Astronomy & Astrophysics manuscript no. output
July 30, 2024

Stellar Activity in Stars with Exoplanets

Presenting the SAITAMA pipeline

Telmo Monteiro[1], S. G. Sousa[2,3], and J. Gomes da Silva[2,3]

[1] Department of Physics and Astronomy, University of Porto, Rua do Campo Alegre 1021 1055, 4169-007 Porto, Portugal
e-mail: up202308183@up.pt
[2] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal
[3] Supervisors of the PEEC

July 30, 2024

ABSTRACT

Aims. In this work, we aimed to develop the SAITAMA pipeline to characterize the stellar activity of stars with exoplanets using spectral data spread in a wide time range. This pipeline is meant to be an easy way to obtain important information about the spectral activity of a star, processing automatically the spectra from ESO's online database.
Methods. We used a total of 18 stars included in the SWEET-Cat catalogue to test the pipeline. SAITAMA's core lies in the ACTIN2 tool, that computes the spectral activity indices for a stellar spectrum, and it uses ACTIN2 to compute activity indices for the CaII H&K, Hα and NaI lines. Through periodogram analysis, the pipeline computes the activity S_{CaII} period. We also converted S_{CaII} to the S_{MW} and $\log R'_{HK}$ indices, as well as estimated the rotation period and chromospheric age of the stars. The pipeline provides statistical information on the activity indices. During the creation of the pipeline, we analysed the spectra used to ensure its proper functioning.
Results. The SAITAMA pipeline was born, providing an easy and quick way to estimate the spectral activity of stars with exoplanets. The pipeline can be tuned by the user to retrieved different activity indices (as long as they are included in ACTIN) for different stars and instruments. Nevertheless, for now only the HARPS, ESPRESSO and UVES spectrographs are configured.

Key words. stellar activity – exoplanet research

1. Introduction

This report describes the work done in the context of the PEEC (Extracurricular Internships Program, in English) "Stellar Activity in Stars with Exoplanets", offered by the Faculty of Sciences of University of Porto in 2024.

The characterization of stars is a crucial step for the study of many astrophysics topics, in particular for the characterization of planetary systems. Specifically, the stellar activity is important to understand the environment of exoplanets, the age of planetary systems and their birth environment. This characteristic also interferes with the detection and characterization of exoplanets using either the RV method or transits, as well as in the characterization of exoplanet atmospheres via transmission spectroscopy. Stellar activity is not a static parameter, and it varies in a time scale that we need to understand for each star. The method to derive stellar activity, through the measure of specific activity spectral indices, has already a tool implemented, in the form of ACTIN[1] (Gomes da Silva et al. (2018), Gomes da Silva et al. (2021)).

In this work, we developed the Stellar ActivITy AuToMatic cAlculator (SAITAMA[2]) pipeline, to characterize the stellar activity of stars with exoplanets. Many of these stars have a significant amount of spectroscopic data which were taken with the goal of finding exoplanets. This spectral data is spread in a wide time range for several stars and can be used to better characterize the stellar activity. SAITAMA ultimate goal is to complement SWEET-Cat (Santos et al. (2013); Sousa et al. (2021)) with this relevant information for the planet-host stars for which we have good spectroscopic data. For this, given a star identifier and the spectrograph name, the pipeline downloads the best spectra from ESO's data base, corrects them by the radial-velocity (RV) and measures specific activity spectral indices with ACTIN. Additionally, the pipeline converts the indices to common literature ones and computes all the relevant statistics, including activity periods through GLS periodograms, chromospheric rotation periods and chromospheric ages, through literature calibrations. The pipeline is available in a public GitHub repository[3].

New links between the properties of known exoplanets and their host stars's properties can revolutionize the understanding of exoplanetary systems. The metallicity correlation identified for giant planets was the first and the most well known link between planets and their host stars. An observation that was soon supported by theoretical prepositions such as the core-accretion idea to explain the formation and evolution of the newly discovered and unexpected giant planets. Several other correlations have been identified and proposed. Some have been confirmed, but most are still disputed. The compiled information about the stellar activity provided by this pipeline can then be used to find correlations between this characteristic and the properties of the planets hosted by these stars.

[1] https://github.com/gomesdasilva/ACTIN2/
[2] Inspired by the main character of "One-Punch Man", a Japanese superhero manga series.
[3] https://github.com/telmonteiro/SAITAMA/

This report is organised as follows. In Sect. 2 we discuss the most popular chromospheric activity indices, ACTIN and a preliminary study using the stars studied in Pepe et al. (2011). The pipeline architecture is explained in Sect. 3. In Sect. 4 we describe the spectral data acquisition module of SAITAMA, while in Sect. 5 we exoplore the data processing and analysis module, which composes the bulk of SAITAMA. In Sect. 6 we describe the products of SAITAMA and in Sect. 7 we present the main results for the test stars. Our concluding remarks follow in Sect. 8. Appendices A and B describe extra studies done on UVES spectra and explore various methods to compute the activity period error, respectively.

2. Chromospheric activity indices

Stellar activity can be detected mainly by spectroscopy or photometry. Rotationally modulated active regions with different contrast from the surrounding photosphere result in variability of the brightness of a star. This activity phenomena together with the inhibition of convection by magnetic fields, which destabilises the granulation pattern, can affect the stellar spectrum. For example, in the visible, there are lines such as CaII H&K and Hα which are sensitive to the chromospheric temperature rise. In the presence of strong magnetic fields, these lines enter in emission and can be used as proxies of magnetic activity.

2.1. Ca II H&K lines

It is known for a long time that the flux on the CaII H&K lines has a direct relationship with the number of active regions in the Sun (Baliunas et al. 1995), and therefore these lines are the most used activity proxies. The use of the CaII H&K lines as an activity index was made popular among activity researchers by the Mt. Wilson "HK program", a project aimed at measuring the long-term activity of solar-like stars which was started in 1966 by O. Wilson (Wilson 1978). Vaughan et al. (1978) introduced the S-index, a dimensionless proxy for the CaII activity measured by the Mt. Wilson Observatory spectrometers. This index was based on the flux integrated in 1.09 Å passbands centred on the H&K lines and normalised to the flux in two 20 Å wide surrounding pseudo-continuum bands centred at 3901 (V band) and 4001 (R band) Å. The S-index can be defined as

$$S = \alpha \frac{F_H + F_K}{F_B - F_V}, \tag{1}$$

where α is a calibration constant, F_H and F_K the fluxes in the H and K lines, and F_B and F_V the fluxes in the two reference lines (Boisse et al. 2009).

The S-index can be used to control the variations of activity of a given star. However, when the S value of stars of different spectral type is compared, one needs to consider the colour and photospheric contributions to the index. Stars of different spectral types have different levels of flux on these spectral regions, affecting the S value. To remove the colour dependence and the photospheric component, Middelkoop (1982) and Noyes (1984) developed a transformation of the S-index into a value R'_{HK} which is a function of B-V and is normalised to the bolometric flux. The photospheric and bolometric corrected R'_{HK} chromospheric emission ratio can be computed from the S_{MW} (index in the Mount Wilson scale) via the Noyes (1984) expression

$$R'_{HK} = R_{HK} - R_{\rm phot}, \tag{2}$$

where

$$R_{\rm HK} = 1.34 \times 10^{-4} C_{\rm cf} S_{\rm MW} \tag{3}$$

is the index corrected for bolometric flux (Middelkoop 1982), $C_{\rm cf}$ is the bolometric correction, and $R_{\rm phot}$ is the photospheric contribution. The photospheric contribution is a function of $B-V$ colour and has been derived by Noyes (1984) as

$$\log R_{\rm phot} = -4.898 + 1.918(B-V)^2 - 2.893(B-V)^3. \tag{4}$$

2.2. Hα line

The CaII H&K lines are easily accessible in FGK dwarfs, but when we consider cooler stars, the energy distribution starts to move to redder wavelengths and the signal-to-noise ratio decreases drastically in these lines. An alternative has been the use of other spectral activity proxies at longer optical wavelengths, for example the core of the Hα line at 6562.808 Å.

A positive correlation between the chromospheric fluxes of the CaII H&K and Hα lines has been suggested in the literature, but the majority of the studies where this relation was observed used averaged fluxes for both lines which were not obtained simultaneously.

The flux in CaII H&K and Hα is known to follow the solar activity cycle, and to correlate well with sunspot number and other activity diagnostics. However, for other stars, the flux in these lines is known to have a wide range of correlations, increasing the difficulty in the interpretation of the signals observed with the Hα line. Gomes da Silva et al. (2022) investigated the effect of the Hα bandpass width on the correlation between the CaII and Hα indices with the aim of improving the Hα index. They found that calculating the Hα index using a bandpass of 0.6 Å maximises the correlation between CaII and Hα, both at short and long timescales. On the other hand, the use of the broader 1.6 Å, generally used in exoplanet detection to identify stellar activity signals, degrades the signal by including the flux in the line wings.

2.3. NaI D1 and D2 lines

The NaI D1 and D2 resonance lines can be observed in the spectra of all stellar types. However, for cooler stars the doublet starts to develop strong absorption wings. For the most active stars, chromospheric emission in the core of the D lines becomes visible, which is an indication of collision-dominated formation processes. Díaz et al. (2007) studied different features of the D lines and defined an index N similar to the Mount Wilson S index: they divided the flux in the core of the D1 and D2 lines by the flux in two redder and bluer pseudo-continuum reference bands. The authors found that when the colour dependence of N and S is taken into account, the correlation between both indices varies from tightly correlated for some stars to cases of no correlation. However, the two indices are well correlated for active stars with emission in the Balmer lines. They conclude that the N index can be useful when comparing the activity variations of individual stars, mainly for later types where little emission is observed in the CaII H&K lines. Earlier stellar types do not show any signs of correlation between both indices. However, the R'_D index was found to correlate well with R'_{HK} for the most active stars which exhibit the Balmer lines in emission, even though some of these stars do not present a line reversal at the core of the D lines.

