Summary of the variability processing and analysis results

B. Holl\textsuperscript{1,}\textsuperscript{†}, M. Audard\textsuperscript{12}, K. Nienartowicz\textsuperscript{3}, G. Jevault de Fombelle\textsuperscript{4,}, O. Marcha\textsuperscript{24}, N. Mowlav\textsuperscript{11}, G. Clementini\textsuperscript{5}, J. De Ridder\textsuperscript{6}, D.W. Evans\textsuperscript{4}, L.P. Guy\textsuperscript{4}, A.C. Lanzafame\textsuperscript{19}, T. Lebzelter\textsuperscript{11}, L. Rimoldini\textsuperscript{2}, M. Roelens\textsuperscript{15}, S. Zucker\textsuperscript{13}, E. Distefano\textsuperscript{10}, A. Garofalo\textsuperscript{12}, I. Lecoqeur-Taibi\textsuperscript{3}, M. Lopes\textsuperscript{40}, R. Molinaro\textsuperscript{40}, T. Muraveva\textsuperscript{36}, A. Panahi\textsuperscript{17}, S. Regibo\textsuperscript{6}, V. Ripepi\textsuperscript{16}, L.M. Sarro\textsuperscript{47}, C. Aerts\textsuperscript{23}, R.I. Anderson\textsuperscript{19}, J. Charna\textsuperscript{3}, F. Barblan\textsuperscript{9}, S. Blanco-Cuaresma\textsuperscript{10,11,11,11}, G. Busso\textsuperscript{47}, J. Cuypers\textsuperscript{27}, F. De Angel\textsuperscript{27}, F. Glas\textsuperscript{21}, M. Grenon\textsuperscript{1}, Á.L. Juhász\textsuperscript{23,29}, A. Kochoska\textsuperscript{23,29}, P. Koubsky\textsuperscript{27}, A.F. Lanza\textsuperscript{10}, S. Leccia\textsuperscript{10}, D. Lorenz\textsuperscript{14}, M. Marconi\textsuperscript{16}, G. Marschalkó\textsuperscript{1,2,5}, T. Mazet\textsuperscript{13}, S. Messina\textsuperscript{10}, F. Mignani\textsuperscript{29}, A. Moitinho\textsuperscript{30}, L. Moln\textsuperscript{23}, S. Morgenthaler\textsuperscript{31}, I. Musella\textsuperscript{10}, C. Ordenovic\textsuperscript{29}, D. Ordoñez\textsuperscript{23}, I. Pagano\textsuperscript{10}, L. Palaversa\textsuperscript{23}, M. Pawlak\textsuperscript{32,33}, E. Plachy\textsuperscript{23}, A. Prš\textsuperscript{25}, M. Riello\textsuperscript{4}, M. Süveges\textsuperscript{14}, L. Szabados\textsuperscript{23}, E. Szegedi-Elek\textsuperscript{23}, V. Votruba\textsuperscript{27}, and L. Eyer\textsuperscript{13}

(Affiliations can be found after the references)

Received; accepted

ABSTRACT

Context. The Gaia Data Release 2 (DR2) contains more than half a million sources that are identified as variable stars.

Aims. We summarise the processing and results of the identification of variable source candidates of RR Lyrae stars, Cepheids, long-period variables (LPVs), rotation modulation (BY Dra-type) stars, δ Scuti and SX Phoenicis stars, and short-timescale variables. In this release we aim to provide useful but not necessarily complete samples of candidates. In the second data release (DR2), Gaia Collaboration et al. (2018b) we extend the variability types for which we published calibrated data of variable objects detected by Gaia to the following list of classes: Cepheids, RR Lyrae stars detected in the Large Magellanic Cloud (LMC), parallaxes. Various methods from classical statistics, data mining, and time-series analysis were applied and tailored to the specific properties of Gaia data, as were various visualisation tools to interpret the data.

