Chapter 7

Variability

Author(s): Laurent Eyer, Leanne Guy, Elisa Distefano, Gisella Clementini, Nami Mowlavi, Lorenzo Rimoldini, Maroussia Roelens, Marc Audard, Berry Holl, Alessandro Lanzafame, Thomas Lebzelter, Isabelle Lecoeur-Taïbi, László Molnár, Vincenzo Ripepi, Luis Sarro, Grégory Jevardat de Fombelle, Krzysztof Nienartowicz, Joris De Ridder, Áron Juhász, Roberto Molinaro, Emese Plachy, Sara Regibo

7.1 Introduction

Author(s): Laurent Eyer

This chapter presents the models and methods used on the Gaia 22 months data to produce the Gaia variable star results for Gaia DR2. The variability processing and analysis was based mostly on the calibrated G, and integrated G_{BP} and G_{RP} photometry.

The variability analysis approach to the Gaia data was described in Eyer et al. (2017), and the Gaia DR2 results are presented in Holl et al. (2018). Detailed methods on specific topics will be published in a number of separate articles, after the data release date. Variability behaviours in colour magnitude diagrams will be presented in Gaia Collaboration et al. (2018c).

This Chapter 7 is organised as follows: the global processing is described in Section 7.2 and subsequent Sections present different data products: the whole sky classification in Section 7.3, RR Lyrae star and Cepheid candidates in Section 7.4, BY Draconis candidates in Section 7.5, short time scale variability in Section 7.6, and long period variable stars in Section 7.7.

7.1.1 Overview

The Variability Processing and Analysis Coordination Unit (CU7) and its associated Data Processing Centre in Geneva (DPCG) gather about 60 people, spread in 18 institutes mostly in Europe (in addition there are contributions of Tel Aviv University, Israel and of Villanova University, USA). The approach to the successive data releases is iterative.

In this second data release we make a significant jump from the first data release: in Gaia DR1, we released 3194 Cepheid and RR Lyrae star candidates and in Gaia DR2 we reach more than 550 000 stars, with 6 variability types. We classified also eclipsing binaries and QSOs that were passed to other coordination units to be analysed and published in Gaia DR3.

7.1.2 Data products in Gaia DR2

Variability products in Gaia DR2 include:

- a whole sky classification of several variability types namely SX Phoenicis/δ Scuti stars, RR Lyrae stars, Cepheids, and long period variables;
- specific object studies for the following variability types: RR Lyrae stars, Cepheids, long period variables, as well as solar-like (magnetic) activity (BY Draconis stars);
- a specific search on the short time scale variability.

The time series in G, $G_{\rm BP}$ and $G_{\rm RP}$ of the CU7 released sources are available in the archive.

The analysis is done with automated methods and the published stars should be considered as *candidates* of variability or of specific variability types.

7.2 Global processing

7.2.1 Introduction

Author(s): Berry Holl, Grégory Jevardat de Fombelle

The variability processing aims at detecting and analysing the variability of the calibrated time series. It consists of multiple processing steps implemented by modules that use various inputs from other CUs and produce various output results.

7.2.2 Properties of the input data

Author(s): Berry Holl, Lorenzo Rimoldini, Krzysztof Nienartowicz, Leanne Guy, Marc Audard, Laurent Eyer, Grégory Jevardat de Fombelle

7.2.2.1 Astrometry

Astrometric information consisting of position and, where available, parallax, proper motion and attributes derived from the parallax was ingested in our catalogues. In this processing the positions have been used for the creation of our cross-match catalogues and the parallax with associated uncertainty in our supervised classification and in all of the specific object modules. The astrometric data reduction available at the time of variability processing was a precursor of the final astrometry published in Gaia DR2, as explained in Figure 7.2. Therefore small deviations in used parallaxes are to be expected, for example in the attributes employed for classification, or the absolute magnitudes for the long period variable module.

7.2.2.2 Photometry

CU5 photometry from July 25, 2018 to May 23, 2016, i.e. 22 months, was the main input for our variability results in Gaia DR2. It contains the G, G_{BP} and G_{RP} photometric bands. Although per-CCD data were available for a subset of the sources, such data were used by the module short time-scale, but excluded from publication in Gaia DR2. The Bronze sources as defined in Section 5.4.3 were not investigated.

7.2.2.3 Spectroscopy (RVS)

RVS instrument data were not available for the variability processing for Gaia DR2.

7.2.2.4 Astrophysical parameters

Astrophysical parameters were not available for the variability processing for Gaia DR2.

7.2.2.5 Source selection criteria

As described in Section 7.2.3.1, we selected sources with either ≥ 2 G-FoV transits or ≥ 20 G-FoV transits for two different processing paths, which partly overlapped in some of the final stages.

7.2.3 Processing steps

Author(s): Leanne Guy, Berry Holl, Marc Audard, Alessandro Lanzafame, Isabelle Lecoeur-Taïbi, Nami Mowlavi, Lorenzo Rimoldini, Joris De Ridder, Luis Sarro, Sara Regibo, Grégory Jevardat de Fombelle

7.2.3.1 Overview

An overview of the variability processing is presented in Figure 7.1. There are two main paths: one starting from ≥ 2 G-FoV transits (left) and one from ≥ 20 G-FoV transits (right). The former results in the published *nTransits:2+* classification results, and the latter results in the published Specific Object Tables of: `vari_short_timescale` (Section 14.3.8) and `vari_rotation_modulation` (Section 14.3.6). The published Specific Object Tables of: `vari_rrlyrae` (Section 14.3.7), `vari_cepheid` (Section 14.3.1), and `vari_long_period_variable` (Section 14.3.5) result from a mixed feed of classification candidates from the published *nTransits:2+* classifier (for sources with a minimum of 12 G-FoV transits) and from the unpublished *nTransits:20+* classifier. The data was published from the highlighted yellow boxes for sources that passed the validation filtering.

Figure 7.2 details how the sources were cross-matched on a preliminary version of the photometry.

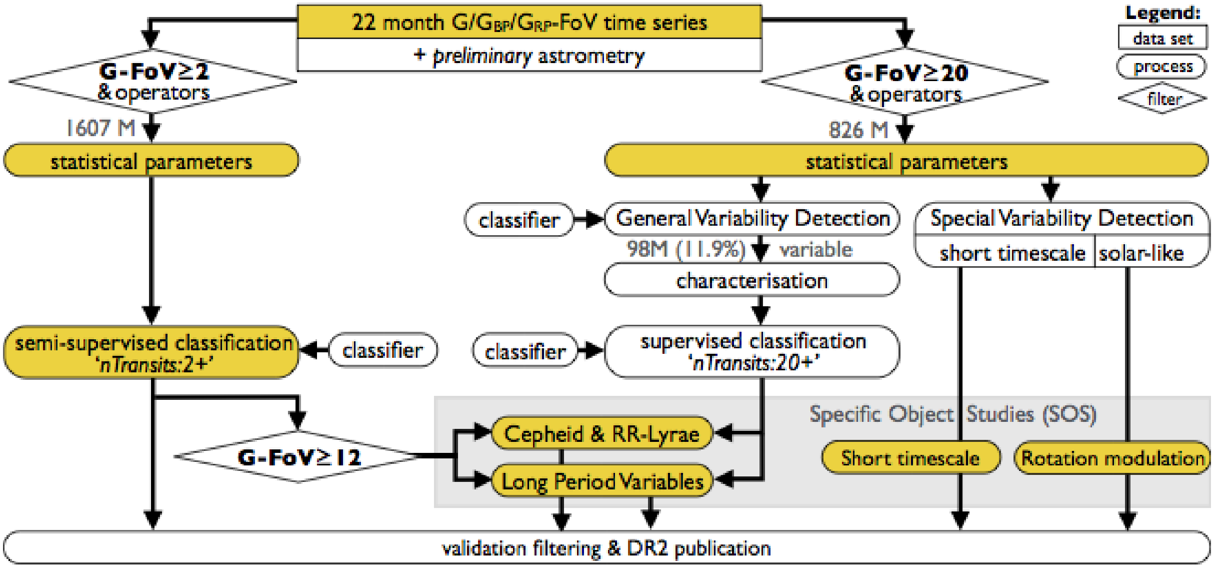

Figure 7.1: Gaia DR2 variability processing overview. The data products appearing in Gaia DR2 (yellow boxes) are either coming from a whole sky classification of *nTransits:2+* or from specific objects studies. Note that *nTransits:20+* classification is not published. Figure 7.2 details how the three classifiers were trained.

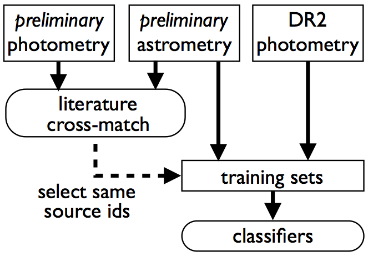

Figure 7.2: Gaia DR2 variability classifier training for the three classifiers in Figure 7.1. The cross-match on external catalogues was performed using a preliminary version of the photometry and astrometry. The final training of the models was performed on the published photometry (though still with preliminary astrometry) on the sources identified using the preliminary photometry.

7.2.3.2 Initial light curve pre-processing

7.2.3.2.1 Definition of observation time Observation times are expressed in units of Barycentric JD (in TCB) −2 455 197.5 days, computed as follows:

1. The observation time is converted from On-board Mission Time (OBMT) into Julian date in TCB (Temps Coordonnée Barycentrique).
2. A correction is applied for the light-travel time to the Solar system barycentre, resulting in Barycentric Julian Date (BJD).

3. Although the centroiding time accuracy of the individual CCD observations is (much) below 1 ms, the per-field-of-view observation times processed and published in Gaia DR2 are averaged over typically 9 CCD observations over a time range of about 44 sec.

Conversion from flux to magnitude In the variability pipeline, magnitudes rather than fluxes are used in the various processing modules. To convert to magnitude, the zero-point magnitudes for G, $G_{\rm BP}$, $G_{\rm RP}$ provided by CU5 in the Vega system are used (Section 5.3.6.6).

Observation filtering The variability processing includes several *operators* that are applied to the ingested and reconstructed photometry. Typical time series operators perform flux to magnitude conversion, outlier removal and error cleaning on the input time series to create derived (transformed and/or filtered) time series suitable for processing by specific algorithms. Chaining of these time series operators creates a hierarchy of derived time series that is used as required by the scientific analyses while ensuring that provenance is preserved.

The following list of operators are applied in sequence to the input photometric time series, a schematic showing the hierarchy of these operators is presented in Figure 7.3:

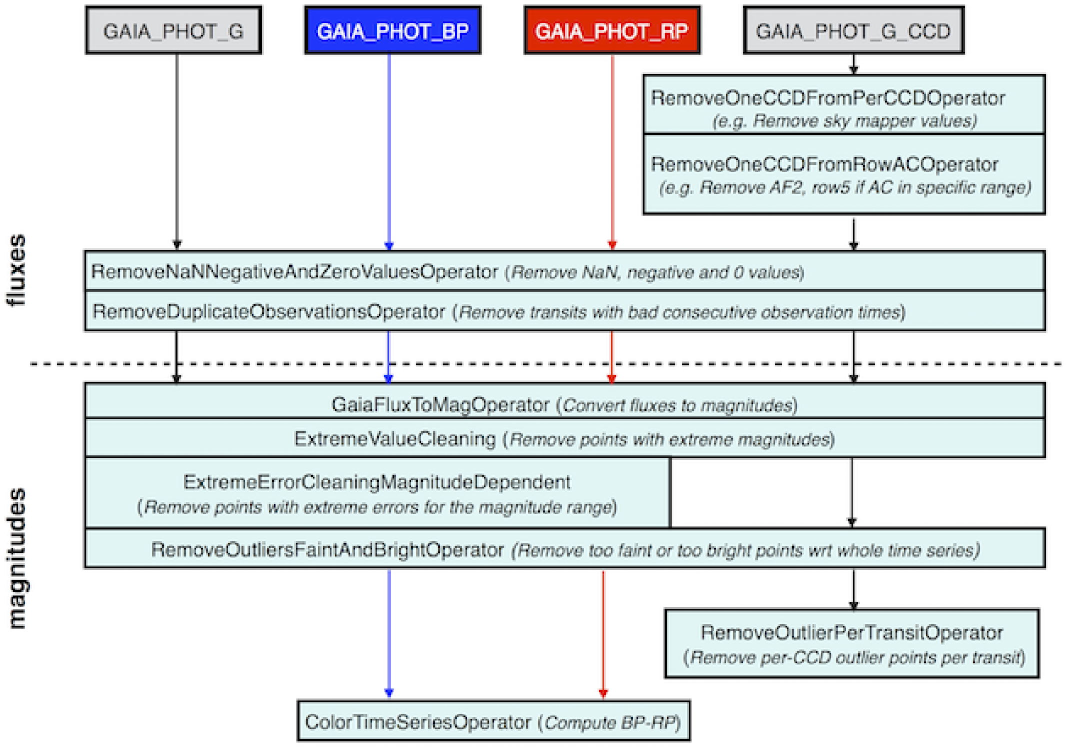
GAIA_PHOT_G
GAIA_PHOT_BP
GAIA_PHOT_RP
GAIA_PHOT_G_CCD
RemoveOneCCDFromPerCCDOperator (e.g. Remove sky mapper values)
RemoveOneCCDFromRowACOperator (e.g. Remove AF2, row5 if AC in specific range)
fluxes
RemoveNaNNegativeAndZeroValuesOperator (Remove NaN, negative and 0 values)
RemoveDuplicateObservationsOperator (Remove transits with bad consecutive observation times)
magnitudes
GaiaFluxToMagOperator (Convert fluxes to magnitudes)
ExtremeValueCleaning (Remove points with extreme magnitudes)
ExtremeErrorCleaningMagnitudeDependent (Remove points with extreme errors for the magnitude range)
RemoveOutliersFaintAndBrightOperator (Remove too faint or too bright points wrt whole time series)
RemoveOutlierPerTransitOperator (Remove per-CCD outlier points per transit)
ColorTimeSeriesOperator (Compute BP-RP)

Figure 7.3: The CU7 operator chain to transform and filter time series.

1. `RemoveNaNNegativeAndZeroValuesOperator`: it removes photometric transits that contain NaN, negative, or 0 flux values.

2. `RemoveDuplicateObservationsOperator`: it removes pairs of transits (within each Gaia band) that are too close in time (within 105 min) to be observations of the same source. Such transits can occur in bright sources for which multiple artefact detections are assigned (these are known as 'far double detections'). Since CU7 did not have access to the flags identifying the double detections transits, this ad-hoc method was applied by which close-in-time pairs of transits were removed.