2.4. ACTIN indices calculation

Following Gomes da Silva et al. (2018) and Gomes da Silva et al. (2021), ACTIN calculates indices by dividing the average flux in the activity sensitive lines by the average flux in reference regions

$$I = \frac{\sum_{i=1}^{N} F_i}{\sum_{j=1}^{M} R_j}, \tag{5}$$

where F_i is the flux in the activity sensitive line i, R_j the flux in reference region j, N the number of activity lines used, and M the number of reference regions. The flux errors only take into account photon noise. A more detailed explanation on this can be found in Gomes da Silva et al. (2021).

The activity indices S_{CaII}, $I_{H\alpha}$ and I_{NaI} come pre-installed with ACTIN, using line parameters and bandpasses described in Gomes da Silva et al. (2011). As these indices are the most commonly used in the literature, we tested and configured SAITAMA to use these indices as default. Nevertheless, the indices used can be changed by the user.

2.5. Stars from Pepe et al. (2011)

As a preliminary study, we chose three stars from Pepe et al. (2011), HD20794, HD85512 and HD192310, all included in SWEET-Cat. These stars allowed to compare the results from ACTIN with the ones from this paper. We retrieved the spectra taken by HARPS corresponding to the same observations used in Pepe et al. (2011) from the ESO archive[4], corrected the spectra by the radial velocity (RV) of the star (this processed is described later) and binned the spectra by night of observation, obtaining average spectra per night of observation. Using ACTIN, we computed the S_{CaII} activity index for the CaII H&K line. Then, we converted S_{CaII} to S_{MW}, using a calibration from Gomes da Silva et al. (2021), and the S_{MW} to $\log R'_{HK}$ using the B-V color of the star from the Simbad database[5] (Wenger et al. 2000) and a calibration from Rutten (1984). Figure 1 shows a comparison between the $\log R'_{HK}$ index from Pepe et al. (2011) and the $\log R'_{HK}$ computed in this work. HD20794 shows some dispersion but the 1-1 relation is within the error bars. HD85512 overall agree but there is a discrepancy for higher values of $\log R'_{HK}$, where our estimates are higher than in Pepe et al. (2011). The opposite is seen in lower values of $\log R'_{HK}$ for HD192310.

3. Pipeline architecture

The complete processing can be separated in three main steps:

1. spectral data acquisition: queries the data, selects it and downloads the spectra;
2. data processing and analysis: corrects the spectra by its radial velocity (RV), obtains the spectral indices and, if one of the indices is CaII H&K, performs a periodogram analysis, converts to literature indices and computes the chromospheric age and rotational period;
3. data compilation and output: computes statistical information, saves detailed results for each instrument, and combining the data into a master FITS file once all instruments have been processed.

[4] https://archive.eso.org/wdb/wdb/adp/phase3_main/form
[5] https://simbad.u-strasbg.fr/simbad/

These steps are successive, as presented in figure 2, a schematic flowchart of the SAITAMA pipeline. Each one of these steps will be described in detail in the next sections. For exemplification purposes, the majority of plots we show in this work concern HD209100, due to the fact that it has spectra available for the three instruments studied (HARPS, ESPRESSO and UVES) and has an extensive number of observations, allowing for a good estimate on the activity period.

The input parameters on SAITAMA are

- *stars*: object identifiers, preferentially in the HD catalogue. Must be in a list format.
- *instruments*: names of the spectrographs to retrieve the spectra from. Must be in a list format.
- *indices*: spectral lines to compute the activity indices. Must be in a list format.
- *max_spectra*: maximum number of spectra to be downloaded. Must be in a float-like format.
- *min_snr*: minimum SNR to select the spectra. Must be in a float-like format.
- *download*; boolean to decide if the spectra are downloaded or not. If not, it is assumed that the spectra is already inside the folders where the pipeline would download the spectra to.
- *neglect_data*: spectra to be manually neglected for whatever reason. Must be a dictionary.
- *username_eso*: username in the ESO data base to download the data. Must be a string.
- *download_path*: folder to where download spectra. Must be a string.
- *final_path*: folder to save products of the pipeline. Must be a string.

4. Spectral data acquisition

The ESO archive is the main source of reduced spectral public data for SWEET-Cat (Sousa et al. (2021)), and consequently the pipeline is formatted to retrieve all its data from it. For this, we used the *astroquery.eso*[6] submodule to search for the required data in the ESO archive.

In this part of the pipeline, it searches in the ESO archive using the stars' Gaia DR3 identifier, as well as the identifier for the instrument (survey, technically) given. An initial quality check is done by applying preliminary cuts: spectra with S/N < 15 were ignored due to low-quality individual exposures, as well as spectra with S/N, too high to avoid saturation of the individual exposures, that varied according to the instrument. For HARPS and UVES, the limit was S/N = 550 and for ESPRESSO it was S/N = 1000. Due to the nature of the UVES instrument, which takes spectra using two different arms, the spectra may have a narrow interval of wavelength (and therefore do not include the CaII H&K line for example) or have an interruption around half of the spectra (near the Hα region). To avoid this last problem, we made it so that the retrieved spectra from UVES have a minimum wavelength of 6400 Å and a maximum wavelength of 6600 Å.

Unlike the conventional approach for stellar fundamental parameters estimation, like what was done in Sousa et al. (2021), we are not interested in combining the spectra into one single high S/N spectrum, but instead in the variation of every single spectrum in time. To achieve this, we need to perform a quality - time span trade-off check, where the spectra downloaded are not necessarily the ones with higher S/N across the total time span.

[6] https://astroquery.readthedocs.io/en/latest/

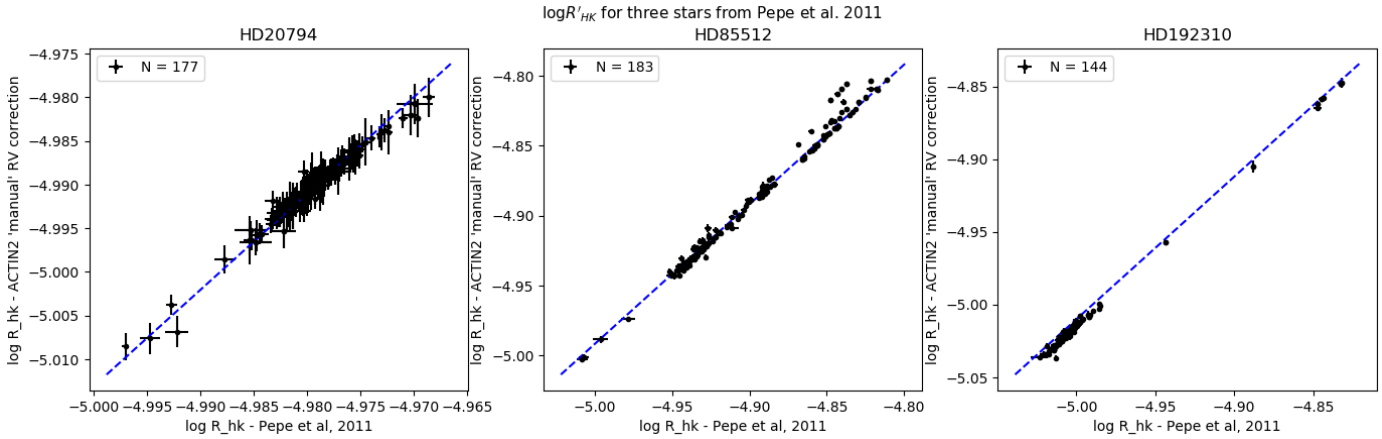

Fig. 1: Comparison of our $\log R'_{HK}$ values with those of Pepe et al. (2011). The blue dashed line is the 1:1 identity and "N" is the number of observations.

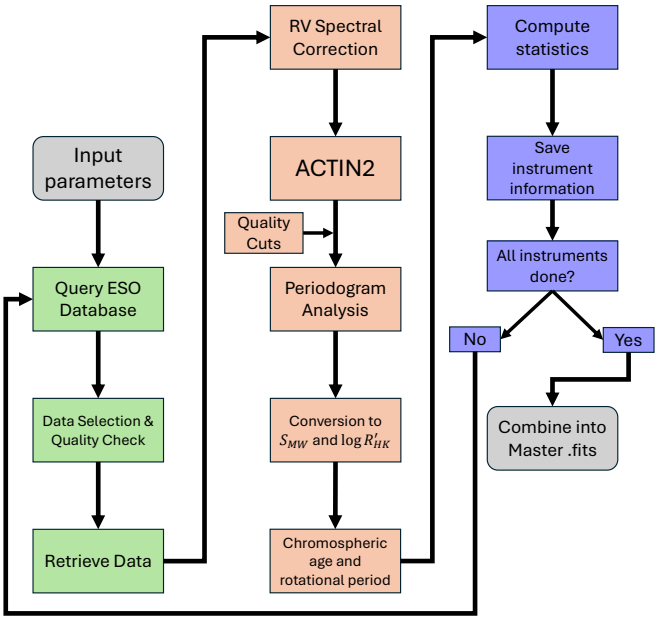

Fig. 2: SAITAMA flowchart. Green color is the spectral data acquisition step, orange is the data processing and analysis step and violet is the data compilation and output step.

Instead, for all the data available after the preliminary cuts, we grouped the observations in months per year and ordered each group by descending S/N. The pipeline then iterates for each group, adding to the new list of spectra to be downloaded the best S/R spectra, until the maximum number of spectra defined by the user is achieved. While testing the pipeline, this maximum number was set to 150, high enough to compute all the other statistical information later on (namely the activity period), but low enough to not burden the computer memory and to ensure a reasonable computing speed.

5. Data processing and analysis

This section describes the stages of correcting the spectra for radial velocity, running the ACTIN2 tool to compute activity indices, analyzing periodograms, and converting the indices to standard scales to estimate chromospheric age and rotational period.

5.1. RV correction

To shift the spectra into it's rest-frame wavelengths, we apply a cross-correlation function, using the *crosscorrRV* function from the PyAstronomy[7] Python library (Czesla et al. (2019)).

