The Stellar Seismic Indices (SSI) data base : Architecture of the pipeline and date base

Réza Samadi, Raphaël Peralta, Mahfoudh Abed, Christian Renié

December 19, 2013

Contents

1 Introduction

Stars are cavities in which waves can propagate and form standing waves (modes). The frequencies of these oscillation modes depend on the internal structure of the star. Their detection and measurement enable us to probe the inner layers of the stars. Oscillations were detected in the Sun for the first time more than 50 years ago. Similar oscillations (thus named solar-like oscillations) were detected in other stars at the end of the nineties. Thanks to CoRoT (CNES) and Kepler (NASA) seismic observations, it is now possible to detect many solar-like oscillations in a very high number of stars. The observed oscillation spectra can be quite complicated to analyse and interpret. Nevertheless, these spectra show characteristic pattern, which can be more simply characterized with what we call *Stellar Seismic Indices*. More specifically, the stellar seismic indices correspond to some characteristic global seismic numbers that are extracted from the oscillation spectra of solar-like pulsating stars. Three main indices are today more and more used:

- the peak frequency, $\nu_{\rm max}$: this is in general the frequency where the oscillation spectrum peaks in the power density spectrum (PDS).
- the mean large separation, $\Delta\nu$: this quantity corresponds to the mean frequency spacing between two consecutive p-modes (with same angular degree).
- the mixed modes period spacings, $\Delta\Pi$: this quantity corresponds to the mean period spacing between two consecutive mixed modes.

The two first seismic indices obey characteristic scaling relations that depend directly on the radius, mass and effective temperature of the star. From the knowledge of the effective temperature and the measure of these two seismic indices, it is then possible to estimate the mass and radius of the star. The third one is expected to

indirectly depend on the mean density of the core and enable us to distinguish between core burning red giants (clumb stars) from stars ascending the read giant branch.

Because they are relatively easy to measure thanks to CoRoT and *Kepler*, it has already been possible to measure seismic indices in about 10 000 red giants and sub-giants, and subsequently to derive their mass, radius and evolutionary status. Theses seismic indices are more and more used in the stellar physic but also for the study of Galactic population (see the review by Belkacem 2012; Belkacem et al. 2013). These seimic indices have opened the way toward what we name now *ensemble asteroseismology*. In this context, we aim at developing a data base containing Stellar Seismic Indices (SSI) derived in a *standard* and *homogeneous* way for a very large set of CoRoT and Kepler targets.

There are actually various ways to define and extract the SSI (see e.g. Verner et al. 2011; Hekker et al. 2011). A study lead by Raphael Peralta (PhD thesis, LESIA) has been undertaken in order to determine the optimal algorithms for extracting these seismic indices from the oscillation spectra. This study is not yet fully completed but converge toward a preliminary choice for the optimal algorithm. Since this study is still ongoing, we will only briefly present the retained algorithms and will rather more focus on the architecture of the pipeline and the structure of the data base.

This document is divided in two parts: the description of the pipeline (Section 2) and of the data base (Section 3).

2 The pipeline

2.1 General description

The main goal of the pipeline is to extract a set of seismic indices from stellar light-curves acquired with CoRoT and *Kepler*. The main inputs of the pipeline will be light-curves (one or several per target) extracted either from the CoRoT base at IAS (`http://idoc-corotn2-public.ias.u-psud.fr/`) or the *Kepler* KASOC data base (`http://kasoc.phys.au.dk/`). A pre-processing of the light-curves will be performed (see Section 2.2), which will provide as output the light-curves corrected for different observational artifacts and long term effects, as well as the associated Power Density Spectrum (PDS). The pre-processing will also have in charge the concatenation of the light-curves acquired at different periods for the same target. The seismic indices will then been extracted from the pre-processed light-curves or the PDS on the basis of different algorithms (see Section 2.3). The figure 1 shows the main data flows and the different data bases handled by the pipeline.

The pipeline is coded with Python language and relies on the following libraries:

- `Numpy`: fundamental package for scientific computing with Python (`http://www.numpy.org/`) ;
- `Scipy`: scientific computing tools for Python (`http://www.scipy.org/`) ;
- `FFTW3`: library for computing discrete Fourier transforms (`http://www.fftw.org/`) ;
- `NFFT`: library for computing nonequispaced discrete Fourier transforms (`http://www-user.tu-chemnitz.de/~potts/nfft/`) ;
- `PyNFFTLS`: a fast algorithm for the Lomb-Scargle periodogram (`http://pypi.python.org/pypi/pynfftls/`).