3. `RemoveOneCCDFromRowAcOperator`: it is designed to remove one CCD point (defined by its CCD number, between 0 and 9, 0 standing for the Sky Mapper and 1 to 9 for the Astrometric Field CCDs) from each transit of G per-CCD data whose measurements correspond to a certain CCD row and whose across-scan (AC) coordinate is outside a certain range (minimum AC=3, maximum AC=1990). It was motivated by the fact that the photometric calibration team reported problematic flux measurements for the second Astrometric Field (AF) CCD of row 5 when AC is greater than 1200. Hence, in Gaia DR2 this operator was tailored to remove the AF2 points for transits with CCD row=5 and AC coordinate > 1200.

4. `GaiaFluxToMagOperator`: it converts fluxes to magnitudes by using the zero-point magnitudes delivered by CU5.

5. `ExtremeValueCleaning`: it removes points above a specific magnitude limit. Cuts were applied at $G = 25$, $G_{\rm BP} = 24$, and $G_{\rm RP} = 22$ mag.

6. `ExtremeErrorCleaningMagnitudeDependent`: it removes individual transits above or below magnitude-dependent values. The values were determined as follows: from a sample of the CU5 photometric catalogue and for each band (limited to 6000 sources per 0.1 magnitude bin), we studied the quantile distribution of the transit magnitude errors. A decision was made to use the 99.7% quantile for the upper value, and the 0.01% for the lower value for G data. For $G_{\rm BP}$ and $G_{\rm RP}$, a cut was applied only above the upper value of the 99.9% quantile. Figure 7.2.3.2.1 shows the distributions of the transit magnitude errors as a function of the transit magnitudes, for the 3 Gaia bands, together with the thresholds used for this operator. The latter was not applied to G per-CCD data.

7. `RemoveOutliersFaintAndBrightOperator`: it removes data points as follows (with configuration parameters described, where relevant, in the respective data product sections).

   (a) A point with a too large error (intrinsically or compared to some number of times the interquartile range of the uncertainties) is an outlier. It is removed before the next step.

   (b) Measurements at the extremes of the magnitude distribution of a time series are identified from their deviations from the median magnitude when these exceed a certain number of times the interquartile range (with different thresholds possible at the bright and faint ends). A point with an 'extreme' magnitude (on the faintest or brightest side, compared to the median magnitude) is an outlier unless it has similar outlying neighbours in time or projected in magnitude.

8. `RemoveOutlierPerTransitOperator`: it removes per-CCD outlier data points per transit. This operator only applies to per-CCD data.

9. `ColorTimeSeriesOperator`: it is applied to the $G_{\rm BP}$ and $G_{\rm RP}$ light curves to compute the $G_{\rm BP}-G_{\rm RP}$ colour.

The Gaia DR2 time series are published in a Virtual Observatory table linked in `gaia_source` via the column `epoch_photometry_url`. The table includes a flag, `rejected_by_variability`, that provides information on which data points in each band were rejected by the hierarchical chain of CU7 operators up to and including `RemoveOutliersFaintAndBrightOperator`. Note that downstream CU7 modules may reject additional points, e.g., by applying stricter thresholds for `RemoveOutliersFaintAndBrightOperator`, however, such rejected points are not flagged in the Gaia DR2 archive. We mention that CU5 flags were not used in variability processing.

However they are available in the Gaia DR2 archive in column rejected_by_photometry of epoch_photometry_url (see also Section 14.3.9).

Published output See Gaia DR2 VO Table linked in column epoch_photometry_url of table gaia_source.

7.2.3.3 Statistical parameter computation

Input All cleaned time series (Section 7.2.3.2.1) in magnitude with at least one field-of-view transit.

Method The first step in the scientific processing chain following conversion from flux to magnitude and basic cleaning (Section 7.2.3.2.1) is the computation of a number of basic descriptive, inferential and correlation statistics of all light curves. These statistics provide a first general overview of the data and their distributions and are used to determine whether variability is present in a time series of Gaia observations.

Descriptive statistics computed on the temporal evolution of the time series include (but are not limited to): the number of observations, time duration of the time series, mean observation time and the min/max time difference between two successive observations. Given the well defined nature of the Gaia scanning law and the angular separation between the 2 telescopes, the latter can be useful in identifying transits assigned to the wrong source.

Parameters that characterise the brightness of the light curve and the associated uncertainty include measures of the min, max, range, mean, median, variance, skewness, kurtosis, point-to-point scatter, interquartile range, median absolute deviation, and the signal-to-noise ratio. Where applicable, unbiased weighted and unweighted estimators as well as robust estimates are computed and compared, as they can be useful in identifying outliers, transits assigned to the wrong source or signatures of variability.

Several inferential test statistics are computed on the time series including the Kolmogorov-Smirnov (K-S) test for equality of continuous distributions, (Kolmogorov 1933; Smirnov 1939), the Ljung-Box test for randomness, (Ljung & Box 1978), the Abbe hypothesis test, (von Neumann 1941, 1942) as well as the chi-squared and Stetson test statistics, (Stetson 1996). These measures are used in the classification of a time series as either constant or variable. Only unbiased, unweighted and robust quantities are available for all Gaia DR2 time series in the Gaia catalogue.

Correlation statistics between all pairs of the three photometric bands are computed for use in the detection of general and special variability (Section 7.2.3.4). Stetson, Pearson and Spearman correlation statistics are computed on all permutations of pairs of the three photometric bands, G, $G_{\rm BP}$ and $G_{\rm RP}$. Computation of the Stetson correlation requires that observations in each band are paired. As each band may have a different number of FoV transits, correct pairing of observations between bands is done by requiring that their time difference is less than 0.05 days. This ensures that paired observations in each band were observed in the same transit. For the Pearson and Spearman correlation statistics, the time series are filtered to remove unpaired observations. The correlation is hence performed on time series of equal length and containing only paired observations.

Run-time configuration parameters The variance, skewness and kurtosis, including weighted, unweighted and robust versions, were all computed with a sample-size bias correction.

Published output See Gaia DR2 table: vari_time_series_statistics.

7.2.3.4 Variability Detection

Description of general and special variability detection strategies.

Input Variability analyses was only performed on field-of-view averaged photometry in G, G_{BP}, and G_{RP} bands.

Method In this data release, General Variability Detection (GVD) employed a supervised classifier trained on a set of identified constants and variables. Variable objects were selected from sources of different variability types derived from the crossmatch with a large number of literature catalogues (Section 7.3.3.2); they included 14 769 sources which covered most of the range of magnitudes of the data in Gaia DR2. On the other hand, constant objects were limited to crossmatched sources from a few catalogues (the least varying sources in OGLE-IV at `ftp://ftp.astrouw.edu.pl/ogle/ogle4/GSEP/maps/`, the Hipparcos constants in ESA 1997, and the SDSS standards in Ivezić et al. 2007), thus they lacked representatives in a significant magnitude range (from about 10 to 15 mag in the G band). A semi-supervised approach was employed to supplement the training set with constant objects identified in a previous iteration of variable versus constant classification, filling the gap in the magnitude distribution and leading to a total sample of 14 424 constants. The selected variable and constant objects were then characterized by time series statistics as well as average photometric quantities in order to train a Random Forest classifier, which returned an estimated completeness of at least 98% and a contamination rate of up to 2%. This classifier was applied to all 826 million sources with 20 or more G-band field-of-view transits.

A source was considered constant or variable when the highest posterior probability class referred to either 'constant' or 'variable', respectively.

No p-value statistics were used or analysed for GVD in this data release.

Run-time configuration parameters The minimum classification probability to consider an object as variable was set to 50%.

Published output No data from this processing step was published in Gaia DR2. The output of this step is used as input to the general classification step (see Section 7.2.3.6).

7.2.3.5 Period search and time series modelling

Input Period search and Fourier modelling were applied to cleaned (Section 7.2.3.2.1) G-band time series (expressed in magnitudes as a function of time in days) with at least five FoV transits, for sources identified as variable (Section 7.2.3.4). These methods rely also on the availability of statistical parameters (Section 7.2.3.3).

Method The process of frequency (or period) search and time series modelling, referred to collectively as *Variability Characterization*, aims to characterize the variability behaviour of time series of Gaia observations using a classical Fourier decomposition approach. The model to fit is given by Equation 7.1. The Characterization process takes as input all time series identified as variable by the preceding *Variability Detection* module (see Section 7.2.3.4). The goal is to produce, in an automated manner, the simplest and statistically most significant model of the observed variability.

The general model of variability that we fit to time series of Gaia observations is given by:

$$y = \sum_{n=1}^{N_f} \sum_{k=1}^{N_h(n)} A_{n,k} \cos(2\pi k f_n t + \psi_{n,k}) + \sum_{i=0}^{N_p} c_i t^i \tag{7.1}$$

where we assume that the reference epoch t_{ref}, the middle of the time series, has already been subtracted from the time points. $N_p \geq 0$ is the degree of the polynomial, $N_f \geq 0$ is the number of detected frequencies, and $N_h(n) \geq 1$ is the number of significant harmonics of frequency f_n. This multi-frequency harmonic model includes a low-order polynomial trend and n frequencies, each with k associated harmonics.

Run-time configuration parameters

1. For frequency search:
   (a) At least ≥5 FoV G transits.
   (b) No de-trending applied prior to the frequency search.
   (c) Frequency searched with the Least Square method.
   (d) Minimum frequency: $1.5\,(\Delta T)^{-1}$ d^{-1} with ΔT denoting the total time span of each time series.
   (e) Maximum frequency: 20 d^{-1}.
   (f) Frequency step: $(10\,\Delta T)^{-1}$ d^{-1} with ΔT denoting the total time span of each time series.
   (g) Refinement of the frequency about the most significant peak was done to a granularity of 10^{-6} d^{-1}.
2. For modelling:
   (a) The polynomial part of Equation 7.1 was limited to degree zero.
   (b) Unweighted observations were used in the fit.
   (c) Non-linear fitting with the Levenberg-Marquard method was applied to the parameters of the final best model.

Published output No data from this processing step is published in Gaia DR2. The results of this step are used as input to the general classification step (see Section 7.2.3.6).

7.2.3.6 Classification

Two classification paths were followed for Gaia DR2.

1. The *nTransits:2+* classification aimed at covering the whole sky. The results of this classifier can be found in the Gaia archive for selected high-amplitude pulsating variable types (δ Scuti/SX Phoenicis stars, RR Lyrae stars, Cepheids, long period variables). This module fed also the specific object modules CEP&RRL and LPV when there were more than 12 G-FoV transits, as shown in Figure 7.1. The details of this classifier are described in Section 7.3.
2. The *nTransits:20+* classification made use of the period search and modelling results from the characterisation module. This classifier fed (as a secondary input) the specific object modules CEP&RRL and LPV, as shown in Figure 7.1. No direct output of this classification is provided in the archive.

This section describes only the *nTransits:20+* classification, while the *nTransits:2+* classification is presented in Section 7.3.

Input The *nTransits:20+* classifier is trained with attributes computed from the results of the Statistical Parameter Computation module (Section 7.2.3.3), the period search and time series modelling modules (Section 7.2.3.5), and it is applied to sources selected by the Variability Detection module (Section 7.2.3.4).

Method The module produces membership probabilities for all sources with at least 20 field-of-view transits. The membership probabilities are obtained in two stages. In the first stage, three different classifiers (Gaussian Mixtures, Bayesian Networks, and Random Forest) produce corresponding membership probabilities based on different attribute sets. In the second stage, a meta-classifier takes as input a set of classification probabilities (denoting for each source the posterior probabilities associated with different types) from the predictions of the individual classifiers to produce the final result. The meta-classifier method is again Random Forest.

Run-time configuration parameters The four classifiers (three in stage 1 and the meta-classifier) define the input attributes via an attribute mapping that transforms the output from the previous modules into suitable classification attributes. The classifiers based on Gaussian Mixtures and Bayesian Networks use the following list of attributes:

1. the first detected frequency;
2. the decadic logarithm of the amplitude of the first and second harmonics of the first detected frequency (two separate attributes);
3. the $G_{\rm BP} - G_{\rm RP}$ (possibly reddened) colour index;
4. the robust percentile-based skewness (as in Eyer et al. 2017);
5. the phase difference between the first two Fourier components of the first detected frequency (after setting the phase of the first term to zero);
6. the decadic logarithm of χ^2_{QSO}/ν as defined in Butler & Bloom (2011);
7. the decadic logarithm of χ^2_{false}/ν as defined in Butler & Bloom (2011).

The Random Forest in the first stage uses the following list of attributes:

1. the first detected frequency;
2. the $G_{\rm BP} - G_{\rm RP}$ (possibly reddened) colour index;
3. the $G_{\rm BP} - G$ (possibly reddened) colour index;
4. the G-band Stetson variability index (Stetson 1996), pairing observations within 0.1 days;
5. the reduced chi-squared statistic of the G-band time series with respect to the constant brightness model;
6. the sample-size (un)biased (un)weighted variance (two attributes) unbiased by Gaussian uncertainties, in the G-band time series (Rimoldini 2014);

7. the median absolute slope (in mag d^{-1}) of the G-band time series within a sliding window of half a day;
8. the median range of the G-band time series within a sliding window of half a day;
9. the interquartile range of the G-band magnitude distribution;
10. the decadic logarithm of χ^2_{false}/ν as defined in Butler & Bloom (2011);
11. the sample-size biased unweighted skewness moment standardised by its variance, in the G-band time series (Rimoldini 2014).

The Random Forest meta-classifier uses as attributes the posterior probabilities for each class estimated by the stage 1 classifiers.

In all cases, the classification scheme comprises the following classes:

1. Cepheid type stars (all subtypes included);
2. RR Lyrae type stars (all subtypes included);
3. Eclipsing binaries (all subtypes included);
4. δ Scuti/γ Doradus sources;
5. Long Period Variables (Semi-regular variables, Mira stars);
6. Quasars;
7. Other (including all other types of variability not included in the previous types).

The combined category of δ Scuti/γ Doradus sources is not separated (despite their different typical periods) due to the significant contamination observed because of aliasing.

Each classifier is defined by a set of parameters the specification of which is out of the scope of this documentation. They include the number of trees in each Random Forest, their maximum depth, the number of attributes used in each node and the minimum number of instances per class at the leaf nodes; for Bayesian Networks and Gaussian Mixtures, the configuration included a multi-stage scheme used for separating the classes and the attributes used in each node; for Gaussian Mixtures, the minimum and maximum number of components was set for each class.

Published output Only the *nTransits:2+* classification results were published in the Gaia DR2 tables: `vari_classifier_definition`, `vari_classifier_class_definition`, `vari_classifier_result`.