To apply the CCF, we need a template spectrum to compare with the stellar spectrum. For this, we used a high S/N (around 1000) day sky solar spectrum provided by Dall et al. (2006)[8]. To minimize the computing time, the search interval is set to a narrow interval in km/s centered in the radial velocity available in the Simbad database. This interval varies for each instrument tested: for HARPS the interval is ±5 km/s, but for ESPRESSO and UVES the interval is much larger, ±110 km/s. This is due to the possibility that the spectra is not corrected for the barycentric radial velocity, so the spectrum's shift is much higher than the expected. If Simbad does not have a radial velocity available, the algorithm will search in an interval of -150 to 150 km/s. The step of search is 0.5 km/s, large enough to make the computing time smaller, but small enough to ensure that the quality of the correction remains reasonable (a 0.5 km/s radial velocity means that the spectrum is shifted 0.011 Å in wavelength).

We tested two options: the first one uses the full spectra to compute the RV, while the second one uses only two 250 Å intervals. For UVES, the intervals are [6400, 6700]Å and [5250, 5500]Å, covering the Hα region and the leftmost side of spectrum, not including the CaII H&K line. For HARPS and ESPRESSO, the intervals were [6400, 6700]Å and [4900, 5250]Å, corresponding to regions that include the Hα and CaII H&K lines. Using this much smaller intervals compared to the whole spectrum resulted in a minimal difference to the first option (often resulting in the same RV values), while making the computing time much smaller. For these reasons, this is the default option in SAITAMA.

Having the RV value obtained, the spectrum is corrected by a Doppler shift

$$\lambda' = \frac{\lambda}{1 + RV/c}, \tag{6}$$

[7] https://github.com/sczesla/PyAstronomy
[8] https://www.eso.org/sci/facilities/lasilla/instruments/harps/inst/monitoring/sun.html

where λ' is the corrected wavelength, λ is the observed (raw) wavelength and c is the light speed.

To be sure that a given spectrum is well corrected, we define α_{RV}, the ratio between the continuum and the center of the line fluxes of a reference line. The chosen line was CaI, due to being a line that is not strongly influenced by stellar activity. α_{RV} is computed as

$$\alpha_{RV} = \frac{F_{\text{continuum}}}{F_{\text{core}}}, \tag{7}$$

where F_{core} is the median flux in the core of the CaI line in an interval of ±0.03Å and $F_{\text{continuum}}$ is the median flux of two windows in the continuum region, defined as [6572.795 − 0.7; 6572.795 − 0.7 + 0.2]Å and [6572.795 + 0.7 − 0.2; 6572.795 + 0.7]Å. The center of CaI is 6572.795 Å and the total window considered was ±0.7Å. The flux in this range was then normalized by dividing it by its median. Figure 3 shows one example of the regions considered for α_{RV}. This spectrum belongs to HD209100, taken with HARPS, and $\alpha_{RV} = 0.2706$.

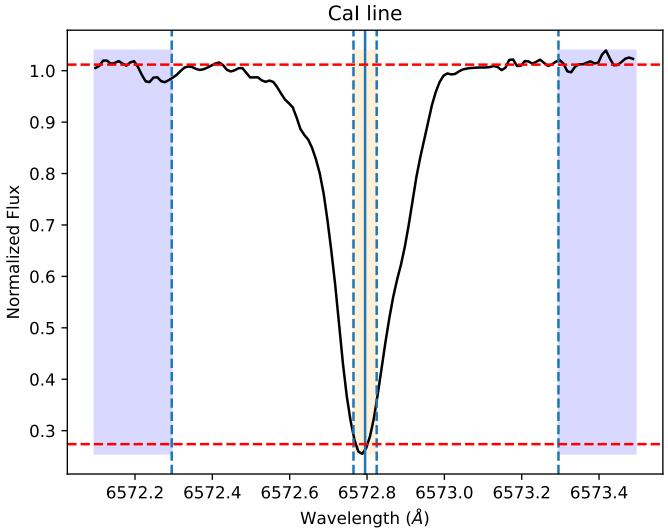

Fig. 3: One of the spectra used for HD209100, taken with HARPS. The blueish regions are the regions used for the continuum flux, while the yellow region was used for the flux in CaI core. The blue dashed lines delimit the continuum and core regions, the solid blue line is the center of CaI and the dashed red lines are the median flux in the continuum and in the core of the line.

We then define β_{RV}, an overall quality indicator of the RV correction. For each spectrum, the algorithm induces different wavelength offsets and if the minimum α_{RV} is not in the interval $[-0.03, 0.03]$Å, then a binary flag γ_{RV} is set to 1. Otherwise, $\gamma_{RV} = 0$. This is exemplified in figure 4, which shows the change of α_{RV} with the wavelength offset. β_{RV} is the ratio between the number of spectra with $\gamma_{RV} = 0$ and the total number of spectra. β_{RV} scale is [0,1], so $\beta_{RV} = 1$ means that all spectra were well corrected, while $\beta_{RV} = 0$ means that none of the spectra was well corrected. For 150 spectra of HD209100 taken with HARPS, $\beta_{RV} = 1$, so all that spectra were well corrected.

To better understand the $[-0.03, 0.03]$Å interval, figure 5 shows the variation of the S_{CaII} index with the offset in wavelength. The mean difference for an offset of ±0.03Å was 4.67×10^{-4} in S_{CaII}, representing a change of around 0.11% in the original (no offset) spectrum. This way, it established that the change

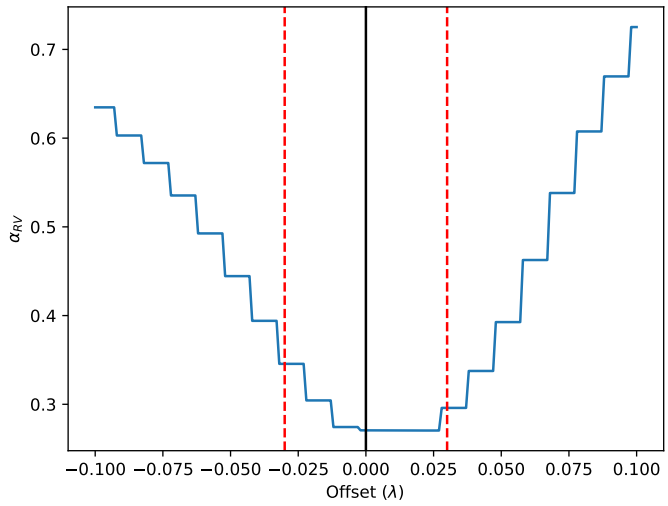

Fig. 4: α_{RV} in function of the wavelength offset. The solid black line is the no offset value and the red dashed lines are the ±0.03Å limits. In this case for HD209100 with a HARPS spectrum, the minimum α_{RV} is inside the ±0.03Å limits so $\gamma_{RV} = 0$.

in S_{CaII} in the interval defined as a well corrected spectrum is not significant.

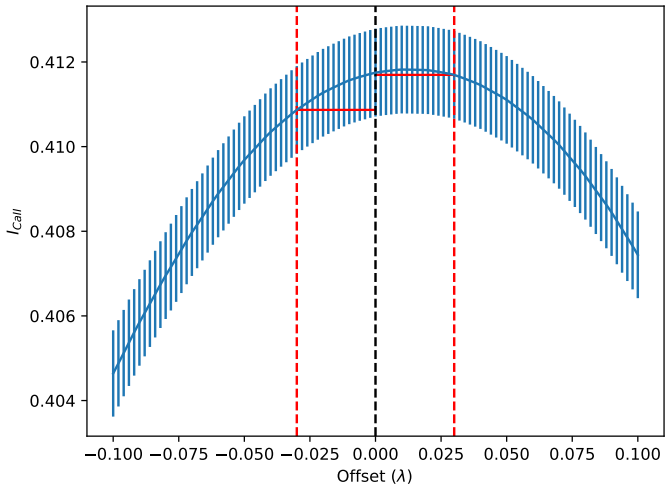

Fig. 5: S_{CaII} in function of the wavelength offset. The red dashed lines are the ±0.03Å limit and the solid red line illustrate the difference in S_{CaII} for the two extremes. The blue bars are the error in S_{CaII}.

5.2. Activity indices and quality cuts

Having shifted the spectra to their rest-frame, SAITAMA runs ACTIN and obtains the activity indices, their errors and Rneg (ratio of negative fluxes in the regions considered by ACTIN). Then, it builds a data frame using the output from ACTIN, adding columns for BJD, the file name, the instrument name, the star identifier, the radial velocity, the SNR and γ_{RV}.

Then, quality cuts are done to the data frame. The first one is to drop any row (spectrum) that has Rneg in at least one of the indices bigger than 0.01, which means that > 1% of the fluxes in

that line is negative. The second cut is to drop any spectra with $\gamma_{RV} = 1$, keeping only the well corrected spectra.

SAITAMA produces plots of known important lines, Hα, CaII H&K, FeII, NaI D1, NaI D2, HeI and CaI, to check manually if the RV correction was done correctly or if there are any visible problems with the spectra.

The final quality cut consists in a 3σ-clip considering the activity indices values, done only if the number of spectra is higher than 10. Some extra care in doing this can be beneficial to avoid cutting any important spectra.

Plots like figure 6 are also saved, showing the variation in time of the activity indices, in this case for CaII H&K, Hα and NaI, and the RV. We remind that, as SAITAMA's goal is to study only stellar activity, the values for RV are not at the precision that a RV oriented analysis would require.

