebert et al. 2011) provides Teff, logg, [M/H], [α/Fe] for $\sim$40000 stars. Internal stability assessed with repeat observations. Comparison to external data with catalogue of 102 reference stars included in RAVE observations + 227 RAVE targets observed at high resolution. Mean offset and dispersion -63, 250 K in Teff, -0.1, 0.43 dex in logg. [M/H] calibrated as a function of derived metallicity, alpha enhancement, Teff, logg, S/N, large dispersion, not able to retreive high metallicties. SEGUE provides APs for 240000 faint stars. 4 papers on SSPP and its validation. Uses several different methods then decision tree to adopt final AP. Teff, log g, [Fe/H] at precisions of 141 K, 0.23 dex, 0.23 dex. Comparison to 125 SEGUE stars observed at high resolution gives external errors of 130 K, 0.21 dex, and 0.11 dex. Mean metallicity of 5 GCs + 3 OCs compared to literature values. SSPP improving with new external data, now 343 SEGUE targets observed at high S/N with aim of 500 but not analysed homogeneously. Residuals AP(SSPP)-AP(HR) show strong systematics, specially in [Fe/H]. (This may be an issues with either the AP estimation of the high resolution spectra – as suggested in the paper – but also in the SSPP algorithms.) Preliminary results of AMBRE on FEROS (R=48000) provided APs for 23000 stars. Comparison to literature is good but essentially limited to solar metallicity dwarfs (Worley et al. 2010). Offsets and dispersion also visible in literature values. Gaia-ESO survey plans to allocate 10 nights to calibrating clusters. Standard fields being defined. Gaia AP reference stars to be included in pointings when possible. Calibration of CU8 algorithm should use as many AP reference stars as possible. All previous surveys and calibration samples will be included in Gaia. Efforts done by other surveys useful for us.

A schedule has to be defined for producing the reference grid. Ulrike Heiter and Caroline Soubiran will inform CU8 how many stars could be analysed and used in the first run of Apsis. For this run, the APs of the reference stars don't need to be of very high accuracy and homogenity, as they will be improved prior to use in producing the final Gaia catalogue. **Action 2**

4 Calibration issues in CU6 (Laurent Chemin)

CU6 needs a set of objects stable in RV over the mission to align internal wavelength scale set by GSIS to absolute frame and to establish the zero-point of RVs. Done with asteroids (ideal but too rare at bright enough magnitude) and stars. Selection of 1420 candidates RV standard stars (Crifo et al. 2010) with $6<V<10.5$, uniform distribution on the sky, Hipparcos stars (homogeneous information on multiplicity, variability etc.), supposedly stable in RV. Stability at 300 m/s checked from ground-based obs + archived data. Pre-launch program to be completed in 2012, then one more measurement during operations. $\sim$3500 GBOG observations to date + 2300 from archives of echelle spectrographs. Less than 10% of the stars show variations above 300 m/s (mean time baseline 3.9 years). $\sim$ 75% stable at the 100 m/s level. M stars showing systematic difference of $\sim$ 500 m/s with catalogue of Nidever et al (2005) to be clarifed. Asteroids used to set all instruments on common scale. For commissioning $\sim$50 brightest stars around NEP observed with NARVAL including RVS range. Similar observations for SEP pending due to lack of adequate instrument.

Flux calibration of RVS does not seem on the agenda anymore. Variation of LST at each observation of a star render the flux calibration pratically impossible. 6 months solution provides averaged LSF and combined spectra. This might be a serious issue for GSP-spec.

5 How to improve the BP/RP simulations (Antonella Vallenari)

There are significant differences between synthetic spectra from different libraries (e.g. Phoenix and Basel): of up to 10% in relative flux. This may well explain the differences in performance seen in CHL-005 (GSP-Phot results) for these libraries. This is because Basel is a low resolution library (10Å sampling) whereas Phoenix (and MARCS) are high resolution libraries (1Å sampling). It is clear that the GSP-Phot algorithms perform better when they are trained on high resolution libraries, and that training on a low resolution library then testing on a high resolution one will give spuriously poor performance.