The pipeline has a multiprocessing capability (the stars are processed in parallel).

2.2 Pre-processing

The pre-processing of the input light-curves consists in handling the following perturbations:

- outliers (e.g. proton impacts): Proton impacts can occur while monitoring a target. This is particularly the case when CoRoT crosses the South Atlantic Anomaly. Most of the outliers are, however, removed by the CoRoT and KASOC pipeline[1].

[1]The corrections applied on the CoRoT light-curves are described in Samadi et al. (2007) while those performed at KASOC on the *Kepler* data are documented in García et al. (2011).

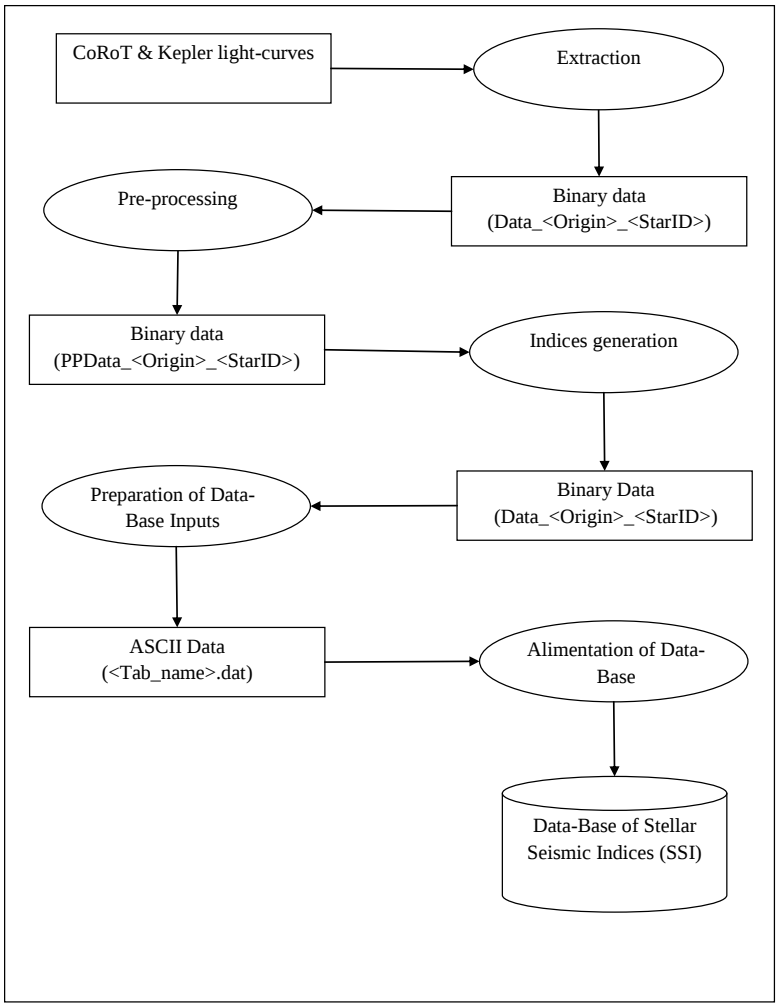

Figure 1: Data flows

- large jumps: big jumps occur some times in the light-curves. They are usually due to the occurrence of a bright pixel, which is typically generated by a particularly energetic proton impact. However, the origin of such jumps are not always clear.
- gaps: the interruptions occurring during an observations sequence introduce aliases in the spectrum, which can alter the determination of the seimic indices.
- flagged measurements: doubtful measurements are in general flagged. For instance for CoRoT, measurements considered as outliers (e.g. proton impacts) are flagged as doubtful measurements by the pipeline. The flagged data are often periodic (for instance the crossing of the South Atlantic Anomaly by CoRoT is periodic). Discarding these measurements then produces aliases in the Fourier spectrum. It is therefore required to replace the flagged measurements by some representative values.
- merging together several light-curves: light-curves acquired during different observing sequences are not comparable in intensity. Therefore, it is not possible to build a single light-curve just by joining together the individual light-curves. It is required to match the individual light-curves in an appropriate manner.
- long term effects: The global gain of an instrument is expected to vary with the temperature. In addition to the variation of the global gain with the temperature, we face long-term decrease of the intensity because of the aging of the optics and of the CCD. All these variations occur at long time-scale (from

weeks to several months). These variations are corrected by the recent version of the CoRoT pipeline, however, some residuals remain, which must be corrected.