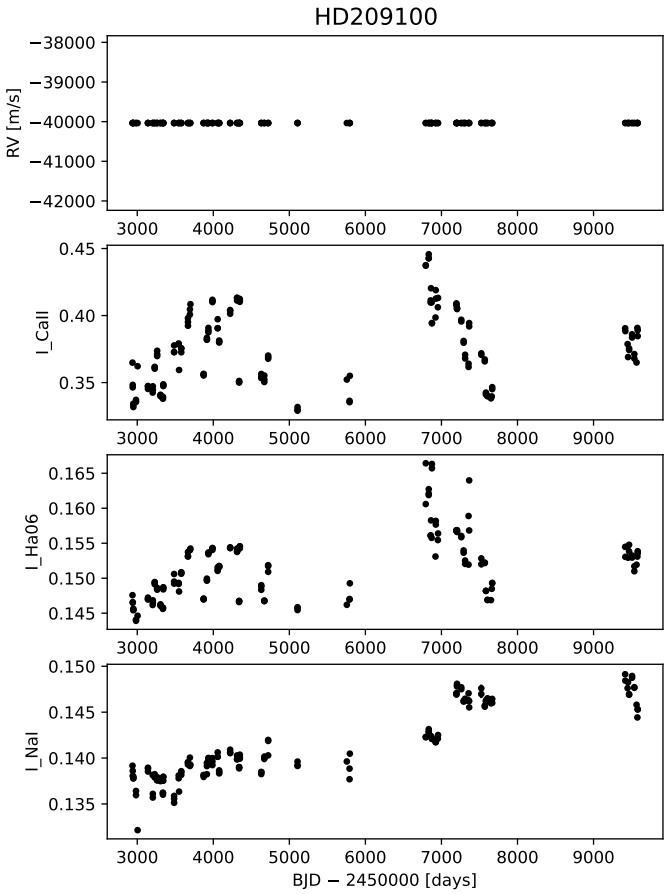

Fig. 6: Activity indices for CaII H&K, Hα and NaI and RV in function of time (BJD - 2450000 days) for HD209100. Spectra from HARPS.

5.3. CaII H&K activity periods

If the data in study is composed of at least 50 spectra in a time span of at least 2 years and includes the I_{CaII} indice, SAITAMA estimates the activity period through Generalized Lomb-Scargle (GLS) periodograms (Zechmeister & Kürster (2009)). The input consists in I_{CaII}, its error, the BJD array, the minimum and maximum periods to compute the periodogram and the step of frequency grid. We used the *LombScargle* function from the Astropy[9] Python library.

To be sure that the peak we retrieve is significant, we applied several quality tests. The first one was to cut all peaks below a False Alarm Probability (FAP) of 1%. Another quality check was made through a Window Function (WF) periodogram, where the BJD array is the same as before but the I_{CaII} array is fixed to 1 everywhere. This provides information about the detected periods in the I_{CaII} periodogram being related to the observation cadence. If we have any period that is around half or a third of some period in the WF, this can be an harmonical signal of that period. A third test consists in analysing the BJD gaps in the data, which is optional.

The GLS significant peaks are retrieved by excluding peaks close to the WF (and to the gaps, if chosen), with a tolerance of 10%. A final quality check is to see if there are harmonics in the significant peaks that include the best period peak. Figure 10 shows the GLS periodogram for HD209100 with HARPS spectra, where the right plot is a fit of the best period using *curve_fit* from the scipy (Virtanen et al. (2020)) Python library. Figure 11 is the Window Function (WF) for this star and instrument.

The error of the period with the maximum power is computed by getting the curvature in the power peak by fitting a parabola $y = aa * x^2$, adapted from the GLS implementation by PyAstronomy. We tested different ways of estimating the period error, but ultimately decided on making the PyAstronomy implementation the default one. A more detailed discussion on the different ways of estimating the error can be found in appendix B.

Having obtained the period and the corresponding error, we developed a color based flag to indicate the quality of the periodogram analysis:

- Green: error below 20% of the period value and no harmonics involving the best period;
- Yellow: error between 20% and 30%, with no harmonics involving the best period;
- Orange: error higher than 30% or no error computed or there are harmonics involving the best period;
- Red: 1 or more peaks with power close to the period power (>85%) or period bigger than the time span of the data;
- Black: period retrieved under 1 year or over 100 years or peak below FAP 1% level or no significant peak obtained;
- White: number of spectra below 50 and time span lower than 2 years.

Table 1 shows a summary of the color based flags for the quality of the periods retrieved. For scientific purposes, we recommend the use of only the periods flagged green or yellow.

5.4. S_{MW}, $\log R'_{HK}$, chromospheric age and rotational period

To convert the I_{CaII} from ACTIN to the common indices used in the literature Mount-Wilson index S_{MW} and $\log R'_{HK}$, described earlier in subsection 2.1, we used calibrations from a GitHub repository[10].

To obtain S_{MW}, we use two different calibrations for HARPS and ESPRESSO:

- HARPS: described in Gomes da Silva et al. (2021), based on S-index calculated with ACTIN, based on 43 stars with $0.105 < S_{MW} < 0.496$ from Duncan et al. (1991) and Baliunas et al. (1995);

[9] https://github.com/astropy
[10] https://github.com/gomesdasilva/pyrhk

Table 1: Summary of the color based flags for the quality of the periods retrieved by SAITAMA. One should think of these as "bottom-up": start by the requirements of the worse flags and iteratively check the boxes until the best flag possible.

Flag	Error (%)	Close periods	P > t_{span}	P < 1 yr or P > 100 yrs	Harmonics	< 1% FAP	N > 50 & t_{span} > 2 yrs	Sig. peak
Green	<20	No	No	No	No	No	Yes	Yes
Yellow	>20 and <30	No	No	No	No	No	Yes	Yes
Orange	> 30 or none	No	No	No	Yes	No	Yes	Yes
Red	-	Yes	Yes	No	Yes	No	Yes	Yes
Black	-	-	-	Yes	-	Yes	Yes	No
White	-	-	-	-	-	-	No	No

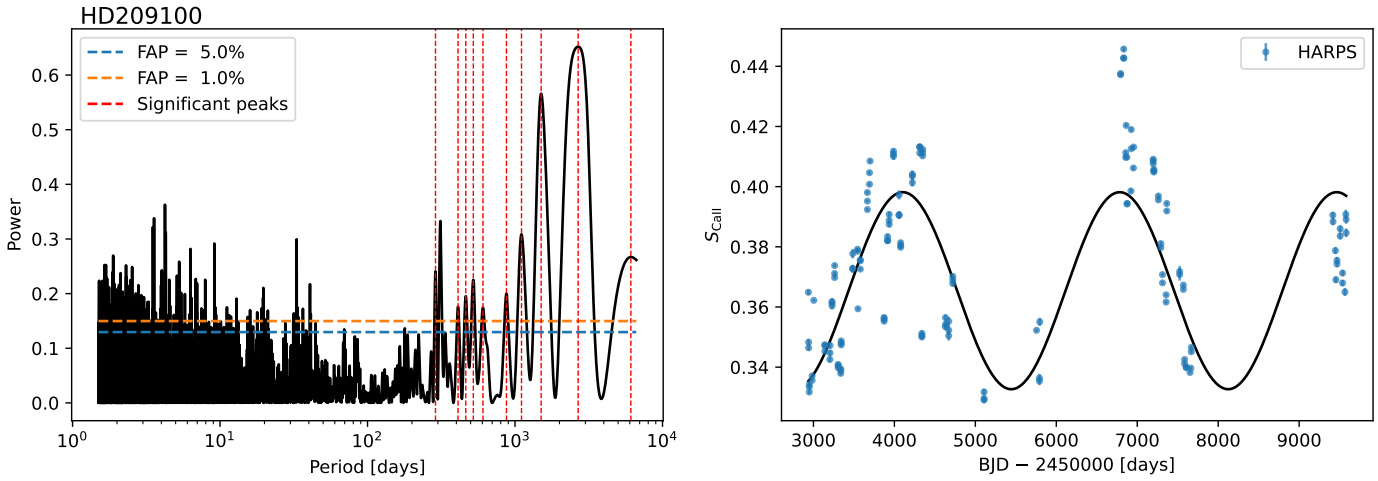

Fig. 7: GLS periodogram for HD209100 using HARPS spectra. The left panel shows the periodogram, where the horizontal dashed lines are the FAP limits and the vertical dashed lines indicate the significant peaks. The right panel is a fit to the data using the best period retrieved.

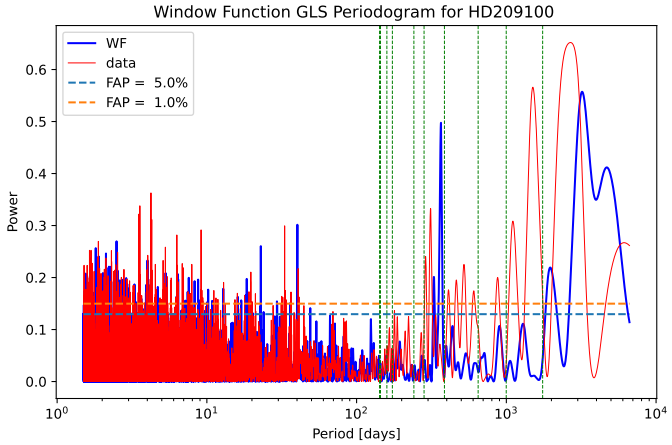

Fig. 8: GLS periodogram Window Function (WF) for HD209100 using HARPS spectra. The horizontal dashed lines are the FAP limits and the dashed vertical lines are the gaps width. The solid red line represents the data shown in figure 10, while the blue solid line is the WF.

– ESPRESSO: preliminary calibration based on 27 stars in common between HARPS and ESPRESSO.

In this part of the pipeline the spectra from UVES were not considered, as none of its spectra covers the CaII H&K lines. The conversion from S_{MW} to $\log R'_{HK}$ implies some caveats, as it requires different calibrations and the stars' B-V color. The B-V color is retrieved from the Simbad database. In the case it is not available in Simbad, we used a conversion between effective temperature and B-V color index based on a black body spectrum and the filter functions presented by Ballesteros (2012). The effective temperature is taken from SWEET-Cat. We used two different calibrations for the bolometric corrections: one from Suárez Mascareño et al. (2016), Suárez Mascareño et al. (2015) applied if the star's $T_{\rm eff}$ < 3700 K (M type), and another from Rutten (1984) otherwise.

The rotation period and the respective error, using $\log R'_{HK}$, are calculated through two different calibrations, from Noyes (1984) or from Mamajek & Hillenbrand (2008), both valid only for $-5.5 < \log R'_{HK} < -4.3$. The chromospheric age, from gyrochronology, is obtained via Mamajek & Hillenbrand (2008), valid for $0.5 < B - V < 0.9$.

6. Data compilation and output

This section focuses on computing statistical information, saving detailed results for each instrument, and combining the data into a master FITS file once all instruments have been processed.