CU5: early in the mission CU5 plans to do a spectral calibration (internal and external) relative to the nominal instrument model (internal and external calibration). This is using the a_{ij} coefficients. This cannot deal with CTI correction and cannot handle crowding properly, but the PEM non-uniformity and gates are taken into account. Only later in the mission (1–2 years of data needed, or more), can the full forward model (AB-020) be applied (both internal and external calibration).

Note that the LSF results in strong correlations between spectral pixels. This needs to be taken into account in the calibration and in the AP estimation (the covariance is used explicitly in q-method, and at the least the impact on the uncertainties is taken into account in ILIUM).

5.1 The problem of interpolation

To date the random grids are produced by interpolating the nominal spectra (i.e. at the AP points used in the stellar models and spectral synthesis codes). This interpolation introduces both additional noise and biases of a few percent. This is small for the strong parameters (Teff and A_0), but catastrophic for the weak parameters ([Fe/H], logg). Fundamentally we should not actually need interpolation for the same of producing extra templates, as they provide no new information. It's also formally redundant as all of the AP algorithms are effectively doing interpolation anyway: SVMs do an inverse interpolation to get a smooth mapping from APs to data; the forward model of ILIUM and q-method for an explicit interpolation.

The very principle of supervised learning from templates relies on the **smoothness assumption**. This is simply the assumption that the fluxes vary smoothly with the APs over some AP length scale. (This is assumed for any method, including conventional line analyses of high resolution spectra, for example.) The relevant issue is, given the intrinsic sensitivity of the BP/RP data (i.e. for the given wavelength coverage, resolution and noise), how large is this length scale. In other words, what is the coarsest sampling of the AP space (from the stellar models) such that the AP variation between these template is smooth enough to be captured by the (possibly implicit) interpolation performed by the AP estimation algorithms.

We can test whether the sampling of any given (nominal) grid is sufficient in the following way. (1) Train – as necessary – an AP estimation algorithm on the nominal grid; (2) generate a set of synthetic spectra directly from the stellar models at AP points "intermediate" to those in the nominal grid; (3) predict the APs at those intermediate points. If the overall residuals are either smaller what we require or are smaller than what we believe is possible to achieve with teh BP/RP data, then we consider the nominal grid density to be sufficiently high (for that algorithm).

Frederic Thevenin and Andreas Korn will ask MARCS people whether something like this is possible, and if so, organize a call amongst the relevant CU8 people for the specifications. Phoenix can also be used for these tests because of its denser sampling. It can be degraded to the MARCS sampling and look at the effect of sampling. **Action 3**

Rene Andrae will, for GSP-Phot, identify the parts of AP space where a higher sampling density is required. **Action 4**

6 Hierarchical colours (Rosanna Sordo)

Rosanna has looked into using hierarchical colours from LL-079. See also AB-017 and PM-004. $\nu_{\rm eff}$ is a scalar parameter calculated from a BP/RP spectra which maps into Teff and A0 space. This could be used for diagnostic purposes.

7 Calibration and benchmark stars (Ulrike Heiter)

Benchmark stars: interferometry to get angular diameters (parallax – from Hipparcos – then gives radius); spectral energy distribution integrations, or single magnitude plus bolometric correction to get bolometric flux. Formal uncertainties are 15–80 K in $T_{\rm eff}$ 0.01–0.45 dex in $\log g$. Most benchmark stars are too bright to be observed by Gaia.

Editorial note: A calibration procedure using reference stars was proposed in CBJ-044 and tested in RKL-002 (see the presentation of Bailer-Jones for a quick summary). We may refer to this as a **data-side calibration procedure**, because it applies corrections to the training data before it is used in an AP estimation algorithm. In contrast is a **AP-side calibration procedure**, in which we apply a correction to the APs resulting from the estimation algorithm. A very simple approach of the latter was used in CBJ-043 Appendix A, where it was shown that it was not possible to correct for a systematic error in [Fe/H]. However, this was only a quick investigation, and making a correction as a function of multiple (*estimated*, not true) APs may be more successful.

Ulrike suggested that an AP-side calibration using these benchmark stars might be the preferable approach. Caroline Soubiran quoted one example of this as used by the RAVE team (see Caroline's presentation).

More tests should be done using a better observed library (e.g. MILES). A high quality library is being built with X-SHOOTER. The optimal number of reference stars for the calibration has to be estimatated.