Once the individual light-curves are processed and merged together, the pre-processing module computes the Power Density Spectra (PDS) associated with the merged light-curve. The time series are in general non-equispaced in time. Accordingly, the PDS is computed using the Lomb-Scargle periodogram, which contrary to the Fast Fourier Transform does not require equispaced measurements. Among the fast available algorithms, we consider the one proposed by Leroy (2012), which relies on the NFFT algorithm (`http://www-user.tu-chemnitz.de/~potts/nfft/`). A Python version of this algorithm named `PyNFFTls` is available at `http://pypi.python.org/pypi/pynfftls/`.

2.3 Generation of the seismic indices

2.3.1 Extraction of the mean large separation, $\Delta\nu$

Depending on the type of stars, we derive two different versions of the mean large separation. For all type of stars, we determine $\Delta\nu$ on the basis of the auto-correlation function of stellar oscillation times series (ACF method hereafter). This method was originally proposed by Roxburgh & Vorontsov (2006) and studied and developed by Mosser & Appourchaux (2009). For red giants, we will derive in future a more precise value from the universal oscillation pattern as proposed by Mosser et al. (2011). At the present time only the ACF method is implemented in the pipeline.

We implemented the ACF method in a similar way as Mosser & Appourchaux (2009). We here briefly describe the implemented algorithm: We multiply the PDS by a filter centred at a given frequency ν_c (not to be confused with the star cut-toff frequency). The filter is a cosinus function with a width equal to NDnu $\times\Delta\nu_{\rm scl}(\nu_c)$, where NDnu is an integer (typically three for red giants) and $\Delta\nu_{\rm scl}(\nu_c)$ is the scaling relation $\Delta\nu(\nu_{\rm max}) = \alpha\,\nu_{\rm max}^{\beta}$ where the coefficients α and β are setup according to the type of star (see e.g. Mosser et al. 2013, and reference therein). We then compute the inverse Fourier transform of the product between the PDS and the filter. This gives an auto-correlation function that we normalize following Mosser & Appourchaux (2009) and note $A(\tau,\nu_c)$, where τ is the correlation time-length and ν_c is the central position of the filter. When the filter lies in the frequency domain of solar-like oscillations, $A(\tau,\nu_c)$ peaks around $2/\Delta\nu$. We then search the maximum reached by $A(\tau,\nu_c)$. We name $A_{\rm max}(\nu_c)$ this maximum and $\tau_{\rm max}$ the associated correlation time-length. However, to prevent catching the wrong maximum, we search the maximum in the interval [$2/\Delta\nu_{\rm scl}$/RAT1,$2/\Delta\nu_{\rm scl}$*RAT1], where the ratio RAT1 is fixed to 1.3. We redo such calculation for NFilter filters with centers ν_c ranging from a minimum frequency (Nu0) up to a maximum frequency (Nu1). We consider $A_{\rm max}$ as a function of ν_c and determine its maximum value. This maximum is reached for a given ν_c, which is considered as our first guess for $\nu_{\rm max}$ ($\nu_{\rm max}^{(0)}$ hereafter).

We then consider NFilter filters centred around our first guess $\nu_{\rm max}^{(0)}$ and ranging in the interval [RAT2$\times\nu_{\rm max}^{(0)}$, RAT3$\times\nu_{\rm max}^{(0)}$]. This provides a new estimate of $\nu_{\rm max}^{(1)}$ and a more precise determination of $\Delta\nu^{(1)}$. The value of $\nu_{\rm max}^{(1)}$ is used as a guess for the determination of $\nu_{\rm max}$ from the fit of the PDS background (see Section 2.3.2 below). This provides a more accurate value of $\nu_{\rm max}$ ($\nu_{\rm max}^{(2)}$ hereafter). We then perform a new calculation of $\Delta\nu$ by considering 2$\times$ NFilter centred around $\nu_{\rm max}^{(2)}$ and ranging in the interval [RAT2$\times\nu_{\rm max}^{(2)}$, RAT3$\times\nu_{\rm max}^{(2)}$]. This yields the final estimate $\Delta\nu^{(2)}$. The 1-σ uncertainty associated with the determination of $\Delta\nu$ is computed as explained in Mosser & Appourchaux (2009).