7.2.3.7 Specific Object Studies

Some variable objects benefit from additional processing that takes into account the specific properties of their variability. The Specific Object Studies (SOS) component of the variability pipeline comprises a number of dedicated modules that aim to compute attributes specific to a variability class, and subsequently publish them in the Gaia DR2 archive. Each SOS module takes as input either the list of candidates of the corresponding variability class, as provided by the classification step (see Sect. 7.2.3.6), using a probability threshold specific to each SOS module or from the selection made in the special variability module.

Details of the selection criteria, processing, and of the output data products of each SOS module are described in the respective data product sections.

Input Source selections depend on specific SOS modules and are described in the relevant data product sections.

Method Methods are described in the relevant data product sections.

Run-time configuration parameters Run-time configuration parameters are described in the relevant data product sections.

Published output See Gaia DR2 tables: `vari_short_timescale` (Section 14.3.8), `vari_rotation_modulation` (Section 14.3.6), `vari_rrlyrae` (Section 14.3.7), `vari_cepheid` (Section 14.3.1), and `vari_long_period_variable` (Section 14.3.5).

7.2.4 Quality assessment and validation

Author(s): Leanne Guy, Laurent Eyer, Grégory Jevardat de Fombelle

7.2.4.1 Verification

Extensive verifications were done on the outputs of the variability processing. A set of 430 verification rules were defined and implemented. It allowed the automatic verification of each output result of each module. Such verifications rules including but not limited to range checks, cardinality, nullity conditions allowed to fix a number of bugs and filter incorrect results. On top of that, each module made supplementary verifications that are explained within each of the following sections.

7.2.4.2 Validation

Validations of period search with external catalogues, validation of general classification with respect to other surveys were done. The validations of the different published variable star catalogues are explained within each of the corresponding sections.

7.3 All-sky classification

Author(s): Lorenzo Rimoldini

The all-sky classification results are published in the Gaia DR2 table `vari_classifier_result` and include candidates for almost two hundred thousand RR Lyrae stars, about one hundred fifty thousand long period variables, more than eight thousand Cepheids and a similar number of SX Phoenicis/δ Scuti stars. A subset of these candidates was further processed by subsequent modules of the CU7 pipeline (Section 7.2.3.7), such as the ones of Cepheid and RR Lyrae stars (Section 7.4) and of long period variables (Section 7.7). Other candidates were verified and validated by means of comparisons with the literature and included known misclassifications, which were nevertheless not removed in order to minimise sample selection effects and maintain the distributions of parameters more homogeneous for statistical analyses. The community is expected to take this cautionary note into account when exploiting this data set.

7.3.1 Introduction

An advance publication of the first Gaia full-sky map of Cepheids, RR Lyrae stars, SX Phoenicis/δ Scuti stars and long period variables is provided by automated classification of all objects with at least two FoV transits in the G band. The results of this classification can be found in the Gaia DR2 archive in the classification table associated with the *nTransits:2+* classifier, although subsequent filtering of sources by CU3 and CU9 increased the minimum number of FoV transits to five (after taking into account also the CU7 observation filtering of the pre-processing step described in Section 7.2.3.2).

7.3.2 Properties of the input data

Machine-learning classifiers were trained with Gaia sources selected from over seven hundred fifty thousand objects crossmatched with the literature, representing a large number of variability types as well as non-varying objects. The training set included about thirty-three thousand sources filtered according to their distribution in the sky, their number of FoV transits, and their median magnitudes in the G band, as described in more details in Section 7.3.3.

All sources with two or more FoV transits in the G band were processed by the classifiers. Photometric time series in the G, $G_{\rm BP}$, and $G_{\rm RP}$ bands were used after the pre-processing steps described in Section 7.2.3.2 and astrometric quantities (such as parallax and proper motion) were employed without specific selections. The results of the Statistical Parameter Computation module (Section 7.2.3.3) provided additional input information which was used directly as classification attributes or in the computation thereof.

7.3.3 Processing steps

The results of all-sky classification were obtained through the following steps.

1. Crossmatch of Gaia with literature to identify objects of known classes (Section 7.3.3.2).
2. Selection of catalogues to crossmatch and their prioritisation (in case of conflictual information on the same objects).
3. Filtering of sources not satisfying simple statistics (such as colour, magnitude, literature period, amplitude, skewness, and Abbe value computed on magnitudes sorted in time as well as in phase) that are typical of class ownership, while allowing for a large range of possible distance, extinction, and reddening.
4. Resampling of sources for a more representative distribution in the sky, in the number of FoV transits, and in magnitude.

5. Pipeline run of the Statistics module on time series pre-processed as described in Section 7.2.3.2.
6. Generation and selection of classification attributes (Section 7.3.3.3).
7. Training of a multi-stage classifier with optimized parameters.
8. Application of the multi-stage classifier to the Gaia data.
9. Improvement of the training set (sources and attributes) including high-confidence classifications and iterating steps 3–6 (Section 7.3.3.5).
10. Training of the improved multi-stage classifier with optimized parameters (Section 7.3.3.4).
11. Pipeline run of the Statistics and the Classification modules on time series pre-processed as described in Section 7.2.3.2.
12. Training of contamination-cleaning classifiers and their application to the results of the previous step, for RR Lyrae stars, Cepheids, and SX Phoenicis/δ Scuti stars (Section 7.3.3.6).
13. Definition of classification scores of the published results (Section 7.3.3.7).
14. Assessment of completeness and contamination of the published results (Section 7.3.4).

7.3.3.1 Classes

The training set included objects of the classes targeted for publication in Gaia DR2 (listed in bold) as well as other types to reduce the contamination of the published classification results. The full list of object classes, with labels (used in the rest of this section) and corresponding descriptions, follows below.

1. **ACEP**: Anomalous Cepheids.
2. ACV: α^2 Canum Venaticorum-type stars.
3. ACYG: α Cygni-type stars.
4. **ARRD**: Anomalous double-mode RR Lyrae stars.
5. BCEP: β Cephei-type stars.
6. BLAP: Blue large amplitude pulsators.
7. **CEP**: Classical (δ) Cepheids.
8. CONSTANT: Objects whose variations (or absence thereof) are consistent with those of constant sources (Section 7.2.3.4).
9. CV: Cataclysmic variables of unspecified type.
10. **DSCT**: δ Scuti-type stars.
11. ECL: Eclipsing binary stars.
12. ELL: Rotating ellipsoidal variable stars (in close binary systems).
13. FLARES: Magnetically active stars displaying flares.
14. GCAS: γ Cassiopeiae-type stars.

15. GDOR: γ Doradus-type stars.
16. **MIRA**: Long period variable stars of the *o* (omicron) Ceti type (Mira).
17. OSARG: OGLE small amplitude red giant variable stars.
18. QSO: Optically variable quasi-stellar extragalactic sources.
19. ROT: Rotation modulation in solar-like stars due to magnetic activity (spots).
20. **RRAB**: Fundamental-mode RR Lyrae stars.
21. **RRC**: First-overtone RR Lyrae stars.
22. **RRD**: Double-mode RR Lyrae stars.
23. RS: RS Canum Venaticorum-type stars.
24. SOLARLIKE: Stars with solar-like variability induced by magnetic activity (flares, spots, and rotational modulation).
25. SPB: Slowly pulsating B-type stars.
26. SXARI: SX Arietis-type stars.
27. **SXPHE**: SX Phoenicis-type stars.
28. **SR**: Long period variable stars of the semiregular type.
29. **T2CEP**: Type-II Cepheids.

7.3.3.2 Crossmatch with literature

Training-set objects are selected from Gaia sources crossmatched with objects associated with known classes in the literature. In order to increase the reliability of crossmatch results, a set of metrics was used in the comparison of Gaia and literature sources, always including the angular separation, and whenever possible also the time-series median magnitude in the G band, the $G_{\rm BP} - G_{\rm RP}$ colour, as well as time series quantities characterising the amplitude of variations in the G band such as the range or standard deviation. Such metrics were combined in a multi-dimensional distance which was minimised in an iterative process in order to allow for the tuning of empirical relations between the Gaia and literature photometric quantities (affected in particular by the different bandwidth coverage and sensitivity). The best matches were projected onto planes for all combinations of crossmatch metrics to inspect the corresponding distributions and reduce the chance of mis-matches by applying thresholds to exclude dubious outliers and excessive tails of the distributions. Although this approach sacrificed completeness in some cases, it was considered appropriate for training purposes, given the large number of sources available.

In order to sample as many regions of the sky as possible, cover most of the range of Gaia magnitudes, and include a large number of variability types, a multitude and variety of catalogues were selected from a larger set, following general reliability considerations, and prioritised in case of conflicting classifications for the same sources. The full list of catalogues employed in the training sets are presented in Table 7.1, including references and crossmatch metrics. Among the over seven hundred fifty thousand crossmatched objects available for training, only a small sample (of about 33 thousand sources) was vetted to train classifiers (Section 7.3.3.4), leaving many reliable crossmatches for the validation of results (Section 7.3.4).

Table 7.1: Crossmatch of (mostly) variable objects from the literature selected for the training set. The Table includes names of surveys and/or variability types (specified by the labels defined in Section 7.3.3.1), references, and crossmatch metrics: angular separation (AS), time-series median G-band magnitude (M) and $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ colour (C), time-series G-band magnitude range (R) and standard deviation (SD).

Description	Reference	Crossmatch Metrics
ASAS All-Star Catalog: solar-like stars	Messina et al. (2010a, 2011)	AS, M
ASAS variables in Kepler	Pigulski et al. (2009)	AS, M
BCEP stars	Stankov & Handler (2005)	AS, M
Catalina cataclysmic variables	Drake et al. (2014a)	AS
Catalina periodic variables	Drake et al. (2014b)	AS, M, R
Catalina RRab stars (paper I)	Drake et al. (2013a)	AS, M, R
Catalina RRab stars (paper II)	Drake et al. (2013b)	AS, M, R
Catalina RRab stars (SSS)	Torrealba et al. (2015)	AS, M, R
CoRoT Rotational Modulation	De Medeiros et al. (2013)	AS, M, C
DSCT and GDOR stars	Bradley et al. (2015); Sarro et al. (2013) Uytterhoeven et al. (2011)	AS, M
EROS-II Beat Cepheids	Marquette et al. (2009)	AS
Gaia DR1 (RR Lyrae & Cepheids)	Clementini et al. (2016)	AS, M
GDOR stars	Debosscher et al. (2007) Kahraman Aliçavuş et al. (2016)	AS, M, C
Hipparcos periodic variables and constants	ESA (1997); van Leeuwen (2007b)	AS, M, C, R
ICRF2 Quasars	Ma et al. (2009)	AS, M
Kepler Flares	Shibayama et al. (2013) Walkowicz et al. (2011); Wu et al. (2015)	AS, M
Kepler Rotational Modulation	Reinhold & Gizon (2015)	AS, M, C
LINEAR periodic variables	Palaversa et al. (2013)	AS, M, SD
M37 Flares	Chang et al. (2015b)	AS
NSVS Red variables	Woźniak et al. (2004)	AS, M, R
NSVS RRab stars	Kinemuchi et al. (2006)	AS, M, R
OGLE-IV Blue large amplitude pulsators	Pietrukowicz et al. (2017)	AS, M
OGLE-IV Cataclysmic variables	Mróz et al. (2015)	AS
OGLE-IV Cepheids and RR Lyrae (LMC, SMC)	Soszyński et al. (2015b,c, 2016b)	AS, M, C, R
OGLE-IV Eclipsing binaries (bulge)	Soszyński et al. (2016a)	AS, M, C, R
OGLE-IV Eclipsing binaries (LMC, SMC)	Pawlak et al. (2016)	AS, M, C, R
OGLE-IV GSEP constant candidates	Soszyński et al. (2012)[a]	AS, M, C, SD
OGLE-IV GSEP variables	Soszyński et al. (2012)	AS, M
OGLE-IV RR Lyrae stars (bulge)	Soszyński et al. (2014)	AS, M, C, R
OGLE-IV Short period binaries	Soszyński et al. (2015a)	AS, M, C, R
Pan-STARRS1 RR Lyrae stars	Sesar et al. (2017)	AS, M
Rotational Modulation	Stauffer et al. (2007) Collier Cameron et al. (2009) Hartman et al. (2009); Meibom et al. (2009) Messina et al. (2010b); Delorme et al. (2011) Meibom et al. (2011a,b); Moraux et al. (2013) Kovács et al. (2014); Meibom et al. (2015) Chang et al. (2015a); Barnes et al. (2015) Douglas et al. (2016); Covey et al. (2016)	AS
RR Lyrae in ω Centauri globular cluster	Braga et al. (2016)	AS, M

Continued on next page

Table 7.1. (Continued)

Description	Reference	Crossmatch Metrics
RR Lyrae in M3	Benkő et al. (2006)	AS, M
RR Lyrae in M15	Corwin et al. (2008)	AS, M
RR Lyrae in ultra-faint dwarf spheroidals	Dall'Ora et al. (2006); Siegel (2006) Kuehn et al. (2008); Greco et al. (2008) Watkins et al. (2009); Moretti et al. (2009) Musella et al. (2009, 2012) Clementini et al. (2012); Dall'Ora et al. (2012) Boettcher et al. (2013); Garofalo et al. (2013) Sesar et al. (2014); Vivas et al. (2016)	AS, M
SDSS DSCT and RR Lyrae stars	Süveges et al. (2012)	AS, M, C
SDSS-PS1-Catalina RR Lyrae stars	Abbas et al. (2014)	AS, M
SDSS Standard stars	Ivezić et al. (2007)	AS, M, C
Solar-like activity in the Pleiades	Hartman et al. (2010)	AS, M
SPB and BCEP stars	Selected by Peter De Cat[b]	AS, M
SPB stars	Niemczura (2003)	AS, M

[a] Selection of the least varying sources at `ftp://ftp.astrouw.edu.pl/ogle/ogle4/GSEP/maps/`.

[b] Selection of P. De Cat available at `http://www.ster.kuleuven.ac.be/~peter/Bstars/`.

7.3.3.3 Classification attributes

About one hundred fifty attributes were computed to characterise sources with photometric (and some astrometric) time series features. Each classifier (described in Section 7.3.3.4) was tested with a varying number of attributes (e.g., Guyon & Elisseeff 2003) and a subset of 40 attributes represented the union of attributes used by all classifiers. The employed classification attributes are defined below, with units quoted in brackets after the attribute name (unless the attribute is dimensionless).

1. ABBE: The Abbe value (von Neumann 1941, 1942) computed from the magnitudes of FoV transits in the G band.

2. BP_MINUS_RP_COLOUR (mag): The possibly reddened colour index from the median magnitudes in the $G_{\rm BP}$ and $G_{\rm RP}$ bands.