Results. The DR2 variability release contains 228,904 RR Lyrae stars, 11,438 Cepheids, 151,761 LPVs, 147,535 stars with rotation modulation, and 3,018 short-timescale variables. These results are distributed over a classification and various Specific Object Studies tables in the Gaia archive, along with the three-band time series and associated statistics for the underlying 550,737 unique sources. We estimate that about half of the new sources are variable stars and the variability type completeness varies strongly as a function of sky position as a result of the non-uniform sky coverage and intermediate calibration level of these data. The probabilistic and automated nature of this work implies certain completeness and contamination rates that are quantified so that users can anticipate their effects. This means that even well-known variable sources can be missed or misidentified in the published data.

Conclusions. The DR2 variability release only represents a small subset of the processed data. Future releases will include more variable sources and variable products; however, DR2 shows the (already) very high quality of the data and great promise for variability studies.

Key words. – Stars: general – Stars: oscillation – Stars: solar-type – Stars:variables: general – Galaxy: stellar content – Catalogs

1. Introduction

The Coordination Unit 7 (CU7) of the Gaia Data Processing and Analysis Consortium (DPAC) is tasked to process the calibrated data of variable objects detected by Gaia. In Gaia Data Release 1 (DR1, Gaia Collaboration et al. 2016), we published the light curves and properties of a sample of Cepheids and RR Lyrae stars detected in the Large Magellanic Cloud (LMC), see Clementini et al. (2016). We refer to Eyer et al. (2017) for a detailed description of the CU7 framework and processing pipeline. In this second data release (DR2, Gaia Collaboration et al. 2018b), we extend the variability types for which we publish light curves to the following list of classes: Cepheids, RR Lyrae stars, long-period variables (LPV), short-timescale variables, stars with rotation modulation, and δ Scuti and SX Phoenicis stars. For some of them, additional specific properties are inferred and published. This constitutes roughly 100 times as many variable sources as were published in DR1 (Eyer et al. 2017; Clementini et al. 2016). Moreover, it covers the whole sky. The variability analysis for this release focused primarily on large-amplitude variable stars and rotation modulated stars. Other classes of variable stars are aimed to be introduced in later releases. We note that eclipsing binaries were identified as well, but will be treated separately and only delivered in future releases. We emphasize that specific selection criteria were applied to each variability type to limit contamination. Completeness within each variability type was not aimed at for DR2.

The various variability type-specific methods that were used for the published DR2 results are discussed in full length in the following dedicated papers:

– DR2: Variable stars in the Gaia colour-magnitude diagram: Gaia Collaboration et al. (2018a)

– DR2: All-sky classification of high-amplitude pulsating stars: Rimoldini & et al. (2018)

– DR2: Specific characterisation and validation of all-sky Cepheids and RR Lyrae stars: Clementini et al. (2018)
This overview paper focuses on presenting the general properties of the exported sample of variable sources in the following way: Section 2 introduces the time-series photometric data and their relevant properties. Section 3 presents a summarised overview of the data processing and analyses. Section 4 gives an overview of the results, including our time-series filters, published time-series, classification, and Specific Object Studies (SOS) tables containing the results. Section 5 presents the final conclusions. Additional information regarding the Gaia archive queries used in this paper is provided in Appendix A.

The data are publicly available in the online Gaia archive at `http://gea.esac.esa.int/archive/` containing the ‘tables’ and ‘fields’ referred to in this rest of this article, as well as all of the ‘DR2 documentation’ and catalogue ‘data model’ listed at `http://gea.esac.esa.int/archive/documentation/`.

2. Data

The results in DR2 are based on the first 22 months of Gaia data, taken between 25 July 2014 and 23 May 2016. For the variability analysis we made use of the photometric time series in the $G$, $G_{BP}$, and $G_{RP}$ bands, using field-of-view (FoV) averaged transit photometry. An exception to this was the analysis of short-timescale variables, for which some per-CCD $G$-band photometry was also analysed, but not published. The photometry is described in detail in Evans et al. (2018) and Riello et al. (2018). We emphasise that owing to the tight DR2 data-processing schedule, we only had access to a preliminary version of the DR2 astrometric solution (Lindegren et al. 2018), on which our results are based. No radial velocity data or astrophysical parameters were available during our processing.