8 What improvements should we make to the scope (diversity) and content (accuracy) of the spectral libraries and simulations? (Alessandro Lanzafame)

8.1 AP estimation via fixed, observed templates

Many stellar phenomena – such as rotation, activity, veiling, lithium depletion, accretion disks, winds, jets – cannot be modelled accurately enough to permit the production of accurate synthetic spectra. Consequently, there are many types of stars (especially young or cool stars) for which we do not have a sufficiently dense or accurate grid of synthetic spectra described by a small number of APs. This makes it difficult to use a standard supervised learning approach to learn the mapping between data and these APs.

A slightly alternative approach is to assign APs to unlabelled objects using a (small) set of fixed objects with known APs ("the templates") which are observed by Gaia (so there is no spectral modelling involved). Only those templates near to the unlabelled object are used in the assignment (generally via some local smoothing or interpolation; in the simplest case this is just a distance-weighted average). This is still a supervised learning method in the sense that APs are assigned to unlabelled objects using the APs of a number of labelled templates. But conceptually it is different because each unlabelled object becomes explicitly and associated with a small number of (local) templates. The objective is to decouple the issue of how we assign templates (and therefore APs) to an unlabelled object, from the issue of assigning APs to the templates in the first place. Procedurally the first issue ("association") is a nearest neighbours classification problem, and the second ("modelling") is the physical or modelling problem of how to assign APs to the templates. The first involves just the data and not the science of the objects, and vice versa for the second. Once the first problem is solved, an unlabelled object becomes permanently assigned to a set of templates, regardless of the solution to the modelling problem. So as our understanding of the science of the objects improves, we reassign APs to the templates and this propagates to updated APs of the unlabelled objects, but without having to go back to the original Gaia data. This is unlike the standard supervised approach, which would require us to retrain the AP estimation algorithms and reapply them to the unlabelled objects (although this is not infeasible and is already envisaged in CU8).

Editorial notes:

- This approach is not fundamentally different from "normal" supervised learning, as we may still have to refit the local smoothing models in the association step when the template APs are reassigned. However, the discrete template methods does permit a more complete decoupling of the association and modelling steps. It also recalls the ideas of MK classification, in which the classifications are defined by permanent standards, the physical interpretation of which may change over time.
- Local interpolation suffers from the well-known *curse of dimensionalty* problem, which is why global interpolation schemes have been so popular and successful in machine learning. Note, however, that the local interpolation is done in the AP space, which probably has dimensionality of a few rather than the data space (which is of order 10^2 for Gaia BP/RP and RVS).
- This method is not entirely unlike a classification method using discrete templates. One (Bayesian) variant of this is described by Bailer-Jones & Hogg in CBJ-061.
- The method also recalls the general idea of *archetypal analysis* by Cutler & Breiman (1994; Technometrics, vol. 36, pp. 338–347), in which projections onto a limited set of templates (archetypes) are used to represent the data.

Antonio Frasca in Catania has implemented a version of this idea in IDL for small datasets. The application to large datasets would require the implementation of some efficient algorithm for the association step. An unsupervised (clustering) method could be used for the association step to identify the "optimal" set of templates in each case. Frederic Thevenin recalled the Minimum Spanning Tree used in CU4 for asteroid classification presented by Christophe Ordenovic and Laurent Gallucio at the CU8 classification meeting in February 2011 at MPIA, Heidelberg.

9 How could we use non-Gaia survey data to improve the calibrations? (Frederic Thevenin)

Use asteroseismology data to estimate $\log g$ of a set of reference stars, perhaps to better than 0.04 dex. This also gives radius and density to a few percent, from the mean seismic variables (see Creevey et al. 2011 A&A). From individual frequencies, we can derive other fundamental quantities too. There are currently 1000 red giants from Kepler (public data) with V=7 to 12, and also Corot data too.

The convective shift has to be taken into account in the RV measurements. Grids of corrections as a function of APs are now being provided. These are mandatory for CU6, could also be used for CU8.

NLTE effects in the RVS range must be taken into account for GSP-spec. Computation of synthetic spectra with NLTE very complicated. Work is on-going (Korn, Lanzafame).