3. BP_MINUS_G_COLOUR (mag): The possibly reddened colour index from the median magnitudes in the $G_{\rm BP}$ and G bands.

4. DENOISED_UNBIASED_UNWEIGHTED_KURTOSIS_MOMENT (mag^4): The sample-size unbiased and unweighted kurtosis central moment of FoV transit magnitudes in the G band, denoised assuming Gaussian uncertainties (Rimoldini 2014).

5. DENOISED_UNBIASED_UNWEIGHTED_VARIANCE (mag^2): The sample-size unbiased and unweighted variance of FoV transit magnitudes in the G band, denoised assuming Gaussian uncertainties (Rimoldini 2014).

6. DURATION (d): The duration of the time series from the first to the last FoV transit observation in the G band.

7. G_MINUS_RP_COLOUR (mag): The possibly reddened colour index from the median magnitudes in the G and G_{RP} bands.

8. G_VS_TIME_IQR_ABS_SLOPE (mag d^{-1}): The unweighted interquartile range of the absolute values of magnitude changes per unit time between successive FoV transits in the G band.

9. G_VS_TIME_MAX_SLOPE (mag d^{-1}): The unweighted 95th percentile of magnitude changes per unit time between successive FoV transits in the G band.

10. G_VS_TIME_MEDIAN_ABS_SLOPE (mag d^{-1}): The unweighted median of the absolute values of magnitude changes per unit time between successive FoV transits in the G band.

11. IQR_BP (mag): The unweighted interquartile magnitude range of FoV transits in the G_{BP} band.

12. IQR_RP (mag): The unweighted interquartile magnitude range of FoV transits in the G_{RP} band.

13. LOG_QSO_VAR: The decadic logarithm of the reduced chi-square of FoV transit magnitudes in the G band with respect to a parameterised quasar variance model, represented by $\log_{10}(\chi^2_{\text{QSO}}/\nu)$ in Butler & Bloom (2011); see Rimoldini et al. (in preparation) for details on the parameter values for the Gaia data.

14. LOG_NONQSO_VAR: The decadic logarithm of the reduced chi-square of FoV transit magnitudes in the G band *not* to follow a parameterised quasar variance model, represented by $\log_{10}(\chi^2_{\text{False}}/\nu)$ in Butler & Bloom (2011); see Rimoldini et al. (in preparation) for details on the parameter values for the Gaia data.

15. MAD_G (mag): The unweighted median absolute deviation from the median magnitude of FoV transits in the G band.

16. MAX_ABS_SLOPE_HALFDAY (mag d^{-1}): The maximum value of the magnitude ranges of FoV transits in the G band within sliding windows of half a day, divided by the time span of the G-band observations within such sliding windows.

17. MEAN_G (mag): The unweighted arithmetic mean magnitude of FoV transits in the G band.

18. MEAN_BP (mag): The unweighted arithmetic mean magnitude of FoV transits in the G_{BP} band.

19. MEAN_RP (mag): The unweighted arithmetic mean magnitude of FoV transits in the G_{RP} band.

20. MEDIAN_ABS_SLOPE_HALFDAY (mag d^{-1}): The unweighted median of the magnitude ranges of FoV transits in the G band within sliding windows of half a day, divided by the time span of the G-band observations within such sliding windows.

21. MEDIAN_ABS_SLOPE_ONEDAY (mag d^{-1}): The unweighted median of the magnitude ranges of FoV transits in the G band within sliding windows of one day, divided by the time span of the G-band observations within such sliding windows.

22. MEDIAN_G (mag): The unweighted median magnitude of FoV transits in the G band.

23. MEDIAN_BP (mag): The unweighted median magnitude of FoV transits in the G_{BP} band.

24. MEDIAN_RANGE_HALFDAY_TO_ALL: The unweighted median of the magnitude ranges of FoV transits in the G band within sliding windows of half a day, divided by the G-band magnitude range of the full time series.

25. MEDIAN_RP (mag): The unweighted median magnitude of FoV transits in the G_{RP} band.

26. NONQSO_PROB: A quantity distributed according to the null-hypothesis distribution of χ^2_{QSO}, given the data, for non-quasar objects, computed from a parameterised quasar variance model with magnitudes of FoV transits in the G band, related to $P(\chi^2_{\mathrm{QSO}}|x$, not quasar) in Butler & Bloom (2011); see Rimoldini et al. (in preparation) for details on the parameter values for the Gaia data.

27. NORMALISED_CHI_SQUARE_EXCESS: The difference between the chi-square of FoV transit magnitudes in the G band and the mean of the chi-square distribution expected for constant objects (i.e., the number of degrees of freedom), normalised by the standard deviation of the chi-square distribution of constant objects (i.e., the square root of twice the number of degrees of freedom).

28. OUTLIER_MEDIAN_G: The absolute difference between the most outlying FoV transit magnitude with respect to the median magnitude in the G band, normalised by the uncertainty of the most outlying measurement.

29. PARALLAX (mas): The parallax value of the source derived from a preliminary astrometric solution (Section 7.2.2.1).

30. PROPER_MOTION (mas yr^{-1}): The proper motion of the source projected in the sky derived from a preliminary astrometric solution (Section 7.2.2.1).

31. PROPER_MOTION_ERROR_TO_VALUE_RATIO: The ratio between the estimated projected proper motion uncertainty and the projected proper motion value of the source, derived from a preliminary astrometric solution (Section 7.2.2.1).

32. RANGE_G (mag): The magnitude range of FoV transits in the G band.

33. REDUCED_CHI2_G: The reduced chi-square of FoV transit magnitudes in the G band.

34. SIGNAL_TO_NOISE_STDEV_OVER_RMSERR_G: The ratio between the sample-size biased unweighted standard deviation of FoV transit magnitudes in the G band and the root-mean-square of their uncertainties.

35. SKEWNESS_G: The sample-size unbiased and unweighted skewness central moment of FoV transit magnitudes in the G band, normalised by the third power of the unbiased unweighted standard deviation of the same time-series measurements.

36. SKEWNESS_PERCENTILE_5: A robust measure of the skewness of the magnitude distribution of FoV transits in the G band, computed as $(P_{95} + P_5 - 2\,P_{50})/(P_{95} - P_5)$ where P_n is the nth unweighted percentile.

37. STETSON_G: The single-band Stetson variability index (Stetson 1996) computed from the magnitudes of FoV transits in the G band, pairing observations within 0.1 days.

38. STETSON_G_BP: The double-band Stetson variability index (Stetson 1996) computed from the magnitudes of FoV transits in the G and G_{BP} bands, pairing observations in different bands within 0.001 days.

39. TRIMMED_RANGE_G (mag): The magnitude range between the 5th and 95th unweighted percentiles of FoV transits in the G band.

40. TRIMMED_RANGE_RP (mag): The magnitude range between the 5th and 95th unweighted percentiles of FoV transits in the G_{RP} band.

7.3.3.4 Classification models

A hierarchical structure of Random Forest (Breiman 2001) classifiers identified objects in progressively more detailed (groups of) classes. For Gaia DR2, we focused on high-amplitude variable stars, so objects with negligible or low amplitude variations were first separated from the high amplitude ones, which were then split into the types and subtypes of interest by subsequent classifiers.

Every Random Forest classifier was configured with unlimited depths and with a minimum number of instances per class at the leafs set to one. Other configuration parameters (number of trees `nTree` and number of tested attributes `mTry` to best split the data at a given node of a tree), the training-set classes to identify (specified by the labels defined in Section 7.3.3.1), and the selected attributes (described in Section 7.3.3.3) are listed below for each classifier. Aggregations of types are denoted by connecting single type labels with an underscore (unless indicated otherwise in brackets).

1. Random Forest classifier configured with `nTree`=400 and `mTry`=10.
   (a) Training set:
      i. 14 684 CONSTANT;
      ii. 3885 LOW_AMPLITUDE_VARIABLE (ACV, ACYG, BCEP, low-amplitude DSCT_GDOR, ELL, FLARES, GCAS, GDOR, OSARG, ROT, SOLAR_LIKE, SPB, SXARI);
      iii. 14 999 OTHER_VARIABLE (ACEP, ARRD, BLAP, CEP, CV, DSCT, ECL, MIRA, QSO, RRAB, RRC, RRD, RS, SR, SXPHE, T2CEP).
   (b) Attributes: BP_MINUS_G_COLOUR, BP_MINUS_RP_COLOUR, DENOISED_UNBIASED_UNWEIGHTED_VARIANCE, DURATION, G_MINUS_RP_COLOUR, G_VS_TIME_MEDIAN_ABS_SLOPE, IQR_BP, IQR_RP, LOG_NONQSO_VAR, LOG_QSO_VAR, MAD_G, MEDIAN_ABS_SLOPE_ONEDAY, MEDIAN_BP, MEDIAN_G, MEDIAN_RP, NONQSO_PROB, NORMALISED_CHI_SQUARE_EXCESS, OUTLIER_MEDIAN_G, RANGE_G, REDUCED_CHI2_G, SIGNAL_TO_NOISE_STDEV_OVER_RMSERR_G, SKEWNESS_PERCENTILE_5, STETSON_G, STETSON_G_BP, and TRIMMED_RANGE_RP.
2. Random Forest classifier configured with `nTree`=321 and `mTry`=4 (not relevant to the classification results published in Gaia DR2, but still described for details on the objects of low-amplitude types employed).
   (a) Training set:
      i. 363 ACV_ACYG_BCEP_GCAS_SPB_SXARI (combination of poorly represented low-amplitude objects characterized by multiperiodic, pulsating, rotating, or irregular light variations);
      ii. 866 DSCT_GDOR_LOW_AMPLITUDE (DSCT, GDOR, and DSCT-GDOR hybrids with low amplitude variations);
      iii. 397 ELL;
      iv. 996 OSARG;
      v. 1247 SOLARLIKE_FLARES_ROT.
   (b) Attributes: BP_MINUS_RP_COLOUR, DURATION, G_MINUS_RP_COLOUR, IQR_RP, LOG_QSO_VAR, MEAN_BP, MEAN_G, PARALLAX, PROPER_MOTION.
3. Random Forest classifier configured with `nTree`=336 and `mTry`=3.
   (a) Training set: 10 BLAP, 711 CEP_ACEP_T2CEP, 518 CV, 1326 DSCT_SXPHE, 3861 ECL, 1945 MIRA_SR, 1996 QSO, 4108 RRAB_RRC_RRD_ARRD, and 500 RS.

   (b) Attributes: ABBE, BP_MINUS_RP_COLOUR, DENOISED_UNBIASED_UNWEIGHTED_VARIANCE, G_MINUS_RP_COLOUR, G_VS_TIME_MAX_SLOPE, MEAN_G, MEAN_RP, MEDIAN_ABS_SLOPE_ONEDAY, MEDIAN_RANGE_HALFDAY_TO_ALL, NORMALISED_CHI_SQUARE_EXCESS, PARALLAX, PROPER_MOTION, PROPER_MOTION_ERROR_TO_VALUE_RATIO, RANGE_G, and SKEWNESS_G.

4. Random Forest classifier configured with `nTree`=202 and `mTry`=3.
   (a) Training set: 2922 RRAB, 969 RRC, 197 RRD, and 20 ARRD.
   (b) Attributes: BP_MINUS_RP_COLOUR, DENOISED_UNBIASED_UNWEIGHTED_KURTOSIS_MOMENT, G_VS_TIME_IQR_ABS_SLOPE, G_VS_TIME_MAX_SLOPE, NORMALISED_CHI_SQUARE_EXCESS, STETSON_G, and TRIMMED_RANGE_G.

5. Random Forest classifier configured with `nTree`=135 and `mTry`=3.
   (a) Training set: 99 ACEP, 455 CEP, and 157 T2CEP.
   (b) Attributes: BP_MINUS_RP_COLOUR, DURATION, LOG_NONQSO_VAR, LOG_QSO_VAR, MAX_ABS_SLOPE_HALFDAY, MEAN_G, MEDIAN_ABS_SLOPE_HALFDAY, and MEDIAN_RP.

7.3.3.5 Semi-supervised classification

Semi-supervised classification was applied to constant objects, RR Lyrae stars, and long period variables, in order to improve their representation in the training set as follows.

1. High-confidence classifications of such classes were selected as candidate training sources.
2. Candidate training objects were filtered by the statistics mentioned in item 3 of Section 7.3.3, except for the literature period and the Abbe value computed on phase-sorted magnitudes (not available for results classified without period computation).
3. Filtered candidate training objects were selected to cover regions in the sky and/or magnitude intervals that lacked proper representation in the training set.

7.3.3.6 Contamination cleaning

The contamination of preliminary classification results was reduced with the help of dedicated classifiers applied to RR Lyrae stars, Cepheids, and SX Phoenicis/δ Scuti stars, separately for each type, as follows.

1. Samples of true positives and false positives (according to crossmatched objects) were selected from the candidates of the previous classification stage.
2. Classification attributes were generated and selected.
3. A binary classifier of true positives versus false positives (in similar amounts) was trained and optimized.
4. The preliminary classification candidates (above some minimal level of classification probability depending on the type) were processed by the binary classifier (item 3) and objects classified as true positives with a minimum probability of 50 per cent were retained.

7.3.3.7 Classification score

The results of the contamination-cleaning classifiers are associated with classification scores which express the confidence of the classifier given the training set, thus such scores should not be interpreted as true probabilities. The scores of Gaia DR2 classification results are obtained by linearly mapping the internal classifier probabilities to values within a range from zero to one (from the weakest to the strongest candidate), for each variability type.

7.3.4 Quality assessment and validation

Author(s): László Molnár, Emese Plachy, Áron Juhász, Lorenzo Rimoldini

The verification of results and their validation are performed by employing:

1. SOS of Cepheids and RR Lyrae stars applied to sources with at least 12 FoV measurements in the *G* band (Section 7.4).
2. SOS of long period variables applied to sources with at least 12 FoV measurements in the *G* band (Section 7.7).
3. The crossmatch of Cepheids and RR Lyrae star candidates with objects in the Kepler/K2 fields (Section 7.3.4.1, Section 7.3.4.2).
4. Crossmatched objects not included in the training set (Rimoldini et al., in preparation).