2.1. Relative completeness

As a consequence of the Gaia scanning law, the sky-coverage is rather non-uniform for the 22 months of DR2 data, resulting in sky-coverage gaps when selecting a certain minimum number of FoV transits. In our processing we used cuts of $\geq 2$, $\geq 12$, and $\geq 20$ $G$-band FoV transits, which we discuss in Sect. 3. Figure 1 shows the effect of these cuts on the sample of all available time series at the time of processing. Assuming that all sources with $\geq 2$ FoV transits are bona fide sources, these figures show that a cut of $\geq 12$ FoV transits results in a relative completeness of $80\%$, and a cut of $\geq 20$ FoV transits results in a relative completeness of only $51\%$. It can generally be expected that these relative completenesses are an upper limit to the absolute (sky) completeness, which is assessed in detail in the follow-up variability papers mentioned in the introduction. Arenou et al. (2018) conclude that the catalogue is mostly complete, in the absolute sense, roughly between magnitudes 7 and 17, and incomplete beyond. For a first-order relative completeness assessment in certain general Galactic directions, we introduce the following regions and compute the Gaia source count in each of them:

- Galactic plane $30^\circ < l < 330^\circ$ & $|b| < 15^\circ$
- Galactic centre $20^\circ > l > 340^\circ$ & $|b| < 15^\circ$
- High Galactic latitudes $|b| > 45^\circ$, excluding a $10^\circ$ radius around the SMC
- LMC region within a radius of $7.5^\circ$ from Galactic coordinates $l = 280.47^\circ$, $b = -32.89^\circ$
- SMC region within a radius of $2.5^\circ$ from Galactic coordinates $l = 302.81^\circ$, $b = -44.33^\circ$

We repeat this for the (approximate) sky footprints of various external catalogues, list the results in Table 1 and display the regions in Fig. 1. It shows that a (maximum) completeness of between $70\%$–$85\%$ for $\geq 12$ $G$-band FoV transits and a relative completeness of between $50\%$–$70\%$ for $\geq 20$ $G$-band FoV transits can be expected. The bottom panel of Fig. 1 shows that various sky regions are not or only partially sampled for $\geq 20$ FoV: the Galactic centre region, for example, has a relative completeness of only $14\%$.

We remark that the numbers in Table 1 and Fig. 1 do not correspond exactly to the number of sources published in the gaia_source table of DR2 because various source filters in the global Gaia processing chain were applied before and after the variability processing. The derived percentages match very well, however, and they are relevant for the presented discussion.

2.2. Outliers and calibration errors

Various issues are visible in the current photometric data. Some of them are caused by instrumental effects, while others are sky related. Improved calibration strategies and careful flagging of these events are being designed and will be implemented in the chain of processing systems that convert the raw data into calibrated epoch photometry for the next Gaia Data Release. Some of these issues are also described in Evans et al. (2018):
- hot pixel columns are not yet treated,
- poor background estimates exist for some observations,
- observations close to bright sources can have biased background estimates or might even be spurious detections due to diffraction spikes,
- spatially close sources can have scan-angle direction dependant outliers,
- isolated sources can still be affected through overlap of the two FoVs.

It is expected that a certain fraction of observation affected by these issues has not been flagged and hence users should be aware of such unflagged outliers in the published data. We discuss variability flagging in more detail in Sect. 4.1.

In addition to the flagging of observations, the Gaia data processing consortium has applied various filters that remove sources that are affected by specific calibration issues; see Gaia Collaboration et al. (2018a) for more details.

3. DR2 variability data processing and analysis

The general variability data processing has been described in detail in our DR1 paper (Eyer et al. 2017), to which we refer here. We here only summarise the main aspects that are specific to the DR2 release. More exhaustive information is provided in the DR2 documentation.

The variability processing is summarised in Fig. 2 and the number of sources involved at various levels of the processing is indicated in the diagram. The processing followed two main paths: one starting from 1607 million sources with $\geq 2$ $G$-band FoV transits (hereafter geq2), and the other from 826 million sources with $\geq 12$ $G$-band FoV transits (hereafter geq12).