6.1. Statistics

After the RV correction, quality cuts, periodogram analysis and conversions to S_{MW} and $\log R'_{HK}$, the final step of the processing module is to compute statistical data that describe summarily the spectra studied. The quantities computed are the maximum, minimum, mean, median, standard deviation and weighted mean of $S_{\rm CaII}$, $I_{H\alpha}$, I_{NaI}, RV, S_{MW}, $\log R'_{HK}$, rotation periods and chromospheric age. The weighted mean is computed by taking into

account the error, by

$$\text{Weighted mean} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}, \quad (8)$$

where N is the observation/spectrum, x_i is the quantity in study and w_i is the weight, defined as $w_i = 1/\sigma_i^2$, where σ_i is the error in the quantity. The number of spectra N_{spectra} is also saved.

6.2. Files saved per instrument

SAITAMA saves a set of files for each different instrument, which includes

- .csv table with the raw (pre-processed) data that includes the activity indices, their error and the Rneg, the RV obtained, γ_{RV}, SNR, BJD and the names of the spectrum file, instrument and star (object);
- .csv table with the statistics computed;
- .pdf plots of important spectral lines, such as CaI, CaII H%K, Hα, FeI, HeI and NaI D1 and D2;
- .pdf plot of the activity indices and RV varying in time, as in figure 6;
- .pdf plots of the GLS periodogram, correspondent fit to data and Window Function, as in figures 7 and 11;
- .txt report on the periodogram analysis, to easily access some relevant information;
- FITS file with a table with the processed data: activity indices and their errors and Rneg, RV, γ_{RV}, SNR, BJD, names of the spectrum file, instrument and star, S_{MW}, $\log R'_{HK}$, rotation periods and chromospheric age and their errors. In the header of this FITS file we save the statistics for each quantity, as well as the name of the star and of the instrument, the minimum and maximum SNR, the time span, the period and error, the period color flag and β_{RV}.

6.3. Master (final) files

When all instruments given as input are treated, SAITAMA combines the data from each instrument and recomputes the periodogram analysis and the statistics. As each individual instrument data set is already processed, the only quality cut made is a 3σ-cut (if there is at least 10 spectra) to ensure that the different data sets combined don't have outliers. Then, a plot like figure 9 is saved, highlighting the data from different instruments. The periodogram analysis is repeated in the same manner as explained before, as well as the computation of the statistics.

The totality of these data is saved as a final "master" FITS file, containing a table and a header for each instrument and a combined table and header, the most important SAITAMA product. Besides the FITS file, plots of the GLS periodogram, fit to data and WF are also saved, seen in figures 10 and 11, along with a .txt report on the periodogram analysis.

7. Test stars results

To choose a set of stars to use in the testing phase of this pipeline, we filtered the stars in SWEET-Cat, choosing the ones that agree with

- V magnitude < 8;
- Declination under +30º
- SWFlag = 1 (indicating homogeneus parameters, described in Sousa et al. (2021)).

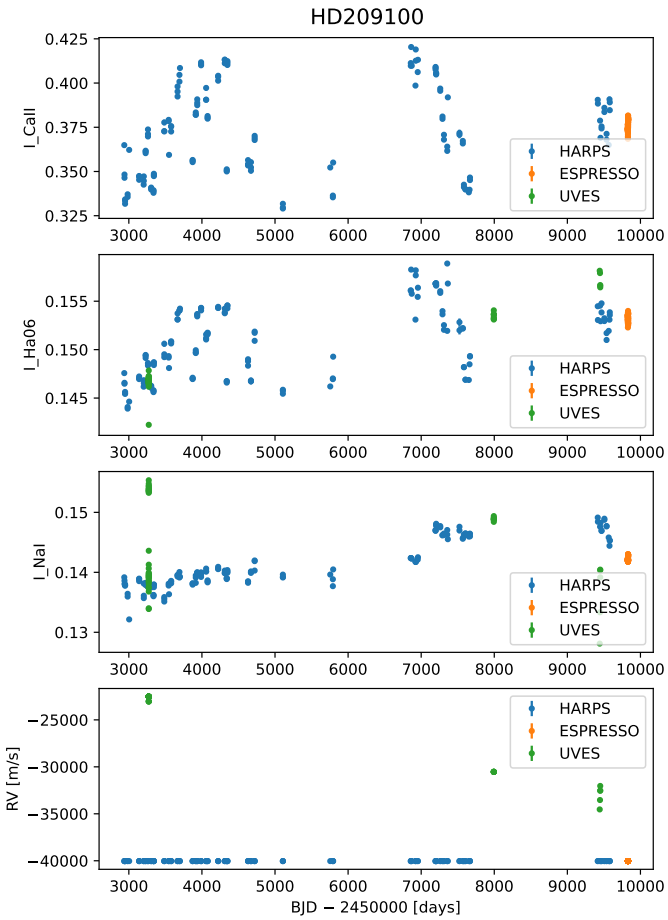

Fig. 9: Activity indices for CaII H&K, Hα and NaI and RV in function of time (BJD - 2450000 days) for HD209100. Combined data from HARPS, ESPRESSO and UVES.

We ordered the stars by effective temperature and chose the stars that had a high number of spectra, with minimum SNR of 15 and maximum SNR of 550, taken with HARPS and with UVES, to ensure a reasonable comparison. We chose 14 stars, adding the 3 stars from Pepe et al. (2011) and a cold star, HD47536, that has a low number of spectra to test the pipeline. This way, we were left with a set of 18 stars with T_{eff} ranging from 4412 K to 6419 K. Table 2 presents the stellar photometric, astrometric and spectroscopic parameters for the test stars.

Table 3 shows some results using the combined data for the 18 test stars. For summary sake, we only show the mean value of the activity indices studied, but SAITAMA gives access to other useful statistical information. From the 18 stars, we could only get reasonable estimates for the activity period (flagged green or yellow) for 6 stars. Regarding the chromospheric age, we retrieved estimates for 14 stars.

8. Summary

In this work we presented SAITAMA, a pipeline for the derivation of spectral activity indices for stars with exoplanets, with a detailed description of the core processing stages of the spectral data acquisition, data processing and analysis, and data compilation and output modules. Additionally, we discussed the particularities related to the functioning of SAITAMA and all the

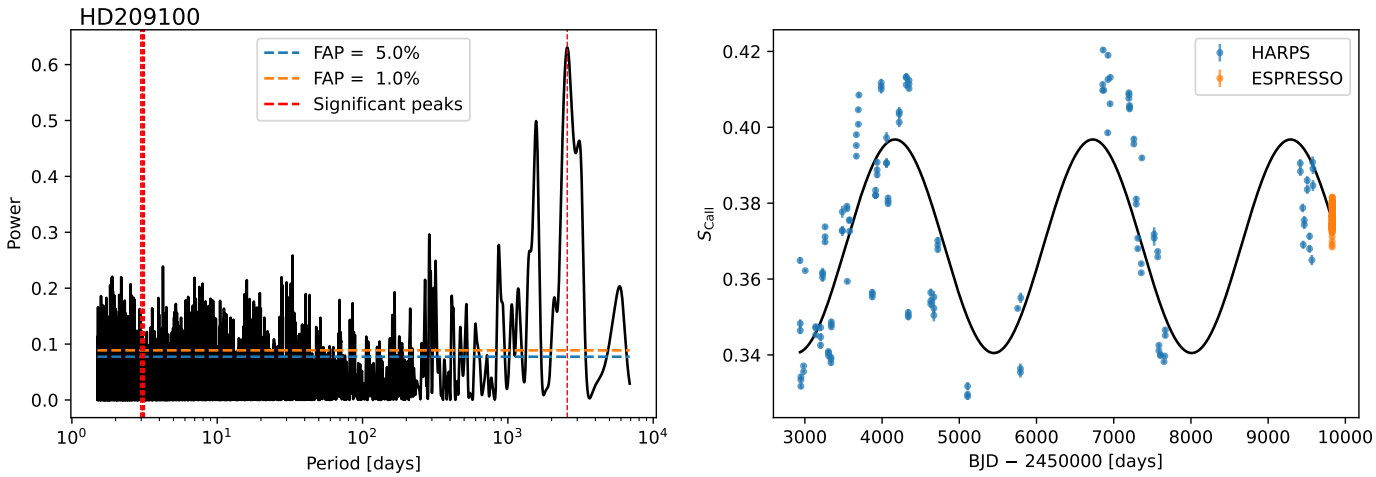

Fig. 10: GLS periodogram for HD209100 using the combined data from HARPS, ESPRESSO and UVES. The left panel shows the periodogram, where the horizontal dashed lines are the FAP limits and the vertical dashed lines indicate the significant peaks. The right panel is a fit to the data using the best period retrieved.

Table 2: Observed and inferred stellar parameters for the 18 test stars, taken from SWEET-Cat.