7.3.4.1 Verification

The verification of RR Lyrae and Cepheid candidates with Kepler/K2 fields is summarised here (for more details, see Molnár et al. 2018). We analysed the Gaia DR2 candidates in circular areas with a 8.5 degree radius centred on the fields of view of the original Kepler mission and the K2 mission observing Campaigns up to Field 13 (Howell et al. 2014, `https://keplerscience.arc.nasa.gov/k2-fields.html`). The prime Kepler mission observed a single field of view towards Lyra-Cygnus for four years. The K2 mission is ordered into campaigns along the Ecliptic; one campaign lasts for 60–80 days and then the spacecraft is reoriented. The Gaia DR2 candidates in these fields were crossmatched with the Kepler Input Catalog (KIC), the K2 Ecliptic Plane Input Catalog (EPIC), and the list of K2 targets selected for observation (Brown et al. 2011; Huber et al. 2016). The resolution of Kepler ($4''\text{pixel}^{-1}$) is much poorer than the one of Gaia, leading to some ambiguity in crossmatching the Gaia sources with the K2 targets. Nevertheless, RR Lyrae and Cepheid variations can be recovered even if the target is blended with another star within the photometric aperture of Kepler. We found no cases where two or more RR Lyrae or Cepheid candidates from Gaia would fall into the same aperture. We did not crossmatch sources from Campaign 9 that targeted the Galactic Bulge as the high source number density and the limited resolution of Kepler lead to strong confusion and data from OGLE was deemed superior to that of Kepler in this region.

We also made a list of known or suspected RR Lyrae stars that were proposed for observation and confirmed by Kepler, so that the completeness of Gaia DR2 candidates could be assessed from the rate of missed identifications.

7.3.4.2 Validation

The validation of RR Lyrae and Cepheid candidates with Kepler/K2 fields is summarised here (for more details, see Molnár et al. 2018). For the Lyra-Cygnus field, we visually inspected the Simple Aperture Photometry (SAP) and Pre-search Data Conditioning (PDC) SAP light curves of each target that was selected for observation in at least one observing quarter (one three-month segment of the original mission). We identified 48 RR Lyrae stars from the Gaia DR2 candidates, four of which were found not to be of the RR Lyrae type. Twelve other known RR Lyrae stars were not among the Gaia DR2 candidates, suggesting a sample completeness of about 78 per cent.

The original Kepler mission also acquired 52 Full-Frame Images (FFI). We extracted light curves for the objects not targeted by the mission from these images using the `f3` code (Montet et al. 2017). We compared the light curves folded with the fundamental periods derived from the Gaia data as well as from the FFI data visually. Out of the 267 additional stars from the Gaia DR2 RR Lyrae candidates, we were able to classify 185 as RRAB or RRC variables (the other ones were either not RR Lyrae stars or associated with unreliable photometry). The combination of this set and the 48 stars described in the previous paragraph suggests a purity of the sample of at least 75 per cent.

In the K2 fields, we checked the light curves available for the targeted stars. These include the SAP/PDCSAP data sets provided by the mission as well as the available community-created light curves for selected campaigns. Out of the 1395 RR Lyrae candidates with counterparts in the K2 fields, 1371 were classified as RRAB or RRC in Gaia DR2, while 24 candidates turned out not to be RR Lyrae variables. The confirmed candidates are part of a larger set of 1816 known RR Lyrae stars in the K2 fields, suggesting a completeness rate around 75 per cent, in agreement with the one estimated from the original Kepler field, and a purity of 98 per cent (with a worst-case lower limit of 51 per cent) for the Ecliptic fields outside the Bulge. The interpretation of the purity value, however, is complicated by the biases in the selection of various targets for the K2 mission. About the classification of RR Lyrae stars into subclasses, 31 of the 1371 confirmed candidates were associated with the incorrect subtype, with misclassification rates of 1, 9, and 50 per cent for RRAB, RRC, and RRD types, respectively.

Cepheids were very sparse in the original Kepler fields. Among the Gaia DR2 Cepheid candidates, we found 38 Cepheid-type stars (ACEP, CEP, T2CEP) in the K2 fields and we were able to confirm 22, and assume 3 more of them (about 66 per cent). In the original field, we confirm the detection of the classical Cepheid V1154 Cyg and the T2CEP HP Lyr, while the semi regular star V677 Lyr was misclassified as T2CEP. However, the low number of targets prevented us from drawing more detailed conclusions.

7.4 Cepheid and RR Lyrae stars

Author(s): Gisella Clementini, Vincenzo Ripepi, Roberto Molinaro

We validate and refine the detection and classification of all-sky candidate RR Lyrae and Cepheid variables provided by the general variable star analysis pipeline from about 22 months of Gaia G, G_{BP}, G_{RP} photometry.

7.4.1 Introduction

We produce a list of confirmed all-sky RR Lyrae and Cepheid stars cleaned from contaminating objects and other types of variables falling into the same period domain. For all stars we provide a number of attributes (with related errors) to be published in the second Gaia Data Release among which, specifically: period, peak-to-peak amplitudes, mean magnitudes and epoch of maximum light in G, G_{BP}, G_{RP} bands (when G_{BP} and G_{RP} are available) as well as Fourier parameters from the G-band light curves. Additionally, for RR Lyrae stars for which the ϕ_{31} Fourier parameter is available we provide a metallicity ([Fe/H]) estimate and, for RRab types we also publish an estimate of the interstellar absorption in the G-band. Also, for Cepheid stars with period shorter than about 6 days we provide an estimate of metallicity ([Fe/H]).

7.4.2 Properties of the input data

Selection criteria:

- sources classified as candidate Cepheid and RR Lyrae variables from the Classifiers;
- a minimum number of 12 G-FoV transits, before applying an outlier removal procedure specifically tailored to Cepheids and RR Lyrae stars to discard obvious wrong epoch data;
- a peak-to-peak amplitude > 0.1 mag in the G-band;
- periods in the range of 0.2-1.0 days for the RR Lyrae variables.

7.4.3 Calibration models

The SOS Cep&RRL processing uses tools such as: period-amplitude (*PA*) and period-luminosity (*PL*) relations in the G-band, as described in the documentation for the processing of RR Lyrae and Cepheid stars released in Gaia Data Release 1, (Clementini et al. 2016). For the Gaia Data Release 2 data processing (Clementini et al. 2018) we also use tools based on the $G_{\rm BP}$ and $G_{\rm RP}$ photometry, such as the period-luminosity in the RP-band and the period-Wesenheit (*PW*) relation in G, $G_{\rm RP}$. Furthermore, we implemented i) use of parallaxes according to the Astrometric Based Luminosity formulation, (i.e. working directly in parallax space; see, e.g., (Gaia Collaboration et al. 2017) and references therein) and applying different *PL*, *PW* relations depending on source position on sky (whether in the Large Magellanic Cloud, in the Small Magellanic Cloud or outside them); ii) calculation of metallicity ([Fe/H]) for the RR Lyrae stars and for δ Cepheid variables with period shorter than about 6 days from the Fourier parameters and, iii) calculation of interstellar absorption in the G-band for the RRab stars from a relation based on G-band peak-to-peak amplitude and period.

7.4.4 Processing steps

The processing includes the following steps common to both RR Lyrae and Cepheid stars (see Fig. 1 in Clementini et al. 2018):

1. Derivation of period and harmonics (amplitudes and phases) by non-linear Fourier analysis,
2. Measurement of light curve parameters (mean magnitudes, amplitudes, epochs of maximum light, etc.),
3. Consistency check of the periods derived from the 3 bands (G, $G_{\rm BP}$, $G_{\rm RP}$),
4. Search for secondary periodicities.

The following additional steps are then applied to sources confirmed as RR Lyrae stars (see Fig. 2 in Clementini et al. 2018):

1. RR Lyrae Double-mode search,
2. Non-linear double-mode modelling,
3. Amplitude ratios,
4. Mode identification,
5. RR Lyrae Classification and validation,
6. Stellar parameters derivation: metallicity,
7. Stellar parameters derivation: absorption in the G-band for RRab stars.

and the following additional steps applied to sources confirmed as Cepheid variables (see Fig. 3 in Clementini et al. 2018):

1. Cepheid Multimode search,
2. Cepheid Type identification,
3. Type II Cepheid Subclassification,
4. δ Cepheid Mode identification,
5. ACEP Mode identification,
6. Cepheid Classification and validation,
7. Stellar parameters derivation: metallicity for δ Cepheid variables with period shorter than about 6 days.

7.4.5 Quality assessment and validation

Quality assessment and validation of the results are performed by crossmatching with catalogues of known RR Lyrae and Cepheid stars from other surveys (OGLE, Catalina, Linear, catalogues of variable stars in globular clusters and dwarf spheroidal galaxies).

7.4.5.1 Verification

Verification is done by crossmatching with catalogues of known RR Lyrae stars and Cepheid stars and comparing source attributes computed by SOS Cep&RRL with those published by OGLE in particular.

7.4.5.2 Validation

Taking advantage of the comparison between properties of RR Lyrae and Cepheid variable candidates derived by the SOS pipeline and those of known objects in the literature, we operated a selection of the candidates that led to the final published catalogue. More in detail, from the original 639 828 RR Lyrae and 72 455 Cepheid candidates, 140 784 and 9 575 objects passed this validation step.

7.5 Solar-Like variables

Author(s): Elisa Distefano, Alessandro Lanzafame, Leanne Guy

The Gaia DR2 provides a list of 147 535 solar-like variable star candidates obtained by the analysis of about 22 months of Gaia photometry. For each of these candidates, the Release supplies different parameters like the stellar rotation period, the amplitude of variability and a list of photometric outliers that could be possible flare events candidates. This section describes the methods and algorithms used for obtaining this list, as well as the verification and validation performed on the obtained sample. All the details on the solar-like analysis methods and results are extensively described in Lanzafame et al. (2018).

7.5.1 Introduction

Solar-like stars are characterised by variability phenomena due to a solar-like magnetic activity that occurs in all the main sequence stars with a spectral type later the F5. The most important variability phenomena exhibited by solar-like stars are the rotational modulation of the stellar flux and the occurrence of flare events. The rotational modulation of the stellar flux is due to the dark spots and bright faculae unevenly distributed over the stellar disk. The stellar rotation modulate the visibility of such surface inhomogeneities and consequently the flux coming from the star. Hence, the period of light curves, for these stars, is coincident with the stellar rotation period. Flare events are sporadic outbursts due to reconnection of magnetic fields with subsequent plasma heating, particles acceleration and emission in several bands, particularly UV and X-rays. A description of solar-like variability phenomena can be found in Distefano et al. (2012) and references therein. The detection and characterisation of solar-like stars is performed by means of the SVD-Solar-Like and the SOS-Rotational-Modulation packages. The first package has the tasks to perform a first selection of solar-like candidates and to identify photometric outliers. The SOS package has the task to detect and characterise rotational-modulation variability on the solar-like candidates.

7.5.2 Properties of the input data

The input sources processed by the SVD-Solar-Like-SOS-Rotational-Modulation pipeline were selected from the catalogue of sources having at least 20 observations in the G band. From this catalogue, we selected the main sequence stars with a spectral type later then F5. This selection has been done by looking at the position of the stars in the $M_{\rm G}$ vs. $G_{\rm BP} - G_{\rm RP}$ diagram, and therfore the parallax of the star is needed. In order to limit the errors on the $M_{\rm G}$ computation, we selected only stars with a relative error in parallax less than 20%. A second criterion used to select the input stars is based on the time sampling of G observations. As described in Distefano et al. (2012), the analysis of solar-like variables requires that the G and the $G_{\rm BP} - G_{\rm RP}$ time-series can be segmented in small sub-series. The SVD-Solar-Like-SOS-Rotational-Modulation pipeline was employed to process only sources whose time-series can be split in at least two segments with a number of observations $N_{\rm G} \geq 12$. See Lanzafame et al. (2018) for more details on the selection procedure.

7.5.3 Processing steps

The main processing steps of the SVD-Solar-Like-SOS-Rotational-Modulation pipeline are the following:

- selection of the input sources

- segmentation of the photometric time-series
- estimate of the linear correlation degree between gmag and $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ observations in each time-series segment
- search for outliers in time-series segments
- search for a periodic signal in each time-series segment
- modelling of the G time-series
- estimate of a magnetic activity index in each time-series segment
- estimate of the stellar rotation period.

The first four tasks are performed by the SVD package whereas the others are performed by the SOS package. The selection of the input candidates is performed according the criteria outlines in Section 7.5.2. The segmentation of photometric time-serie is required because the typical lifetime of spots and faculae is of the order of several months (see e.g. Lanza et al. 2003; Messina et al. 2003, and reference therein). A detailed description of the adopted segmentation algorithm can be found in Lanzafame et al. (2018). A well defined linear correlation between colour and magnitude measurements is expected for variability due rotational modulation in solar-like stars (see e.g. Messina et al. 2006). The SVD-Solar-Like package estimates the Pearson Correlation Coefficient r between G and $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ observations in each time-series segment. This index can be regarded as an indicator of variability due to rotational modulation. The closer r is to ± 1 the higher the probability that rotational modulation is occurring. The SVD-Solar-Like package performs also a robust linear regression between G and $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ observations. The robust regression procedure permits also to identify possible outliers i.e. points whose location in the G vs. $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ scatter plot, is significantly distant from the straight-line best-fitting the data. The identification of photometric outliers is also performed by searching for the observations satisfying the condition

$$(G_{\mathrm{BP}} - G_{\mathrm{RP}})_i < \overline{G_{\mathrm{BP}} - G_{\mathrm{RP}}} - 5\sigma_{G_{\mathrm{BP}}-G_{\mathrm{RP}}} \tag{7.2}$$

The points identified as outliers can be regarded as candidate flare-events. The SOS-Rotational-Modulation package performs a period search in each time-series segment by computing the generalised Lomb-Scargle periodogram as implemented by Zechmeister & Kürster (2009). The period with the highest power in the periodogram is selected by the pipeline and a False Alarm Probability (FAP) is associated with it. The formulation used to compute the FAP is that prescribed by Baluev (2008). A period is flagged as valid if the associated FAP is less then 0.05. If a significant period is detected in a given segment, the pipeline performs also a data modelling and fits the time-series segment to the function:

$$G(t) = A + B \sin\left(\frac{2\pi t}{P}\right) + C \cos\left(\frac{2\pi t}{P}\right) \tag{7.3}$$

where t is the observation time referred to the reference epoch t_{start} that is the time at which starts the segment and P is the period detected in the segment. The SOS-RotationalModulation package estimates the rotation period of a solar-like candidate by analysing the distribution of the periods recovered in the different time-series segments. The mode of the distribution is taken as best estimate of the stellar rotation period. The amplitude of rotational modulation can be regarded as an index of the stellar magnetic activity and is widely used to study solar-like activity cycles (see e.g. Rodonò et al. 2000; Ferreira Lopes et al. 2015; Lehtinen et al. 2016). For a given we computed an estimate of this Activity Index (AI) by means of the equation:

$$AI = G_{95th} - G_{5th} \tag{7.4}$$

where G_{95th} and G_{5th} are the 95-th and 5-th percentiles of the G magnitudes measured in the segment. Note that an alternative estimate of the amplitude associate with rotational modulation can be inferred from the fit coefficients of Equation 7.3 through the relationship:

$$A_{fit} = 2\sqrt{B^2 + C^2} \tag{7.5}$$

A more detailed description of the reduction pipeline can be found in Lanzafame et al. (2018).