- DR2: Rotation modulation in late-type dwarfs: Lanzafame et al. (2018)
- DR2: Short-timescale variability processing and analysis: Roelens et al. (2018)
- DR2: The first Gaia catalogue of long-period variable candidates: Mowlavi et al. (2018)
- DR2: Validation of the classification of RR Lyrae and Cepheid variables with the Kepler and K2 missions: Molnár et al. (2018)
sources with ≥ 20 G-band FoV transits (hereafter geq20). The ≥ X means a G-band time series with X non-null and positive FoV flux observations, but before the variability filters were applied that are described in Sects. 4.1 and 4.2. The first path resulted in the published nTransits: 2+ classification results, as described in Sect. 4.3 as well as in a split-off path with ≥ 12 G-band FoV transits (hereafter geq12) that served as one of the inputs for the SOS modules: SOS Cep and RRL (Clementini et al. 2016; 2018) and SOS LPV (Mowlavi et al. 2018).

The second path leads to the majority of the pipelines and SOS results, as described in Sect. 4.4. It had two different variability detection methods. The first is a ‘general’ variability detection, defined by a classifier trained on known cross-matched constant objects1 from OGLE4 (Soszyński et al. 2012), Hipparcos (ESA 1997), the Sloan Digital Sky Survey (SDSS) (Ivezić et al. 2007), and classified constants, as well as many additional catalogues for variables (Rimoldini & et al. 2018). The second consists of multiple dedicated variability detection methods: one tuned to detect short-timescale variables and another to rotation modulation effects, which were subsequently followed by their associated SOS modules: SOS short timescale (Roelens et al. 2018) and rotation modulation (Lanzafame et al. 2018), analysing these sources in more detail. The general variability detection was followed by a generic multi-harmonic Fourier modelling based on the periodogram peak-frequency of an unweighted least-squares period search. Various classification attributes were derived from this model and the time-series statistical parameters, on which the classifier for the geq20 path was trained. The classified RR Lyrae stars, Cepheids, and LPVs were subsequently fed into the corresponding modules SOS Cep&RRL and SOS LPV, being their second input source. All modules of the variability pipeline include validation rules to ensure that the output values are within acceptable ranges, which allowed us to identify issues early on in the processing.

An accounting of the number of sources published in the classification and SOS modules is given in Fig. 3. It shows that of the 550 737 unique variable sources, 363 969 (66%) appear in the classification results (Sect. 4.3) and 390 529 (71%) in SOS (Sect. 4.4). For several variability types, 203 681 sources (37%) appear in both, while the rest of the sources appear only in either classification (160 288, i.e. 29%) or SOS results (186 768, i.e. 34%). Note that the number appearing in both classification and SOS tables (203 681) is 721 more than the sum of overlapping types shown in Fig. 3 because these are cases of different assigned types in classification and SOS tables, which are omitted from the figure.

Depending on the type of object, there might only exist either a classification or SOS result. SOS results have generally lower contamination than the classification results and contain various type-specific astrophysical parameters that can be used to obtain more detailed information (such as the period), although this information might be available for a less sky-uniform sample due to the minimum number of G-band FoV observations of 12 or 20, depending on the SOS type. The classification contains only a label and ‘score’ and has generally higher contamination rates, but it contains larger samples that are (more) uniformly distributed over the sky due to the minimum number of G-band FoV observations, which are 5 for RR Lyrae, 6 for δ Scuti and SX Phoenicis, 9 for Cepheids, and 12 for LPVs. These differences between classification and SOS in contamination rate and sample size are part of the pipeline design in which classification results feed into most of the SOS modules, as shown in Fig. 2 and discussed in more detail in the DR1 paper Eyer et al. (2017).
– RemoveNaNNegativeAndZeroValuesOperator flagged and removed transits with NaN, negative, or zero flux values.
– RemoveDuplicateObservationsOperator flagged and removed pairs of transits with too close observation times (essentially one transit from the pair was incorrectly assigned to the source). This operator was responsible for the majority of the transit removals for $G$ magnitudes brighter than $-8$; see Mowlavi et al. (2018) for more details.
– GaiaFluxToMagOperator converted the Gaia fluxes into magnitude using the Gaia zero-point magnitudes.
– ExtremeValueCleaning flagged and removed transits with unrealistically faint magnitudes. The cut values were chosen to be $G \geq 25$, $G_{BP} \geq 24$, and $G_{RP} \geq 22$.
– ExtremeErrorCleaningMagnitudeDependent flagged and removed transits with extreme magnitude errors. For $G$, we used both a lower (0.01%) and upper (99.7%) threshold based on the observed distribution of transit magnitude errors (as a function of magnitude), while for $G_{BP}$ and $G_{RP}$, we used only an upper threshold (99.9%).
– RemoveOutliersFaintAndBrightOperator flagged and removed the time-series outliers of the source. Different criteria were used to account for the distribution of magnitudes or of the magnitude errors in the time series.