HD	RA [°]	Dec [°]	V [mag]	π [mas]	$T_{\rm eff}$ [K]	log g [cgs]	[Fe/H] [dex]	M [$M_\odot$]
142A	00 06 19.17	-49 04 30.68	5.70	38.19 ± 0.03	6419 ± 30	4.45 ± 0.04	0.20 ± 0.02	1.34 ± 0.01
1461	00 18 41.86	-08 03 10.80	6.46	42.73 ± 0.02	5812 ± 29	4.50 ± 0.06	0.21 ± 0.02	1.06 ± 0.01
10647	01 42 29.31	-53 44 27.00	5.52	57.64 ± 0.04	6178 ± 20	4.49 ± 0.03	−0.01 ± 0.01	1.10 ± 0.01
13445	02 10 25.93	-50 49 25.41	6.17	92.92 ± 0.04	5110 ± 34	4.39 ± 0.07	−0.31 ± 0.02	0.75 ± 0.00
16141	02 35 19.92	-03 33 38.18	6.78	26.50 ± 0.02	5790 ± 19	4.14 ± 0.03	0.15 ± 0.02	1.14 ± 0.01
16417	02 36 58.60	-34 34 40.71	5.78	39.29 ± 0.03	5848 ± 17	4.14 ± 0.03	0.14 ± 0.01	1.19 ± 0.01
20794	03 19 55.65	-43 04 11.21	4.26 ± 0.01	165.52 ± 0.07	5422 ± 39	4.53 ± 0.08	−0.42 ± 0.03	0.80 ± 0.01
22049	03 32 55.84	-09 27 29.73	3.73	310.57 ± 0.13	5053 ± 46	4.45 ± 0.10	−0.14 ± 0.03	0.75 ± 0.01
46375	06 33 12.62	+05 27 46.52	7.84	33.87 ± 0.02	5274 ± 54	4.26 ± 0.12	0.20 ± 0.04	0.87 ± 0.01
47536	06 37 47.61	-32 20 23.04	5.26 ± 0.01	7.99 ± 0.05	4412 ± 46	2.06 ± 0.12	−0.67 ± 0.03	2.50 ± 0.02
85512	09 51 07.05	-43 30 10.02	7.65	88.67 ± 0.01	4718 ± 126	4.08 ± 0.61	−0.40 ± 0.05	0.67 ± 0.02
102365	11 46 31.07	-40 30 01.27	4.89 ± 0.01	107.30 ± 0.08	5629 ± 17	4.43 ± 0.03	−0.30 ± 0.01	0.87 ± 0.00
108147	12 25 46.26	-64 01 19.52	7.00	25.75 ± 0.01	6262 ± 24	4.45 ± 0.03	0.18 ± 0.02	1.22 ± 0.01
115617	13 18 24.31	-18 18 40.30	4.74 ± 0.01	117.17 ± 0.14	5556 ± 19	4.31 ± 0.03	−0.01 ± 0.01	0.91 ± 0.00
160691	17 44 08.70	-51 50 02.58	5.15	64.08 ± 0.09	5797 ± 26	4.26 ± 0.05	0.30 ± 0.02	1.15 ± 0.01
179949	19 15 33.22	-24 10 45.66	6.25	36.31 ± 0.03	6282 ± 21	4.49 ± 0.04	0.23 ± 0.02	1.24 ± 0.01
192310	20 15 17.39	-27 01 58.71	5.72 ± 0.01	113.48 ± 0.051	5041 ± 61	4.34 ± 0.12	−0.06 ± 0.03	0.77 ± 0.01
209100	22 03 22.00	-56 47 10.00	4.70	274.84 ± 0.09	4694 ± 115	4.40 ± 0.33	−0.17 ± 0.04	0.70 ± 0.01

caveats associated. We tested the pipeline with a total data set of 18 stars, included in SWEET-Cat.

As discussed in the introduction, the study of the chromospheric activity of stars with exoplanets is crucial in the scope of planetary detection and characterization. With SAITAMA, we aim to provide a quick and painless way to obtain relevant information on the spectra activity of a star.

There are some caveats regarding the functioning of SAITAMA. There can be problems in a given spectrum that are not noticed by the pipeline, for example low spectral resolution, resulting in dubious results. This was particularly seen in UVES spectra. Another possible problem arises from the combination of spectra from different instruments, because there can be discrepancies in the spectra due to instrumental characteristics. In Appendix A we studied the UVES spectra, comparing them to HARPS spectra, but did not find out significant correlations across all stars. In the future, a more thorough analysis on this discrepancy between instruments would be very useful. Nevertheless, the 3σ-clip included in SAITAMA should minimize the impact of this issue.

All of the 18 test stars consist in FGK type stars, also called "solar type". Continuing testing the pipeline with stars of different types, such as A-type or M-type (higher or lower $T_{\rm eff}$ than FGK), may provide new insights to improve the pipeline, as their spectra will have different features. Regarding the error of the activity period via periodogram analysis, we tested four different methods, all described in Appendix B. For now, we opted for an implementation of the Zechmeister & Kürster (2009) method included in the source code of PyAstronomy, but some more studies on the nature of the error estimation methods can enlighten us on what the best method should be.

All taken into consideration, we will continue working on SAITAMA in the near future so that it can be used without reservation in the studies about stellar activity. Specifically, we hope that this pipeline is useful in expanding the contents of the SWEET-Cat (Santos et al. (2013); Sousa et al. (2021)) catalogue, open way for studies with the goal of finding correlations between stellar activity and the properties of the planets hosted by these stars.

Table 3: Partial results using the combined data for the 18 test stars. "Instr." refers to the instrument used, where 1 is HARPS, 2 is ESPRESSO and 3 is UVES. The time span $t_{\rm span}$, activity period $P_{S_{CaII}}$ and mean rotation periods from Noyes (1984) or from Mamajek & Hillenbrand (2008) are in days. N_{CaII} is the number of spectra including the CaII H&K lines and this number is the same for S_{MW}, $\log R'_{HK}$, the rotation period and the ages (given in Gyr).

HD	Instr.	SNR	$t_{\rm span}$	$P_{S_{CaII}}$	$P_{\rm Flag}$	$\langle S_{CaII}\rangle$	N_{CaII}	$\langle I_{H_\alpha}\rangle$	$N_{H_\alpha,NaI}$	$\langle I_{NaI}\rangle$	$\langle S_{MW}\rangle$	$\langle \log R'_{HK}\rangle$	$\langle P_{\rm rot,\,N84}\rangle$	$\langle P_{\rm rot,\,M08}\rangle$	$\langle {\rm Age}_{M08}\rangle$
142A	1,3	73 - 524	5184	0 ± 0	white	0.122	30	0.103	36	36	0.154	-4.946	11.861	12.276	2.439
1461	1,3	60 - 336	6495	3770 ± 143	green	0.128	142	0.101	142	142	0.161	-5.028	29.648	31.730	5.800
10647	1,3	82 - 468	5916	3027 ± 78	green	0.159	130	0.111	130	130	0.198	-4.706	7.677	7.424	1.161
13445	1,3	38 - 436	2265	1416 ± 47	red	0.215	122	0.130	123	123	0.265	-4.749	31.002	30.142	3.784
16141	1,3	59 - 239	4627	0 ± 0	white	0.117	45	0.100	54	54	0.147	-5.132	36.423	40.285	8.113
16417	1,3	76 - 471	5074	235 ± 2	black	0.122	149	0.100	149	149	0.154	-5.062	28.019	30.370	5.734
20794	1,3	160 - 542	7039	6 ± 0	black	0.135	134	0.105	134	134	0.170	-4.993	32.153	33.930	5.989
22049	1,2,3	107 - 546	5900	1250 ± 9	red	0.392	148	0.167	188	188	0.477	-4.490	16.075	15.497	1.097
46375	1,2,	40 - 150	1453	0 ± 0	black	0.151	91	0.116	91	91	0.189	-4.975	43.461	45.552	7.452
47536	1,3	66 - 341	848	0 ± 0	white	0.093	16	0.097	22	22	0.119	-5.495	100.128	90.001	—
85512	1	81 - 194	7342	3674 ± 53	green	0.363	141	0.158	141	141	0.442	-4.928	47.752	49.178	—
102365	1,2,3	62 - 469	5659	245 ± 1	black	0.139	145	0.102	145	145	0.174	-4.947	26.333	27.269	4.582
108147	1,3	89 - 371	5635	0 ± 0	white	0.145	41	0.106	50	50	0.181	-4.788	9.532	9.341	1.546
115617	1,2,3	117 - 500	6407	2798 ± 180	green	0.133	141	0.103	142	142	0.167	-5.003	31.330	33.201	5.923
160691	1,2,3	122 - 412	4093	46 ± 0	black	0.119	143	0.100	143	143	0.150	-5.107	34.364	37.784	7.443
179949	1	72 - 449	0	0 ± 0	white	0.168	3	0.113	35	35	0.209	-4.669	7.369	7.108	0.994
192310	1	57 - 407	4717	3417 ± 81	green	0.166	147	0.120	147	147	0.207	-4.951	44.170	45.914	—
209100	1	15-550	5900	2562 ± 25	green	0.166	147	0.120	147	147	0.207	-4.951	44.170	45.914	—

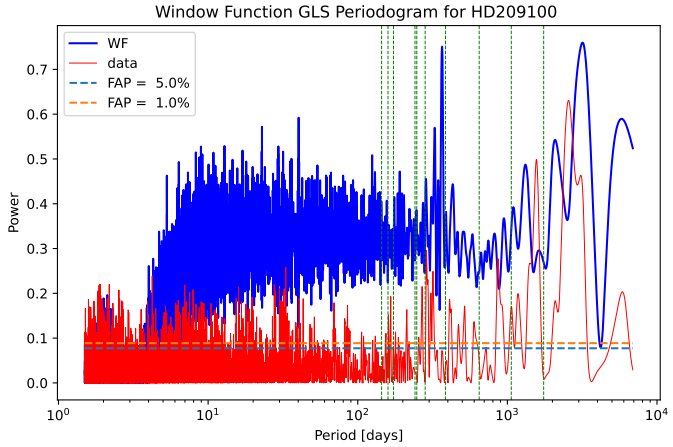

Fig. 11: GLS periodogram Window Function (WF) for HD209100 using the combined data from HARPS, ESPRESSO and UVES. The horizontal dashed lines are the FAP limits and the dashed vertical lines are the gaps width. The solid red line represents the data shown in figure 10, while the blue solid line is the WF.