7.5.4 Quality assessment and validation

Quality assessment and validation of the results were performed by means of three different methods:

- Cross-match between Gaia DR2 results and solar-like variables with a known rotation period;
- statistical analysis of the stellar parameters inferred by the pipeline;
- visual inspection of folded light-curves of a few hundred selected examples.

7.5.4.1 Verification

The optimal method to verify the SVD-Solar-Like-SOS-Rotational modulation pipeline should be the comparison of Gaia DR2 results with surveys dedicated to solar-like variables like Kepler of Corot. Unfortunately the number of observations and the Gaia sampling in the sky fields covered by Kepler Corot do not allow the selection of input sources in those sky areas. A comparison between Gaia results and these surveys will be possible only for the DR3 release. In spite of everything, there are several studies on rotational modulation in open-clusters stars that can be used to verify Gaia DR2 data. The Pleiades field, for instance, has been well studied and the rotation period has been estimated for a few hundred of stars belonging to this field (see e.g. Hartman et al. 2010). See Lanzafame et al. (2018) for details on the comparison between Gaia DR2 results and those of Hartman et al. (2010).

7.5.4.2 Validation

The SVD-Solar-Like-SOS-Rotational-Modulation pipeline was able to detect 723 315 solar-like candidates. The statistical analysis of the parameters inferred by the pipeline and the visual inspection of hundreds of folded light curves showed that a certain fraction of the selected candidates was doubtful. In many stars there was a strong discrepancy between the parameters AI and A_{fit} (defined in Equation 7.4). In certain segments, where a significant period was detected, the visual inspection of the folded light curve revealed that the phase coverage of the data is really poor making the detected period doubtful. The folded light curves of some stars have the typical shape of other variable objects like Cepheids stars or Eclipsing Binaries. In order to deal with these issues, we applied four different filters to the sample of solar-like candidates. The first filter takes into account the ratio R between AI and A_{fit}. We rejected all the stars satisfying one of the conditions:

$$R \geq 1.4 \tag{7.6}$$

$$R \leq 0.5 \tag{7.7}$$

where the values 0.5 and 1.4 correspond the 5-th and 95-th percentile of the R distribution. The second filter is based on the Phase-Coverage (PC) and the Maximum-Phase-Gap (MPG) parameters. The PC parameter measures how uniformly the observations are distributed over phase when the data are folded with the period detected by the period-search module. The observations collected in a given segment are folded according the period detected in that segment and their phases are binned in 10 equally spaced intervals in the range [0,1]. The number of bins that contains at least one observation is divided by the total number of bins, to obtain a phase coverage number in [0,1]. If every bin in the phase-coverage histogram contains at least one observation, the phase coverage will be 1, indicating that for the given model, the data are quasi-uniformly distributed in phase. At the other extreme, if all the observations fall into the same bin, a value tending to 0 will be obtained. The MPG parameter is defined as the maximum gap in phase between the data i.e. as

$$MPG = \max(\Delta\phi_{i,j}) \tag{7.8}$$

where

$$\Delta\phi_{i,j} = \phi_i - \phi_j \tag{7.9}$$

where ϕ_i and ϕ_j are the phases of the i-th and j-th observations when folded according to the detected period.

We applied a filter that flags a candidate as valid only if the requirements:

$$PC >= 0.4 \tag{7.10}$$

$$MPG <= 0.3 \tag{7.11}$$

are satisfied in one segment at least. Finally, visual inspection of the folded light curves revealed that some of the detected variables were not solar-like stars but Cepheids. This can happen if the star has an over-estimated parallax and, consequently, un under-estimated luminosity. In such a case the location of the star in the magnitude-colour diagram can fall in the region used for the selection of the input sources. In order to avoid these problems, we rejected all the stars classified as Cepheids from the CU7-Classification package. By applying all these filters, the final number of solar-like candidates reported in Gaia DR2 is 147 535.

7.6 Short time scale variables

Author(s): Maroussia Rolens, Laurent Eyer

The Gaia DR2 provides a first list of suspected periodic short-timescale candidates with periods below 0.5–1 day from about 22 months of Gaia photometry. This section describes the methods and algorithms used for obtaining this list, as well as the verification and validation performed on the obtained sample. All the details on the short timescale analysis methods and results are extensively described in Roelens et al. (2018).

7.6.1 Introduction

The short-timescale SOS work package aims to produce a list of suspected periodic short-timescale candidates with periods between a few tens of minutes to one day. This candidate list results from the analysis of Gaia time series (in G CCD, G FoV, G_{BP} and G_{RP}), for faint sources, with sufficient number of transits in the G band, and for which per-CCD time series showed a significant degree of variability at the transit level and for the majority of the transits of the source.

7.6.2 Properties of the input data

The input sample comprised only those sources for which per-CCD data was available in the Cycle 2 photometry provided by CU5 and which satisfied ≥20 G-FoV transits, to ensure a reliable variogram analysis. At this time of the Gaia processing, CU7 receives per-CCD time series only for sources with more than half of their FoV transits identified as ‘noisy’ according to their p-value from the 9 CCD measurements of the considered transit (the limit defining a noisy transit being p-value below 0.01).

The analysis focused on faint sources with a mean G magnitude in the range 16.5–20 mag as it is in this range that a relevant and validated detection criterion can be obtained for short timescale candidates based on the variogram analysis.

7.6.3 Calibration models

The short-timescale candidate selection criteria are based on the variogram analysis (see Section 7.6.4) with crossmatched catalogues of known variables (including both short and longer timescale sources) and known constant / standard stars, from OGLE catalogues. The idea here is to define a relevant detection threshold γ_{det} that can be compared to the variogram values of each investigated source. This threshold corresponds to the level of variability above which the observed variability is considered as not spurious. A magnitude-dependent detection threshold is defined based on the variogram analysis of crossmatched sources and on the simulation work done previously to assess the power of the variogram method for short timescale variability detection with Gaia (see Roelens et al. 2017).

As mentioned previously, for Gaia Data Release 2, the aim is to focus only on periodic variability with periods below 0.5–1 day. Thus, CU7 also uses the crossmatched catalogues of known constant and variable sources, to define additional criteria to select suspected periodic short-timescale candidates, taking advantage of the period search performed on sources flagged as short timescale candidates from the variogram analysis (see Section 7.6.4). Those additional criteria are basically ‘boxes’ on various metrics, be it classical statistics or specific parameters calculated in the short timescale framework.

Additional criteria are verified by running ‘blindly’ on a subsample of the sources to be investigated, and then are refined to remove some spurious candidates and focus on bona fide on short-timescale suspected periodic candidates, as detailed in Section 7.6.4 and Section 7.6.5.

7.6.4 Processing steps

The short-timescale processing starts with the variogram analysis, similarly to what is described in Roelens et al. (2017). In short, the flagging of short timescale candidates is based on the comparison of the variogram values of the considered source with a magnitude-dependent detection threshold $\gamma_{\text{det}}(\bar{m})$, completed with an upper limit of 0.5 d on the detection timescale τ_{det} (which is the shortest lag for which the variogram value goes above the detection threshold). However, a different formulation of the variogram is used here, based on the IQR and not on the variance. For more details about the variogram approach in the Gaia context, see Roelens et al. (2018).

To define the appropriate detection threshold $\gamma_{\text{det}}(\bar{m})$, the variograms associated to the Gaia light-curves of known OGLE periodic variables (including short timescale sources as well as long timescale ones), and constant sources, with more than 20 FoV transits in G, are calculated. By comparing the maximum variogram values of short timescale, longer timescale and constant sources, as it is done in Roelens et al. (2017), it is possible to retrieve a relevant detection threshold, enabling to separate constant sources from variable stars on the basis of their variograms, and also eliminating a significant fraction of longer period variables. In the end, the detection threshold used is simply a scaled version of the detection threshold deduced from simulations in Roelens et al. (2017): $\gamma_{\text{det}} = 10\gamma_{\text{det,simu}}$. At this point, the recovery rate of short timescale variables is around 50%, contamination from false positives about 2%, and contamination from variable sources with period greater than 1 d around 20%.

For the candidates passing the variogram short timescale selection, a Least-Square period search algorithm is run on the per-CCD time series, searching the frequency range 10min –1d.

The short timescale analysis also relies on classical statistics calculated in the corresponding statistics module, such as the Spearman correlation between the three Gaia photometric bands or the Abbe value on those time series. Additional statistics are defined, such as the ratio of IQRs between the different photometric bands (G, G_{BP} and G_{RP}), or the ratio between the median of variogram values at CCD lags (i.e. up to 40s) and the median of variogram values at FoV lags (i.e. above 40s). They are specific to short timescale analysis, and mostly not published in the Gaia DR2 archive.

So as to both focus the analysis on short-timescale suspected periodic candidates and reduce the contamination from false positives and long period variables, the short timescale analysis (variogram analysis, period search, and complementary statistics calculation) was performed on a few hundred known constant and variable (periodic and non-periodic) sources, not only from the OGLE survey but also from other crossmatched catalogues from the literature (LINEAR, Catalina, etc...). From this analysis, additional cuts on the statistics mentioned above are defined to focus on short period candidates. This series of selection criteria (variogram + cuts on statistics) is refered to as the preliminary selection criteria, and will be refined afterwards (see Section 7.6.5).

7.6.5 Quality assessment and validation

The short timescale suspected periodic selection criteria relies on the analysis of known constant and variable sources from OGLE catalogues. In order to validate the analysis, sources from other catalogues of variable stars such as Catalina, LINEAR, ASAS, AAVSO, etc as well as other resources from the literature are crossmatched with the Gaia data using the Simbad crossmatch tool. Finally, visual inspection of candidate light-curves together with complementary follow-up of some short period variable candidates enabled us to further refine the selection criteria and clean the suspected short period sample.

7.6.5.1 Verification

By applying the preliminary short-timescale selection criteria to all Gaia sources with G CCD photometry available, having more than 20 FoV transits in G, and G a magnitude between 16.5 and 20 mag (which is the range where the variogram detection criterion has been validated), 16 703 sources are selected as preliminary short period candidates. Visual inspection of light-curves of a few hundred randomly selected examples enables to identify several unexpected and probably spurious behaviours, such as G light-curves switching between two discrete magnitude level, or sources exhibiting incompatible behaviours in G, G_{BP} and G_{RP}.

To filter out such spurious variability, cleaning of the sample based on the candidates' environment over the sky (in a similar way as to what is done by Wevers et al. 2018), removing e.g. candidates possibly contaminated by bright nearby sources, have been necessary.

An additional time series cleaning operator has also been applied, specific to the short timescale analysis and based on the expected amplitude of the variation in the G band, to remove the possibly remaining G_{BP} and G_{RP} outliers.

Finally, thanks to extra-cuts on the number of observations, skewness, median variogram ratio and correlation values in G, G_{BP} and G_{RP} bands, the remaining spurious variable candidates have been efficiently excluded.

7.6.5.2 Validation

At this stage, some further validation and black-listing of the short timescale candidates sources has been necessary.

First, a few tens of sources in the sample are reported as showing excess flux features in $G_{\rm BP}$ +$G_{\rm RP}$ compared to G, which have been removed.

Additionally, a few hundred candidates are overlapping with the bona fide eclipsing binaries sample provided by the eclipsing binaries work-package (whose analysis were performed as a test case, but whose results were not made public for Gaia DR2) to CU4 for further analysis and characterization. The publication of new eclipsing binaries identified and characterized from Gaia data is planned only from Data Release 3 and onwards. Hence those few hundred sources are excluded from the published short timescale candidates list.

Finally, after applying all the filtering and refinements described in the previous and current sections, the published list of short timescale, suspected periodic candidates should contain 3018 bona fide sources. This list includes about 138 known variables from the literature catalogues used for quality assessment and validation, with about three quarters of them being period variables with periods below 1 d. All the non-periodic variable and constant sources from these catalogues have been removed from the published short timescale suspected periodic candidates sample. Hence, there is a contamination of about 19% of the sample from longer period variables. However, those sources have periods around a few days, and relatively high amplitudes, hence not being short period variables per se, but whose detection at the short timescale level is justified.

When compared to all the OGLE short period variables processed as part of the global short timescale variability search for Gaia DR2, the completeness of the short timescale suspected periodic candidates sample published is assessed around 0.05%.

Further contamination estimation is performed, using the OGLE photometric database: the Gaia DR2 short timescale sample of 3018 sources is crossmatched with this OGLE catalogue in the Magellanic Clouds, then the OGLE and Gaia time series are compared to check if the features observed in the later are compatible with the former. From this analysis, the real contamination from spurious or non-periodic variability is assessed around 10–20% is those regions.

More details on the Gaia DR2 short timescale analysis results, efficiency and quality, are available in Roelens et al. (2018).

7.7 Long period variables

Author(s): Nami Mowlavi, Isabelle Lecoeur-Taïbi, Thomas Lebzelter

7.7.1 Introduction

The Gaia DR2 provides the first Gaia all-sky catalogue of Long Period Variable (LPV) candidates. They are the result of the processing of a selection of sources classified as `MIRA_SR` with the *nTransits:2+* classifier and the *nTransits:20+* classifier.

Because of the time series properties inherent to the data published in Gaia DR2, i.e. the limited range of observation durations (22 months, to be compared to LPV periods that can be larger than a thousand days) and the small number of observations (a mean of 26 observations per source for all LPV candidates, knowing that the majority of these objects are multi-periodic), we restrict the search of Gaia LPV candidates in Gaia DR2 to sources that satisfy certain criteria. In particular, we consider only large amplitude LPVs (variability amplitudes larger than 0.2 mag in G), and therefore exclude all small amplitude red giants such as OSARGs. Moreover, we do not aim at completeness in Gaia DR2. The selection procedure is described in Section 7.7.2, and the LPV parameters published in Gaia DR2 are described together with the processing steps in Section 7.7.3 and Section 7.7.4. The quality assessment and validation procedures are then presented in Section 7.7.5.