2.3.2 Extraction of the peak frequency, $\nu_{\rm max}$

The peak-frequency $\nu_{\rm max}$ is obtained using a method commonly used with various variants (see e.g. Verner et al. 2011; Hekker et al. 2011; Mathur et al. 2011, and references therein). In such a method, the smoothed PDS is fitted with an empirical model composed by several components. Our implementation considers three components:

$$F(\nu) = \frac{P_{\rm gran}}{1+(2\pi\tau_{\rm gran}\nu)^{\alpha_{\rm gran}}} + A_{\rm env}\,\exp\left[-4\ln(2)\left(\frac{(\nu-\nu_{\rm max})}{\delta_{\rm env}}\right)^2\right] + W \qquad (1)$$

where the first component is an Harvey-like model representing the contribution of the stellar granulation, the second component is a Gaussian function modeling the envelope of the solar-like oscillation spectrum, and the

third component represents the white noise at high frequency. In Eq. 1, $\tau_{\rm gran}$ is the characteristic timescale of the granulation, $P_{\rm gran}$ the height of the granulation component, $A_{\rm env}$ the amplitude of the envelop, $\delta_{\rm env}$ its full width at half maximum (FWHM), and W a constant. The guess values for $A_{\rm env}$ and $\delta_{\rm env}$ are taken from the scaling relations obtained by Mosser et al. (2010) using $\nu^{(1)}_{\rm max}$, while those for $P_{\rm gran}$ and $\tau_{\rm gran}$ are determined from the scaling relations derived by Mathur et al. (2011) using $\nu^{(1)}_{\rm max}$. Finally, the guess value for W is obtained by computing the median of the PDS around $3\,\nu^{(1)}_{\rm max}$ and over an interval of size $6\,\Delta\nu^{(1)}$. Prior doing the fit, the PDS is smoothed by averaging over a width of size $3\,\Delta\nu^{(1)}$. The fit is performed by means of the least-squares and yields our final estimate $\nu^{(2)}_{\rm max}$ with its associated 1-σ uncertainty. We keep and store in the data base the estimated values of $P_{\rm gran}$, $\tau_{\rm gran}$, and $\alpha_{\rm gran}$. We also compute and store the signal to noise ratio (SNR) defined as:

$$SNR = \frac{A_{\rm env}}{\mathcal{P}_{\rm gran}(\nu^{(2)}_{\rm max}) + W}\,, \tag{2}$$

where $\mathcal{P}_{\rm gran}(\nu)$ is the granulation component (*i.e.* the first term in the RHS of Eq. 1).

2.4 Post-processing

The post-processing consists of the analysis of the generated seismic indices and their insertion into binary files. The analysis aims at determining the validity of each seismic index. Valid seismic indices are those for which `status` is equal to zero, otherwise `status` takes one of the values given in table 1.

Table 1: Values taken by the `status` of the `indice` table.

Bit	Value	Explanation
1	0	valid indice
1	1	$A_{\rm max}$ is lower than 8
2	2	the envelop component failed to converge
3	4	the granulation component failed to converge
4	8	the fit failed to converge
5	16	the degree of freedom of the fit is lower than 100
6	32	some errors occurred when processing the star
7	64	some errors occurred when *pre*-processing the star

2.5 Inputs

The inputs of the pipeline are either CoRoT or *Kepler* light-curvse. Three different types of CoRoT light-curves are handled:

- `AN2_STAR`: light-curve from the asteroseismology channel (sampling: 32s) ;
- `EN2_STAR_MON`: monochromatic light-curve from the Exo-planet channel (sampling 32s, 512s, or a mixed of both cadences) ;
- `EN2_STAR_CHR`: chromatic light-curve from the Explonat channel (sampling 32s, 512s, or a mixed of both cadences).