Frederic will think of a better link of FLAME to and to GSP-phot/q-method. **Action 5**
Editorial note: It is in principle straight forward to generalize q-method to work directly in terms of the primary parameters like mass and age, although in practice is non-trivial.

Note that CU5 and CU7 provide information on flux variability. This is important additional information, since variable stars populate the instability strip in HR- diagram. Ideally GSP-Phot would use that information.
Editorial note: This is not currently planned in GSP-Phot (although it is planned for DSC), and may be better done as post-DPAC processing.

10 Improving the CU8 stellar spectral libraries (Andreas Korn)

SWOT analysis of Gaia spectral libraries. Little progress in the past two years reflected in current/upcoming simulations. HRD coverage essentially complete, but some parameter extensions needed to optimize the codes parametrization (e.g. carbon supergiants, PMS stars, $[\alpha/Fe]$, spacing). Strength: DPAC expertise to compute all but coolest objects. Weaknesses: classical models too simplistic, lack manpower and have inconsistencies between subgrids. Opportunities to take with other surveys, e.g. Gaia-ESO, to test capabilities. One identified risk is the difficulty to guarantee in the long-term the capacity to produce specific updates or extensions of grids.

11 Status of ground-based observations (Ulrike Heiter)

Observations for candidate benchmark stars: HR spectra for 75% with HARPS, SARG and NARVAL + Nearby Stars programme (Luck and Heiter 2007) for 10 stars + many archived spectra available.

Observations for candidate AP reference stars: 30 'good' field stars from PASTEL + 47 stars in eight open clusters. Spectra of a same star on different instrument exhibit large scale features. Analysis has to be made on small segments. AK informs us that a large number of COROT targets have been observed with NARVAL and will be published soon.

12 The SME code to analyse reference stars (Thierry Jacq)

Spectroscopy Made Easy (SME) calculates synthetic spectra of stars and fits them to observed spectra. The first release was by Valenti and Piskunov (1996). Needs IDL and external shared library. Requires atomic and/or molecular line data in VALD format. SME job defined through GUI or script, with input and output in IDL structures. Execution time from few second to a few days depending on number of lines synthesized and free parameters. Now working on acceptable line lists and corresponding masks, and defining procedures for partial automation. Problem with SME code not fully polished.

13 Optimization of line lists for the analysis of reference stars (Nathalie Brouillet)

Work is being done on NARVAL spectra of Arcturus and the Sun with several line lists available in order to define the one most suited for the analysis of AP reference stars already observed with that instrument. This involves much tedious work of visualization to keep or reject lines, and select continuum windows. Definition of good fit to be refined as well as objective criteria to select enough lines for a high quality analysis. Another goal is to select select more lines in Sun and Arcturus which are a priori rejected because they are strongly damped, but which become more relevant in metal poor stars.

14 The status of the ground-based data for M stars, brown dwarfs and OBA stars, for calibration purposes in CU8 (Yves Fremat)

For UCD, three databases are available for MLT stars (Leggett, Keck LRIS, SpeX Prism) with T_{eff} spanning 700 to 3000 K, and logg from 3.0 to 6.0 for 63 stars with known APs and LR spectrum above 6000 Å (20 have known parallax). On the hot side, several criteria have been used to build a list of potential calibration OBA stars: not binary, not variable, 8–10 mag, parallax uncertainty lower than 30%, echelle spectrum available. This gives 54 A stars with known APs, 33 with a good parallax, 9 with 1 echelle spectrum. For OB stars, the numbers are respectively 175, 9, 16. A reference sample per instrument configuration might be needed (e.g. RVS data HR/LR, BP-RP ?). SEP/NEP data should be exploited.

15 Calibration of OBA stars (Ronny Blomme)

The VLT Flames Survey of Massive Stars by Evans et al. (2005) is a very good dataset for calibration of OBA stars. Stellar parameters determined by Dufton et al. (2007), with good coverage in (Teff, logg) for OB, but there is a lack of A stars. The Gaia-ESO Survey will include open clusters with massive stars. Expect about 1500 stars. Lots of science topics with OBA stars in GES. The Evans et al. sample also includes LMC-SMC stars with different abundances than MW. Complication around the Halpha line which can be variable.