The time series following the above last operator was then the input for the different variability modules, and all transits that were flagged and filtered in the operator chain can be identified in the published time series by the rejected_by_variability flag (see Sect. 4.2). The resulting number of selected (i.e. not-rejected) transits can be found in the num_selected_g_fov/bp/rp fields of the vari_time_series_statistics table. In Figs 4 and 5 the rejected transits are plotted as crosses. However, some SOS modules (e.g. SOS Cep and RRL, SOS rotation modulation) applied individual stricter conditions that were tuned for their analysis. We note that the SOS short timescale module also used another operator, RemoveOutlierPerTransitOperator, to remove CCD outliers per transit.

The sky distribution of the selected number of transits in the $G$, $G_{BP}$, and $G_{RP}$ time series are shown in Fig. 6 for all 550,737 published sources. The distribution of the mean values per sky-pixel is dominated by the scanning law, which is also visible in the large masked regions of the $\geq 20$ FoV panel of Fig. 1. Figure 10 shows a histogram of the time-series lengths. The minimum number of transits in $G$ is 5.

4.2. Time series, statistics, and variability flag

The full list of variable stars published in DR2 can be easily identified in the gaia_source table by the field phot_variable_flag, which is set to VARIABLE, while for all other sources it is set to NOTAVAILABLE. The time series of all variable stars are published in $G$, $G_{BP}$, and $G_{RP}$. In contrast to DR1, the time series are no longer provided in a separate archive table, but instead are accessible in a Virtual Observatory Table (VOTable) linked via the field epoch_photometry_url in the gaia_source table. For various reasons, a fraction of the transit observations was not photometrically processed in DR2, resulting in null values. The variability processing flags per FoV transit (see previous section) are provided in the column rejected_by_variability of the VOTable epoch photometry. Additional flags from the photometric processing are available for transits in the column rejected_by_photometry. Unchanged from DR1 is the expression of observation time in units of barycentric JD (in TCB) in days $-2,455,197.5$, see DR2 documentation or Eyer et al. (2017) for more details.

All time series are provided in both flux and magnitude, where the flux is provided with the associated standard error. The magnitude-transformed error value is omitted given its nonsymmetric nature. The Vega-system flux to magnitude zero-points are defined in Evans et al. (2018).

Several basic statistical parameters of the cleaned time series are listed in the vari_time_series_statistics table. The variable stars will furthermore have an entry in at least one of the vari_* tables, see Fig. 5 for their numbers. An overview of time-series statistics per Gaia band and types is provided in Tables 4 and 5. The mean magnitudes provided in gaia_source (phot_g/bp/rp_mean_mag) differ from the mean magnitudes in vari_time_series_statistics (mean_mag_g_fov/bp/rp) because they are calculated differently (see Evans et al. 2018) and because the filtering of the light curves was different; furthermore, the gaia_source mean photometry can be absent, but present in the vari_time_series_statistics table (together with the related time series in epoch_photometry_url) because various filters were applied by the photometry pipeline. The sky and median magnitude distributions of these sources are shown in more detail in Figs. 4 and 5.