7.7.2 Properties of the input data

The input sources for LPV processing are the ones classified as `MIRA_SR` by the Classification processing (either with *nTransits:2+* or with *nTransits:20+* classifiers) which meet the following selection criteria:

- a minimum number of 12 data points in the G-band. The maximum number of data points is 238. The limit on the number of good data points led to the exclusion of several nearby and bright LPVs.
- a minimum number of 9 data points in the $G_{\rm RP}$-band.
- a colour $G_{\rm BP} - G_{\rm RP} > 0.5$ mag; (colour computed as median($G_{\rm BP}$) - median($G_{\rm RP}$))
- a correlation between the G-band variability and the $G_{\rm BP} - G_{\rm RP}$ colour larger than 0.5 mag. The correlation is computed using the Spearman algorithm.
- a variability amplitude range in G larger than 0.2 mag. The variability amplitude is quantified by the trimmed range (at 95% level).
- a minimum Abbe value (von Neumann (1941) and von Neumann (1942)) of 0.8 on the smoothed G light curve. The smoothed G light curve is computed in an iterative way by merging successive pairs of observations with time difference less than 5 days.

7.7.3 Calibration models

- **Identification of red supergiants.** Long period variables consist of both red giant (mainly on the asymptotic giant branch – AGB) and supergiant stars. According to Wood et al. (1983) red supergiants can be identified based on their absolute bolometric magnitude M_{bol} and main period P. If the LPVs are plotted in a M_{bol}-P diagram, two distinct regions are found: one occupied by AGB stars and one occupied by supergiants. AGB stars never exceed a certain value max($M_{bol,P}$) that depends on the period. Therefore the following relation can be seen as upper luminosity limits:

  $$\max(M_{bol,P}) = -5.62787 - 0.00383 * P + 1.875 * 10^{-6} * P^2 \tag{7.12}$$

  All stars brighter than max($M_{bol,P}$) have been classified as supergiant.

- **Computation of the bolometric correction** $BC(G)$ **for the** G **band** using the $G_{\rm BP} - G_{\rm RP}$ colour (which is computed as mean($G_{\rm BP}$) - mean($G_{\rm RP}$)).

  The procedure distinguishes three different cases:

  1. Supergiants. Based on the compilation of M_{bol} values for red supergiants by Levesque et al. (2005) a mean value of BC_G =-0.71±0.3 mag was chosen for all red supergiants, where the bolometric error represents the standard deviation around the mean value.

2. LPVs with G amplitude larger than 3 mag (Mira-like). The G amplitude is computed as the 5-95% trimmed range, using the LEGACY strategy of commons-math to compute the percentiles In Gaia DR2, no distinction was made between M/S- and C-stars. The correction function is based on synthetic spectra computed with MARCS. A fixed value of BC_G =-2.2±0.005 mag was used in Gaia DR2, based on Kerschbaum et al. (2010).
3. Other LPVs. The following relation is used, based on synthetic spectra of hydrostatic M-star models (Aringer et al. 2016):

$$\begin{aligned} \mathrm{BC}_G &= 0.2438 - 0.25155 * (G_{\mathrm{BP}} - G_{\mathrm{RP}}) - 0.11433 * (G_{\mathrm{BP}} - G_{\mathrm{RP}})^2 \\ &\quad +0.00154 * (G_{\mathrm{BP}} - G_{\mathrm{RP}})^3 \end{aligned} \tag{7.13}$$

   However, if the uncertainty in G_{BP} or G_{RP} is larger than 4 mag, the bolometric correction is computed as if $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ = 3.25 mag, i.e. BC_G = -1.729±1.892 mag.

The colour has not been corrected for extinction since no reddening correction was available during our processing.

It is important to note that these bolometric corrections are preliminary results. Improved values will be provided in Gaia DR3.

7.7.4 Processing steps

The SOS LPV processing applied the following steps to the pre-selected LPVs as shown in Figure 7.5:

- Computation of a variability period based on the G time series.
  The period search method used is Least Square, applied to a frequency range from 0.001 c/d to 0.1 c/d with a frequency step of $5 * 10^{-5}$ c/d. In Gaia DR2, only sources with periods greater than 60 days are published, in particular because of the aliasing in the period search.

- Computation of bolometric correction.
  As described in Section 7.7.3.

- Determination of the absolute bolometric magnitude M_{bol}.
  Using the Gaia parallax measurement (which was still preliminary at the time of our processing), the mean G-band magnitude, and the bolometric correction, the absolute bolometric magnitude was calculated using (with ϖ in arcsec):

  $$M_{bol} = m_G + \mathrm{BC}_G - A_G + 5 \log \varpi + 5 \tag{7.14}$$

  where m_G is the mean G-band magnitude, BC_G the bolometric correction, A_G the interstellar extinction which has been forced to 0 as it was not available as input and ϖ the parallax given in arcseconds. The uncertainty of the absolute bolometric magnitude is derived via:

  $$\sigma_{M_{bol}} = \sqrt{\sigma^2_{m_G} + \sigma^2_{BC} + \sigma^2_{A_G} + 4.715\, \varpi^{-2}\, \sigma^2_{\varpi}} \tag{7.15}$$

  Note that the uncertainty on the extinction has been forced arbitrarily to 0.05 mag. We point out that the absolute bolometric magnitude and its error is depending on the calculation type of the mean G magnitude and its error. Due to the variability of LPVs the values of m_G and σ_{m_G} could be misleading. The derived light amplitudes in the G band could also be underestimated compared to the real ones depending on the coverage and gaps of data points and hence on the quality of the lightcurve.

- Set of red supergiant flag.
  Stars brighter than M_{bol_P} (cf. (7.12)) are flagged as red supergiant.

Finally, each LPV candidate is published in the `vari_long_period_variable` table (Section 14.3.5) with the following attributes:

- one frequency (and the associated uncertainty),
- the bolometric correction (and the associated uncertainty),
- the absolute bolometric magnitude (and the associated uncertainty),
- a red super giant flag.

7.7.5 Quality assessment and validation

7.7.5.1 Verification

The good execution of the processing steps shown in Figure 7.5 is tested in this section through a series of graphical representations of the results. The discussion of the results is presented in more details in Mowlavi et al. (2018). All results presented here concern only the subset of LPV candidates in Gaia DR2 that have LPV-specific results, excluding the candidates published as part of the classification results that would not have specific LPV results.

The period distribution is shown in Figure 7.6. All LPV candidates with LPV-specific results have periods larger than 60 d, in agreement with the filter criterion imposed for Gaia DR2. Among them, 164 (0.1%) have periods above 1000 d, the upper limit set in the Least Square period search algorithm. This is due to the application of a non-linear period refinement algorithm on the most significant period found by the Least Square algorithm. At this stage, all these periods above 1000 d have to be taken with caution, and be confirmed with subsequent Gaia data releases which will cover larger observation durations.

The colour-magnitude diagram of the LPV candidates published in the *vari_long_period_variable* table is shown in Figure 7.7. Three groups of stars are visible. The main group with the largest concentration of stars extends from $G_{\rm BP} - G_{\rm RP} \simeq 2$ mag to ~7.5 mag. The extension of the colours to such large values results from reddening. The second group, at $G \simeq 16$ mag and with colours extending from $G_{\rm BP} - G_{\rm RP} \simeq 1.8$ mag to ~3.5 mag, represents LPV candidates from the Large and Small Magellanic Clouds. The presence of the Clouds is further noticeable in the histogram of G magnitudes shown in Figure 7.8. Finally, a third group of stars is visible in Figure 7.7 below the main stream, with G magnitudes above ~17 mag. These stars have large $G_{\rm BP}$ uncertainties, as witnessed by the $G_{\rm BP}$ uncertainties colour-coded in the figure, leading to very uncertain $G_{\rm BP} - G_{\rm RP}$ values that explain the presence of this group of stars below the main stream.

The bolometric corrections are shown in Figure 7.9 as a function of $G_{\rm BP} - G_{\rm RP}$ colour. We remind here that the colours used in the LPV processing pipeline of Gaia DR2 are computed from the mean values of $G_{\rm BP}$ and $G_{\rm RP}$, and that they are not corrected for extinction. The relation given by (7.14) is clearly visible in the figure, as well as the constant values of the bolometric corrections adopted in the specific cases of red supergiants, of LPVs with large amplitudes, and of sources with large uncertainties on either $G_{\rm BP}$ and/or $G_{\rm RP}$ (see Section 7.7.3). The bolometric correction versus the 5%-95% trimmed range of the G-band time series is shown in Figure 7.10. It illustrates the adoption of a value of -2.2 mag for all LPVs having a G variability amplitude measured by this trimmed range larger than 3 mag. It must be noted that the trimmed range used in the figure uses the $\rm R_1$ strategy of commons-math to compute the percentiles, while the computation in the LPV code was adopting the LEGACY strategy.

This explains why some LPV candidates in Figure 7.10 also have their bolometric correction set to 2.2 mag while having a 5%-95% trimmed range of *G* smaller than 3 mag.

The bolometric magnitudes published in Gaia DR2 are shown in Figure 7.11. It must be stressed that the parallaxes used in the figure are those used for the computation of the bolometric magnitudes of LPV candidates published in Gaia DR2, which are not the final parallaxes published in Gaia DR2, which were not available at the time of the variability processing of LPV candidates. The improvements brought to the parallax computations are illustrated in Figure 7.12.

7.7.5.2 Validation

The periods derived in Gaia DR2 are validated against the periods published in OGLE-III for the LMC, SMC and the region towards the Galactic bulge, for the Gaia DR2 LPV candidates that crossmatch OGLE-III LPVs. Likewise, the degree of completeness and percentage of new candidates *relative to* OGLE-III can be estimated by comparing the fraction of OGLE-III LPVs that are cross-matched with Gaia LPV candidates and the fraction of Gaia LPV candidates that are absent from OGLE-III catalogues, respectively. About 30% of OGLE-III LPVs in the LMC and SMC are present in Gaia DR2, while less than a few percent of the Gaia DR2 LPV candidates towards the Clouds are not present in OGLE-III. The results of these comparisons are presented in Mowlavi et al. (2018).

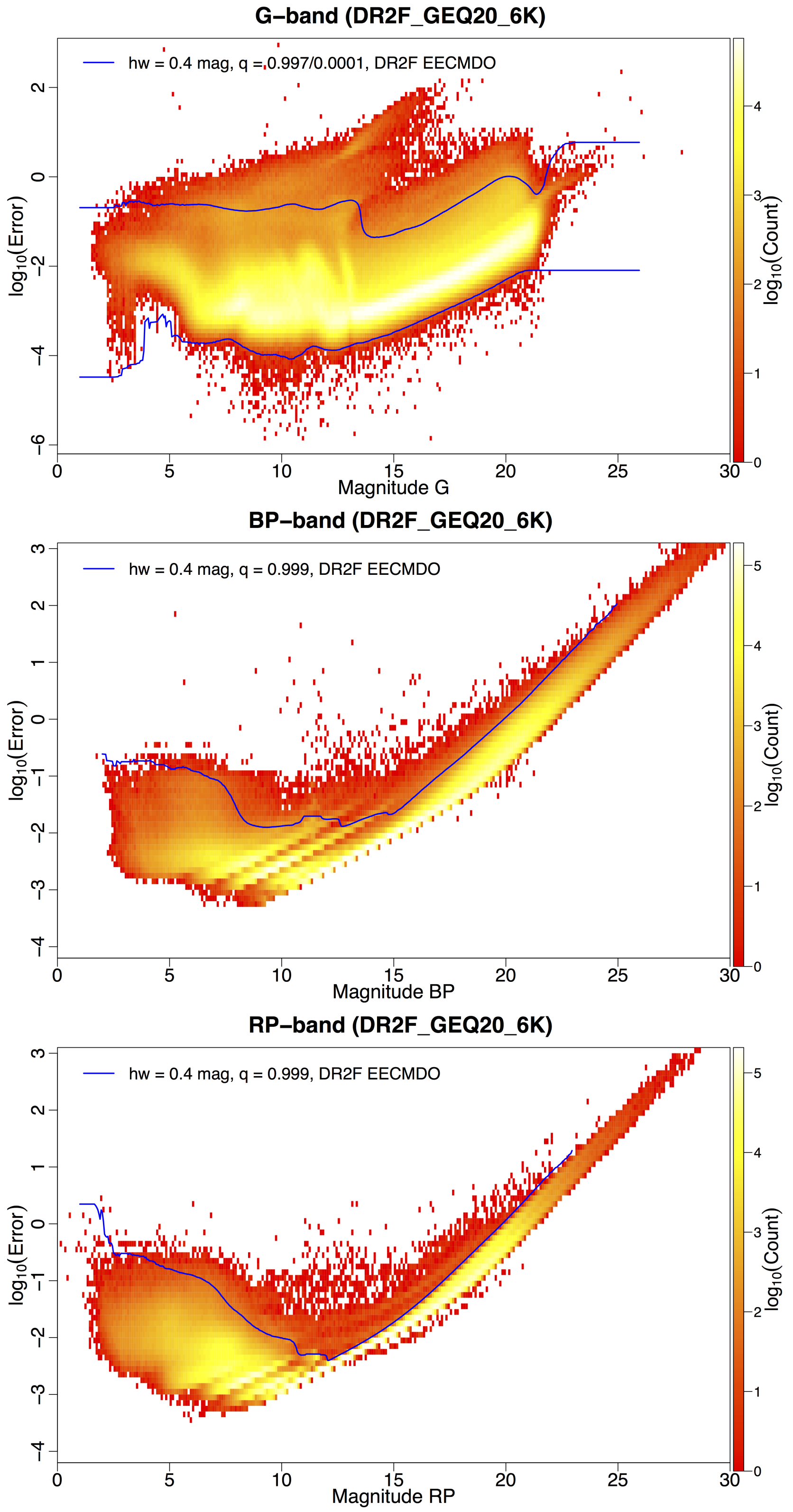

Figure 7.4: Transit magnitude error distributions vs transit magnitude for G, $G_{\rm BP}$ and $G_{\rm RP}$, using the ≥ 20 G-FoV input data set. In blue are shown the thresholds for the `ExtremeErrorCleaningMagnitudeDependent` operator.