References

Baliunas, S. L., Donahue, R. A., Soon, W. H., et al. 1995, ApJ, 438, 269. doi:10.1086/175072

Ballesteros, F. J. 2012, EPL (Europhysics Letters), 97, 34008. doi:10.1209/0295-5075/97/34008

Boisse, I., Moutou, C., Vidal-Madjar, A., et al. 2009, A&A, 495, 959. doi:10.1051/0004-6361:200810648

Czesla, S., Schröter, S., Schneider, C. P., et al. 2019, Astrophysics Source Code Library. ascl:1906.010

Dall, T. H., Santos, N. C., Arentoft, T., et al. 2006, A&A, 454, 341. doi:10.1051/0004-6361:20065021

Díaz, R. F., Cincunegui, C., & Mauas, P. J. D. 2007, MNRAS, 378, 1007. doi:10.1111/j.1365-2966.2007.11833.x

Duncan, D. K., Vaughan, A. H., Wilson, O. C., et al. 1991, ApJS, 76, 383. doi:10.1086/191572

Gomes da Silva, J., Santos, N. C., Bonfils, X., et al. 2011, A&A, 534, A30. doi:10.1051/0004-6361/201116971

Gomes da Silva, J., Figueira, P., Santos, N., et al. 2018, The Journal of Open Source Software, 3, 667. doi:10.21105/joss.00667

Gomes da Silva, J., Santos, N. C., Adibekyan, V., et al. 2021, A&A, 646, A77. doi:10.1051/0004-6361/202039765

Gomes da Silva, J., Bensabat, A., Monteiro, T., et al. 2022, A&A, 668, A174. doi:10.1051/0004-6361/202244595

Mamajek, E. E. & Hillenbrand, L. A. 2008, ApJ, 687, 1264. doi:10.1086/591785

Middelkoop, F. 1982, A&A, 107, 31

Noyes, R. W. 1984, Advances in Space Research, 4, 151. doi:10.1016/0273-1177(84)90379-X

Pepe, F., Lovis, C., Ségransan, D., et al. 2011, A&A, 534, A58. doi:10.1051/0004-6361/201117055

Rutten, R. G. M. 1984, A&A, 130, 353

Santos, N. C., Sousa, S. G., Mortier, A., et al. 2013, A&A, 556, A150. doi:10.1051/0004-6361/201321286

Sousa, S. G., Adibekyan, V., Delgado-Mena, E., et al. 2021, A&A, 656, A53. doi:10.1051/0004-6361/202141584

Suárez Mascareño, A., Rebolo, R., González Hernández, J. I., et al. 2015, MNRAS, 452, 2745. doi:10.1093/mnras/stv1441

Suárez Mascareño, A., Rebolo, R., & González Hernández, J. I. 2016, A&A, 595, A12. doi:10.1051/0004-6361/201628586

VanderPlas, J. T. 2018, ApJS, 236, 16. doi:10.3847/1538-4365/aab766

Vaughan, A. H., Preston, G. W., & Wilson, O. C. 1978, PASP, 90, 267. doi:10.1086/130324

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261. doi:10.1038/s41592-019-0686-2

Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9. doi:10.1051/aas:2000332

Wilson, O. C. 1978, ApJ, 226, 379. doi:10.1086/156618

Zechmeister, M. & Kürster, M. 2009, A&A, 496, 577. doi:10.1051/0004-6361:200811296

Appendix A: UVES spectra study

One should be careful when combining data from different instruments, as some instrumental differences can propagate to the spectra retrieved, changing the outcome of stellar activity computations. At first, we noticed that the spectra from UVES often retrieved different values for the activity indices (mainly $I_{H\alpha}$) when compared with spectra from HARPS (or ESPRESSO).

To further analyse this discrepancies, we tried to find correlations between several quantities:

- SNR in function of time (BJD);
- RV in function of SNR;
- $I_{H\alpha}$ in function of SNR;
- Spectral Resolution R in function of BJD, including HARPS and UVES spectra.

We plotted these relations for each star, as seen in figure A.1 an example for HD209100. We also plotted $I_{H\alpha}$ in function of BJD for both UVES and HARPS spectra, and marked by visual inspection data points from UVES that we deemed outliers. We also analysed some meteorological conditions, like ambient temperature, mean airmass during the observation, relative humidity and wind speed, as a function of BJD, as seen in figure A.2 for HD209100.

Unfortunately, we did not find a definite conclusion on the origin of the discrepancies between the instruments, but we did find some possible reasons in individual cases. Nonetheless, one prevalent factor seemed to be that lower spectral resolution results in a higher estimate for $I_{H\alpha}$. Here we leave some remarks about some of the test stars analysed:

- HD1461: one outlier with RV close to 0, isolated in time from other observations and with a spectral resolution much lower than the other spectra (<50 000);
- HD16141: only one spectrum not outlier and all of the outliers were observed in the same night, with a much lower spectral resolution (<50 000);
- HD16417:the outlier has spectral resolution lower than 20 000;
- HD102365: all of the outliers are concentrated in the same night, where only one spectrum is not an outlier. All of the spectra from UVES have spectral resolution a bit lower than HARPS;
- HD115617: the outliers have SNR between 100 and 250, the outliers have low spectral resolution but so does the outlier, so it is inconclusive;
- HD160691: only one is non-outlier, the one with lower SNR;
- HD192310: outliers have spectral resolution under 50 000;
- HD142A: no outliers;
- HD10647: no outliers;
- HD13445: no outliers;
- HD22049: no outliers;
- HD47536: no outliers;
- HD108147: no outliers;
- HD179949: no outliers;
- HD20794: inconclusive;
- HD209100: inconclusive.

We also inspected the individual spectra classified as outliers or non-outliers and compared with HARPS spectra. Additionally, we also analysed the flux in the reference regions described in ACTIN in function of $I_{H\alpha}$.

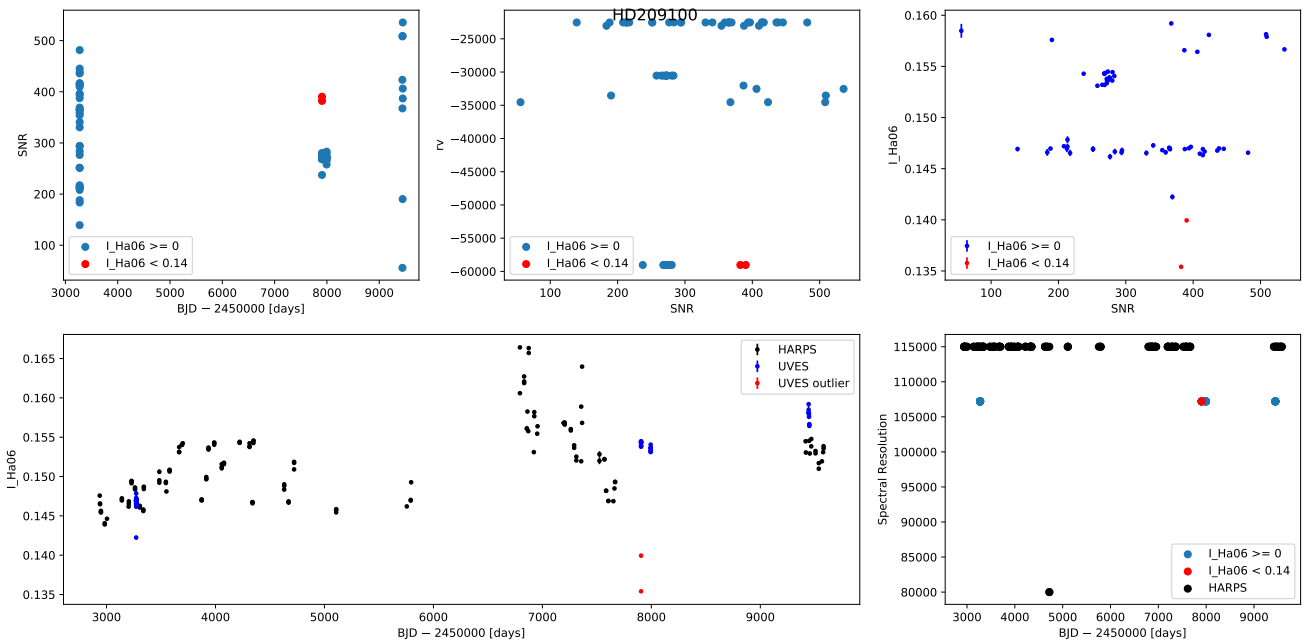

Fig. A.1: Comparison of UVES and HARPS data for HD209100. Top left plot: SNR in function of BJD. Top middle plot: RV in function of SNR. Top right plot: $I_{H\alpha}$ in function of SNR. Bottom right plot: Spectral Resolution R in function of BJD, including HARPS and UVES spectra. Bottom left plot: $I_{H\alpha}$ in function of BJD for both UVES and HARPS spectra. Based on this last plot, we marked the UVES outliers in red, the non-outliers in blue and HARPS spectra in black.

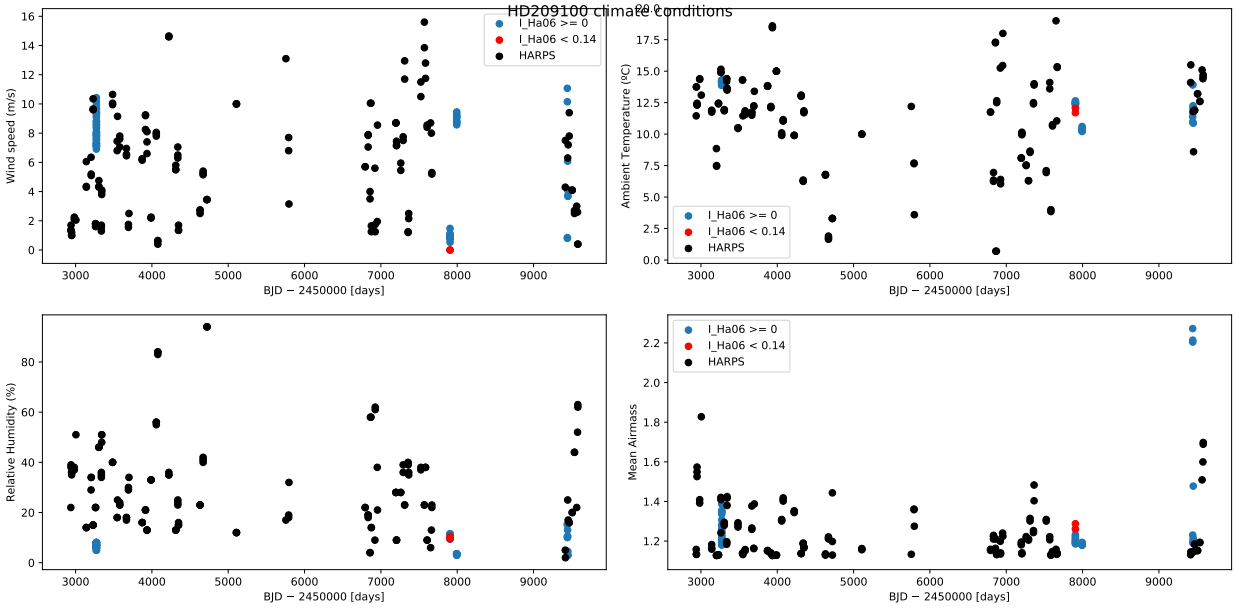

Fig. A.2: Comparison of meteorological conditions for UVES and HARPS spectra: wind speed (m/s), ambient temperature (ºC), relative humidity (%) and mean airmass during the observation. UVES spectra are colored based on $I_{H\alpha}$ in function of BJD for both UVES and HARPS spectra, as described in figure A.1.