Note that we consider only the white flux of the chromatic light-curves (*i.e.* the sum of the three colors). The formats of the CoRoT light-curves are described in the document available at `http://idoc-corotn2-public.ias.u-psud.fr/jsp/doc/DescriptionN2v1.5.pdf`

The *Kepler* light-curve contain two type of fluxes: `SAP_FLUX` and `PDCSAP_FLUX`. The former corresponds to the flux acquired with the optimal aperture pixels while the latter corresponds to the flux corrected for instrumental perturbations (see *Kepler* Archive Manual and *Kepler* Data Processing Handbook). We consider the corrected flux. The cadence is either $\sim$ 1 mn (short candence) or $\sim$ 30 mn (long candence). Both cadences are not mixed in the same file.

3 Structure of the data base

The date base contains five tables (see Figure 2) with the following meaning:

- `Star`: characteristics of the stars (e.g. name, origin, position, magnitude, spectral type ...) ;
- `Indice`: the seismic indices (e.g. value, precision) ;
- `File`: characteristics of the set of files (light-curves) that have been merged together for a given target ;
- `Code`: versions of the different components of the pipeline ;
- `Use`: this table ensures the links between the indices and the files.

The data base is developed under `postgresql`, an open source data base system (`http://www.postgresql.org/about/`). We only describe below the tables named `Star` and `Indice`. A more complete description will be given in a forthcoming document.

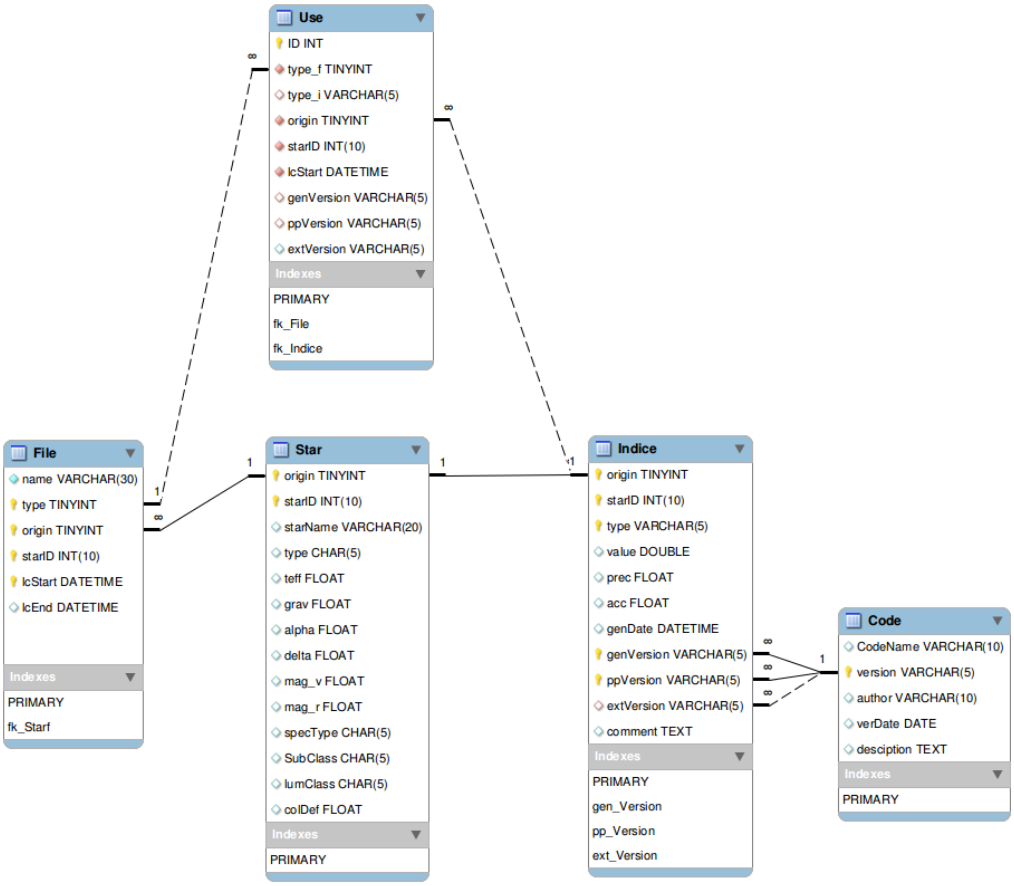

Figure 2: Structure of the date base

3.1 `Star` table

This table contains informations related with the star:

- `Origin`: origin of the data ;

  = 0 : for an unknown origin

= 1 : for CoRoT data

= 2 : for Kepler data

= 3 : for OGGLE data

- `starID`: the star ID (CoRoTID in the case of a CoRoT target, and the KIC number in the case of a Kepler target) ;
- `type`: type of light-curve (depending of the instrument, the light-curves can be of different type. For instance for CoRoT: monochromatic or chromatic light-curves, or light-curves from the asteroseismology channel) ;
- `grav`: gravity of the star in log decimal ;
- `alpha`: right ascension (equatorial coordinate) ;
- `delta`: declination (equatorial coordinate) ;
- `mag_v`: magnitude in the V band (depends on the origin and type of the light-curve) ;
- `mag_r`: magnitude in the R band (depends on the origin and type of the light-curve) ;
- `specType`: spectral type ;
- `subClass`: spectral type sub-class ;
- `lumClass`: luminosity class.

3.2 `Indice` table

This table contains the seismic indices (values and associated data):

- `Origin`: as in the `Star` table ;
- `starID`: as in the `Star` table ;
- `type`: name of the seismic indice (see below) ;
- `value`: value of the seismic indice (Negative if unknown) ;
- `prec`: precision (1-σ uncertainty) with which the seismic indice was measured (Negative if unknown) ;
- `acc`: accuracy with which the seismic indice was measured (Negative if unknown) ;
- `status`: status of the indice, `status`=0 for a valid indice otherwise a non zero value ;
- `genDate`: generation date ;
- `genVersion`: version of the generation pipeline ;
- `ppVersion`: version of the pre-processing ;
- `extVersion`: version of the extraction program ;
- `comment`: any comment associated with the indice.

The names of the currently available seismic indices are:

- `Deltanu`: the mean large separation $\Delta\nu^{(2)}$ [in μHz] obtained by the ACF method (see Sect. 2.3.1) ;
- `Amax`: the maximum of the auto-correlation function ;
- `numax`: the peak frequency $\nu^2_{\rm max}$ [in μHz] obtained by the envelop fitting (see Sect. 2.3.2) ;
- `A_env`: the amplitude of the envelop [in ppm^2/μHz] ;

- `tau_gran`: the granulation timescale [in sec] ;
- `P_gran`: the granulation power [in ppm^2/μHz] ;
- `alpha_gran`: the slope of the granulation component ;
- `snr`: the characteristic signal-to-noise ratio (see Eq. 2) ;

References

Belkacem, K. 2012, in SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. S. Boissier, P. de Laverny, N. Nardetto, R. Samadi, D. Valls-Gabaud, & H. Wozniak, 173–188, arXiv:1210.3505

Belkacem, K., Samadi, R., Mosser, B., Goupil, M.-J., & Ludwig, H.-G. 2013, in Progress in physics of the sun and stars: a new era in helio- and asteroseismology, Vol. to appera in ASP Conference Series

García, R. A., Hekker, S., Stello, D., et al. 2011, MNRAS, 414, L6

Hekker, S., Elsworth, Y., De Ridder, J., et al. 2011, A&A, 525, A131

Leroy, B. 2012, A&A, 545, A50

Mathur, S., Hekker, S., Trampedach, R., et al. 2011, ApJ, 741, 119

Mosser, B. & Appourchaux, T. 2009, A&A, 508, 877

Mosser, B., Belkacem, K., Goupil, M. J., et al. 2011, A&A, 525, L9

Mosser, B., Belkacem, K., Goupil, M.-J., et al. 2010, A&A, 517, A22

Mosser, B., Samadi, R., & Belkacem, K. 2013, ArXiv e-prints

Roxburgh, I. W. & Vorontsov, S. V. 2006, MNRAS, 369, 1491

Samadi, R., Fialho, F., Costa, J. E. S., et al. 2007, ArXiv Astrophysics e-prints

Verner, G. A., Elsworth, Y., Chaplin, W. J., et al. 2011, MNRAS, 415, 3539