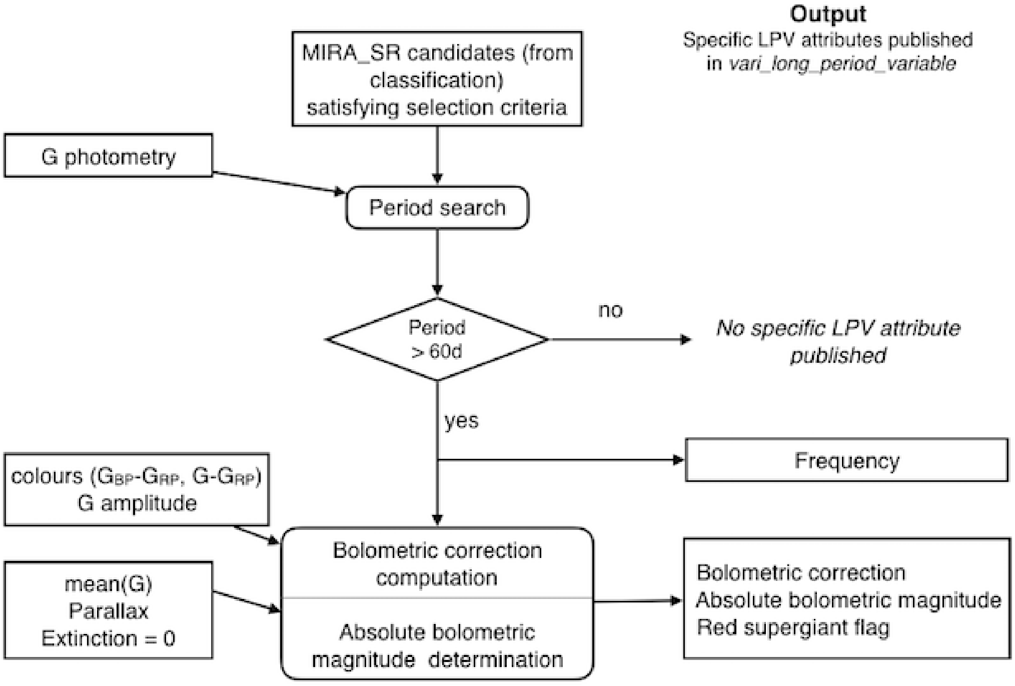
Output
Specific LPV attributes published in *vari_long_period_variable*
MIRA_SR candidates (from classification) satisfying selection criteria
G photometry
Period search
Period > 60d
no
No specific LPV attribute published
yes
Frequency
colours (G_{BP}-G_{RP}, G-G_{RP}) G amplitude
mean(G) Parallax Extinction = 0
Bolometric correction computation
Absolute bolometric magnitude determination
Bolometric correction Absolute bolometric magnitude Red supergiant flag

Figure 7.5: Gaia DR2 Long Period Variable processing overview.

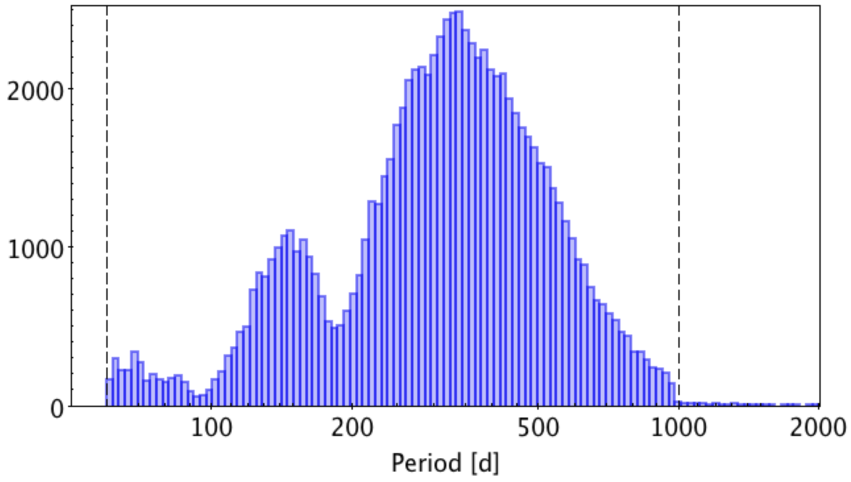

Figure 7.6: Period distribution of LPV candidates published in Gaia DR2 table `vari_long_period_variable` (Section 14.3.5). The X-axis has been limited to an upper limit of 2000 d for clarity of the figure, 24 sources being omitted in the figure for having periods longer than this upper limit. A vertical dashed line is drawn at the period of 60 days above which LPV candidates are published in table `vari_long_period_variable`. A second vertical dashed line, at 1000 days, gives the upper range limit considered for period search. See text for details.

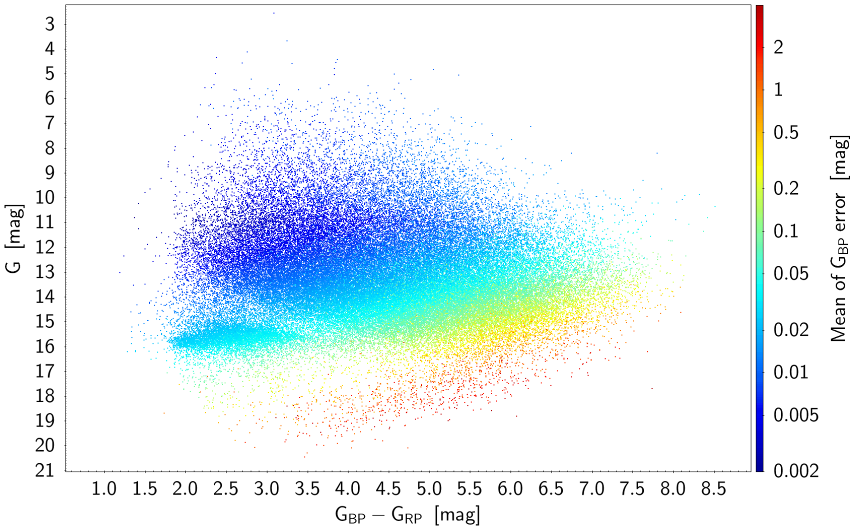

Figure 7.7: Colour-magnitude diagram of LPV candidates published in Gaia DR2 table `vari_long_period_variable`. Medians values of G, G_{BP} and G_{RP} are used. Each point is colour-coded relative to the mean uncertainty of the G_{BP} magnitude according to the colour scale shown on the right of the figure. The LMC/SMC population forms the elongated clump seen at G magnitudes around 16 and $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ colours between about 2 mag and 3.5 mag.

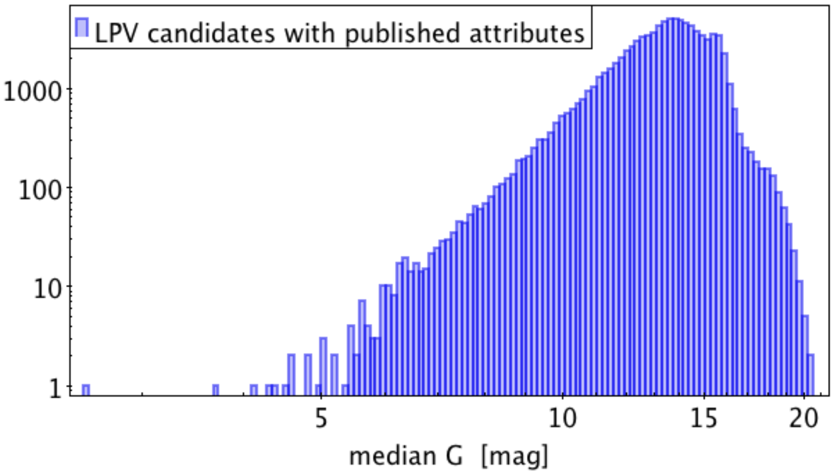

Figure 7.8: Magnitude distribution of LPV candidates published in Gaia DR2 table `vari_long_period_variable`. The median value of G is used.

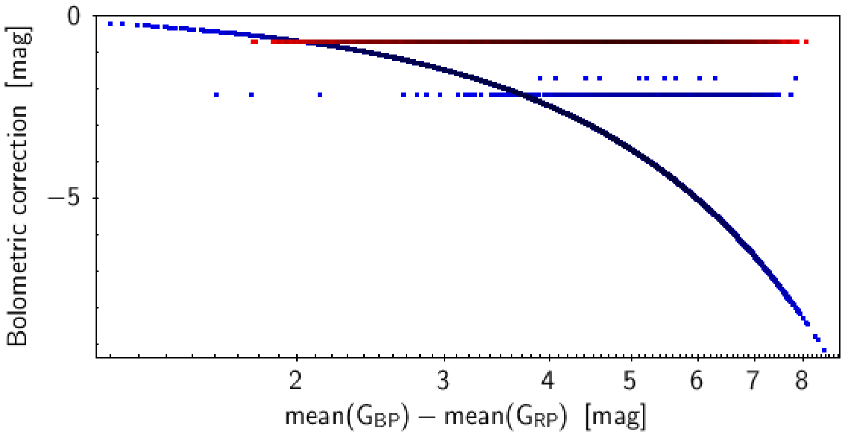

Figure 7.9: Bolometric correction versus $G_{\rm BP} - G_{\rm RP}$ colour of LPV candidates in Gaia DR2, using the mean magnitudes of $G_{\rm BP}$ and $G_{\rm RP}$ to compute the colour $G_{\rm BP} - G_{\rm RP}$, as was done for the computation of Gaia DR2 bolometric corrections. Red supergiants are identifiable at -0.71 mag (shown in red in the figure).

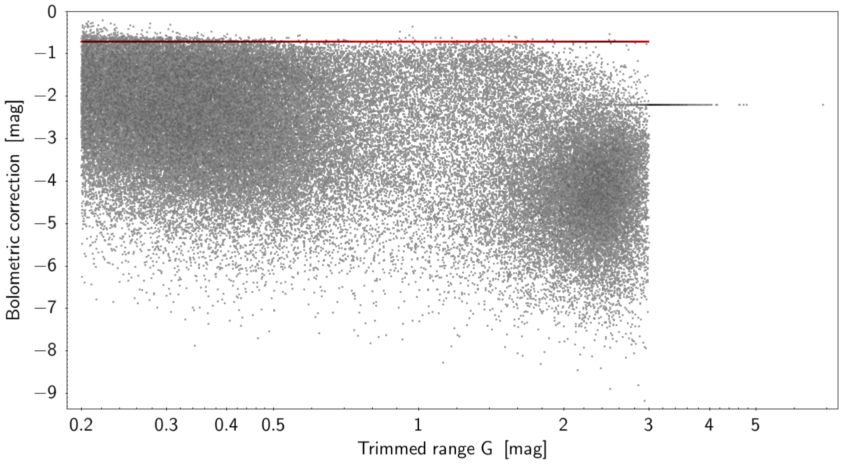

Figure 7.10: Bolometric correction versus G amplitude of variability of LPV candidates in Gaia DR2. The variability amplitude is measured by the 5%-95% trimmed range of the G magnitude time series. Red supergiants are identifiable at -0.71 mag (shown in red in the figure).

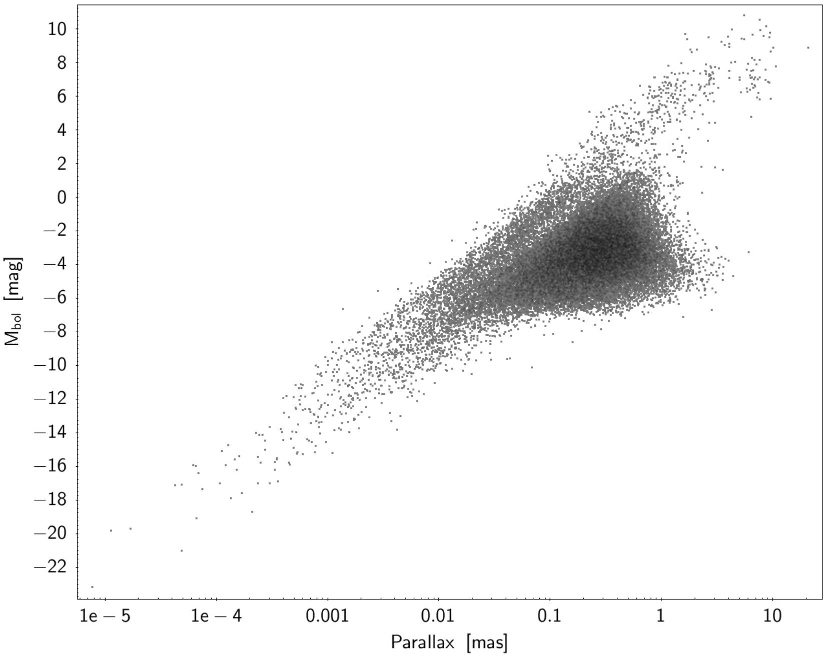

Figure 7.11: Bolometric magnitude versus parallax in milliarcsec of LPV candidates in Gaia DR2. It must be noted that the parallaxes used in the figure are those used for the computation of the bolometric magnitudes published in Gaia DR2, which are not the final parallaxes published in Gaia DR2.

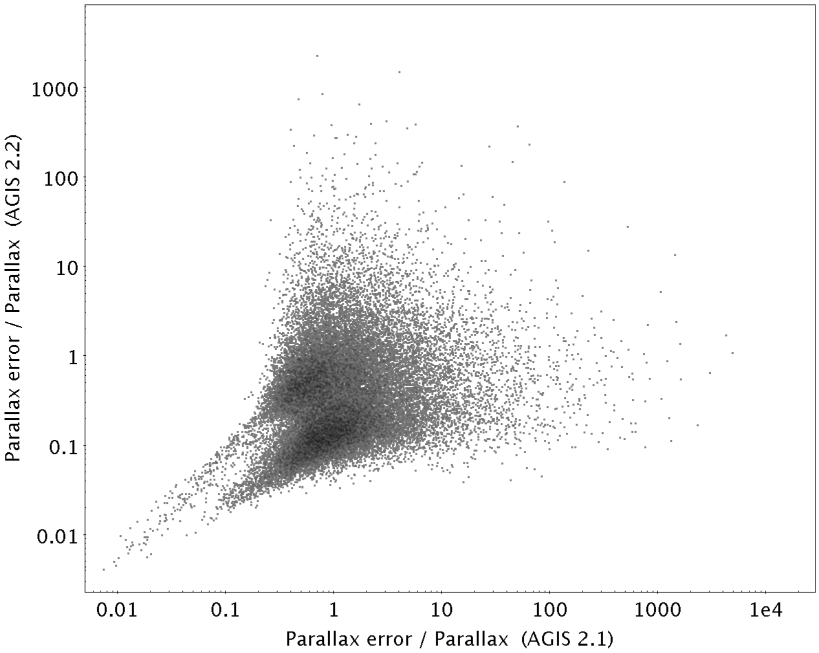

Figure 7.12: Comparison of the relative parallax errors between two versions of the Astrometric Global Iterative Solution used to compute parallaxes: version 2.1 available at the time of variability processing of LPV candidates on the X-axis versus version 2.2 published in Gaia DR2 on the Y-axis. Only sources with positive parallaxes are reported in the diagram. The vertical dashed (horizontal dotted) line locates the region in the diagram where the relative parallax uncertainty is 10% with AGIS version 2.1 (2.2). The diagonal blue line is drawn as an eye guide. Sources located below the diagonal line have an AGIS 2.2 relative parallax uncertainty smaller than the relative uncertainty that was available in AGIS 2.1.