Appendix B: Methods to estimate error of period

To estimate the error of the period retrieved by the GLS periodograms described in Sect. 5.3, we studied four different methods with HARPS data. Figure B.1 shows 1-1 comparisons between the four methods for the 6 stars that had their period estimate flagged either "green" or "yellow". As we can see, the method retrieving the most discrepant results is the Gaussian fit.

Appendix B.1: GLS implementation by PyAstronomy

As the Astropy implementation of the GLS periodograms does not include a way to retrieve the error on the best significant period peak and the PyAstronomy implementation does not allow for WF periodograms, we decided to implement the Astropy implementation in SAITAMA, while manually implementing the PyAstronomy function to compute the error. Following the source code in the PyAstronomy repository[11] based on Zechmeister & Kürster (2009), we get the curvature in the power peak by fitting a parabola of the form $y = aa * x^2$. If the index of the power peak k is between 1 and the length of the frequency array minus 2, we shift the parabola origin to the power peak by defining

$$xh = (\text{freq}[k-1:k+2] - \text{freq}[k])^2 \tag{B.1}$$

and

$$yh = p[k-1:k+2] - p_{\max}. \tag{B.2}$$

Where $p_{\max}$ is the power corresponding to the best peak. Then we calculate the curvature (final equation from least square)

$$aa = \frac{yh \cdot xh}{xh \cdot xh} \tag{B.3}$$

and then the error in frequency is

$$\text{Frequency Error} = \sqrt{\frac{-2}{N}\frac{1}{aa}(1 - p_{\max})}. \tag{B.4}$$

Where N is the number of data points. The error in period is then defined as

$$\text{Period error} = \frac{\text{Frequency Error}}{\text{Best frequency}^2}. \tag{B.5}$$

The frequency array is by default the same used in the GLS periodogram, but if no frequency array is given, the function computes it via Nyquist frequencies and the time array (BJD).

Appendix B.2: Fitting with curve_fit

Another method we implemented is by using the *curve_fit* function from the Scipy Python library. It computes the error of the period obtained from the GLS periodogram by fitting a sinusoidal model, restrained enough so that the period obtained is approximately the same as the one from the periodogram. We define a sinusoidal model

$$y = A \sin\left(2\pi \frac{t}{P} + \phi\right) + \omega + m \times t. \tag{B.6}$$

A is the amplitude of the signal, y is the activity index S_{CaII}, t is the BJD time, P is the period from the GLS periodogram, ϕ is the phase, ω is an offset and m is the slope of a linear function of t. As bounds for the parameters, A varies between 0 and $+\infty$, ϕ, ω and m vary from $-\infty$ to $+\infty$, and the period is restrained to 99.999% to 100.001% of the best period from the GLS periodogram. Different restrains on the period result in a variation in the error obtained. The errors for each parameter are retrieved from the covariance matrix given by *curve_fit*.

[11] https://github.com/sczesla/PyAstronomy/blob/master/src/pyTiming/pyPeriod/gls.py

Appendix B.3: Equation 52 from VanderPlas (2018)

Equation 52 from VanderPlas (2018) provides a scaling on the standard deviation of the frequency of the best peak, basically fitting a Gaussian curve to the (exponentiated) peak. This dependence comes from the fact that the Bayesian uncertainty is related to the width of the exponentiated periodogram, which depends on $\text{Power}_{\max}$, the height of the peak.

The author considers a periodogram with maximum value $\text{Power}_{\max} = \text{Power}(f_{\max})$, so that $\text{Power}(f_{\max} \pm f_{1/2}) = P_{\max}/2$. The Bayesian uncertainty comes from approximating the exponentiated peak as a Gaussian

$$\exp\left[\text{Power}(f_{\max} \pm \delta f)\right] \propto \exp\left[-\delta f^2/(2\sigma_f^2)\right]. \tag{B.7}$$

From this we can write $\text{Power}_{\max}/2 \approx \text{Power}_{\max} - f_{1/2}^2/(2\sigma_f^2)$ or $\sigma_f \approx f_{1/2}/\sqrt{\text{Power}_{\max}}$. In terms of signal-to-noise ratio $\Sigma = \text{rms}[(y_n - \mu)/\sigma_n]$, where rms is the root mean square. A well-fit model gives $P_{\max} \approx \hat{\chi}_0^2/2 \approx \Sigma^2 N/2$ which leads to the expression in Equation 52 of VanderPlas (2018)

$$\sigma_f \approx f_{1/2}\sqrt{\frac{2}{N\Sigma^2}}. \tag{B.8}$$

To obtain $f_{1/2}$, we compute the half maximum power and define a left boundary corresponding to the frequency to the left of the peak where the power is approximately is half of the maximum power. $f_{1/2}$ is then simply the subtraction between the frequency of the maximum power and the left boundary. We use left and not right boundary because for some cases, when considering period and not frequency, the right side of the peak is not well.

To compute the SNR Σ we fit a sinusoisal model using *curve_fit* of the form

$$y = A \sin\left(2\pi \frac{t}{P_{\max}} + \phi\right) + \omega, \tag{B.9}$$

where $P_{\max}$ is the period of the best peak, fixed in the fitting of the function. The expected value μ is then the value retrieved by applying this model to our data and σ_n is the error in S_{CaII}. The error in frequency σ_f is converted to period by

$$\text{Period error} = \frac{\text{Frequency Error}}{\text{Best frequency}^2}. \tag{B.10}$$

Appendix B.4: Gaussian fitting

The fourth method we implemented consists in fitting a Gaussian of the form

$$\text{Gaussian(x)} = H + A\exp\left[-(x - x_0)^2/(2\sigma^2)\right], \tag{B.11}$$

where H is an offset, A is the amplitude, x is the period array from the GLS periodogram, σ is the standard deviation and x_0 is the expected value of the Gaussian. We used *curve_fit* in a region between 90% and 110% of the period corresponding to maximum power peak. As bounds, H is restricted between 0 and

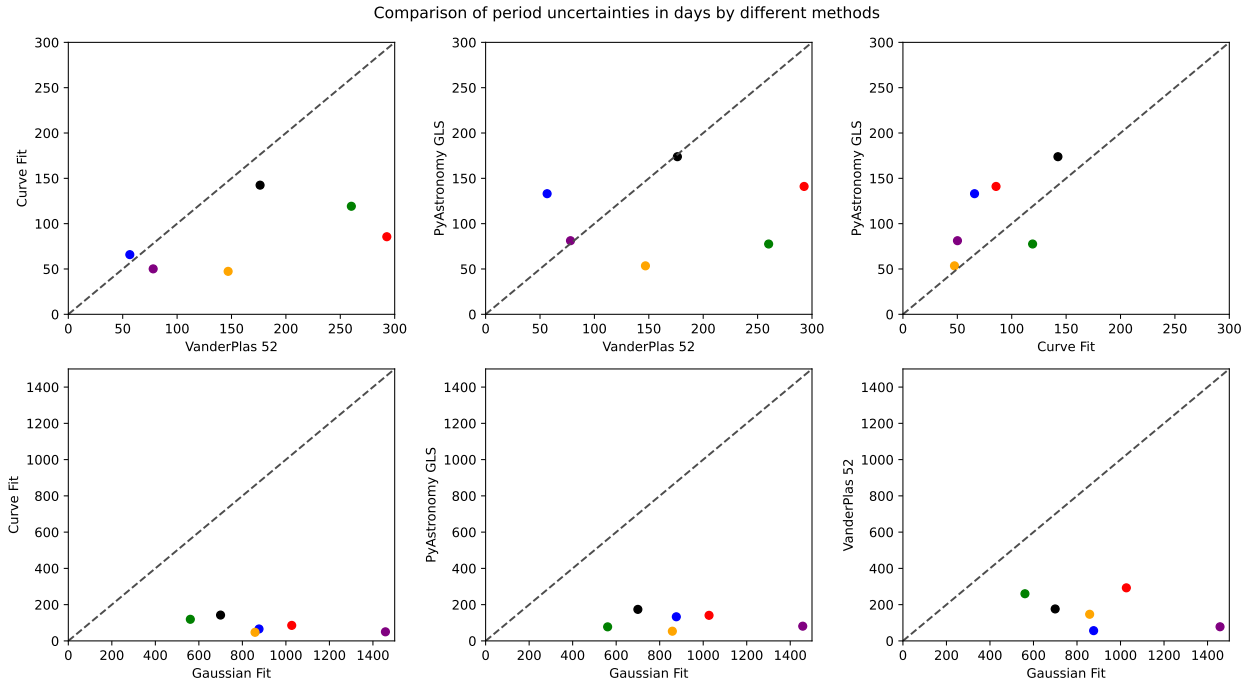

Fig. B.1: 1-1 comparisons between the four methods studied to estimate the period error, using HARPS data. The blue point is HD209100, the black point is HD115617, the red point is HD1461, the green point is HD10647, the orange point is HD85512 and the purple point is HD192310.

$+\infty$, A is restricted between $0.5\times P_{\rm max}$ and $1.5\times P_{\rm max}$, x_0 between $0.99\times$Period$_{\rm max}$ and $1.01\times$Period$_{\rm max}$ and σ is restricted between 0 and $0.2\times$ Period$_{\rm max}$ (to try to obtain a reasonable value). The standard deviation σ of the Gaussian is taken as the error for the activity period. A major issue with this method is that it retrieves very high values for σ, probably because the region where we fit the Gaussian is very narrow, having only a few points (depending on the step defined for the GLS periodogram).