4.3. Classification

In DR2 we introduce the output of the semi-supervised classifier targeted to identify high-amplitude variable stars over the whole sky, resulting from the geo2 path described in Sect. 3. This is the aim of the nTransits:2+ classification. Owing to the limited 22 months of data and rejected observations, the number of FoV transits per source can be very small, prohibiting the use of Fourier modelling parameters for a large portion of the sources. Previous studies such as (Ivezić et al. 2000 and Hermscheik et al. 2016) have shown that it is still possible to identify good RR Lyrae star candidates with just a few observations. Our classifier was trained to identify high-amplitude variables of the type of RR Lyrae stars (anomalous RRd, see loszynski et al. 2016; RRd, RRab, RRc), LPV (Mira-type and semi-regulars), Cepheids (anomalous Cepheids, δ Cepheids, type-II Cepheids), δ Scuti and SX Phoenicis stars, amongst the many other (non-)variable source types that exist. These classes can be found in the Gaia archive vari_classifier_result table with best_class_name = ARRd, RRd, RRab, RRc, MIRA_SR, Acep, Cep, T2CEP, and DSCT_SXPH, respectively. Each entry has an associated best_class_score between 0 and 1. Additional (non-published) classes were included in the prediction types, see Rimoldini & et al. (2018) for all details. Descriptions of the classifier and the published classes can be found in the Gaia archive vari_classifier_definition and vari_classifier_class_definition table, respectively.

The sky and magnitude distribution of the published classes, grouped by main type, is shown in the left panels of Figs. 3 and 5 and some time-series examples are shown in Figs. 4 and 6.

3 For tutorials on using this datalink, see the Gaia archive help webpage http://gea.esac.esa.int/archive-help/ Time-series bulk download is possible from http://cdn.gea.esac.esa.int/Gaia/
The published source samples were cleaned to reduce contamination levels caused by data and processing artefacts, as well as by genuine variability of other types. The results and details of this procedure are fully detailed in Rimoldini & et al. (2018).

The output of this classifier for sources containing at least 12 measurements in their $G$-band time series was used as one of the two inputs to the SOS of RR Lyrae stars, Cepheids, and LPV, as shown in Fig. 2 hence accounting for the overlap between classification and SOS results in Fig. 3. Some SOS modules may reclassify the type in the SOS tables, see Sect. 4.6. An overview of the inferred completeness and contamination with respect to several cross-match catalogues of the different types is provided in Table 2.

### 4.4. Specific Object Studies

In the geq20 path described in Sect. 3, potential variable type candidates are identified in the supervised classification and special variability detection (SVD) parts of the pipeline were analysed in SOS modules, which often derived a model and/or additional astrophysical properties. The SOS modules can also decide whether an object is not of the expected type and as a result not produce any output for it. We recall that a sub-sample of the sources processed in the geq2 path was processed by some SOS modules (see Fig. 2). We present short descriptions of the SOS modules used for DR2 below.

The sky and magnitude distribution of each published type is shown in the right panels of Figs. 4 and 5 and example light curves are presented in Figs. 6, 7, and 8. Because SOS modules can be fed by (partially) independent pipeline-paths, it is not excluded that a source is listed in multiple SOS tables, as discussed in Sect. 4.5 or has a different classification and SOS type, as discussed in Sect. 4.6. Despite extensive validation efforts, the nature of our processing is an automated and statistical characterisation of the sources, hence identified sources that are not yet known in the literature should be considered to be candidates, unless they are specifically validated in one of the SOS articles referenced below.

An overview of the inferred (absolute) sky completeness and contamination with respect to existing catalogues of the different types is provided in Table 2; see sections below and the references therein for more details. Many of the listed reference catalogues have a (very) limited sky coverage, and hence the contamination might be locally biased. To assess contamination in the whole SOS published samples, we performed the following additional visual inspection for each SOS output separately: we took a random sample of 500 sources from the published results, as well as a sky-uniform 6 sample of 500 sources (non-overlapping with the random sample). The former sample follows the general sky distribution as shown in Fig. 4, while the latter achieves an almost sky-uniform sampling and hence also draws samples from very low density regions. The results are summarised below for each SOS result:

- **Cepheids**: 5% of the random sample was visually rejected, of which 3% might be due to binaries or ellipsoidal. Most of this sample is located in the Magellanic Clouds. Of the sky-uniform sample, 15% was visually rejected, of which 11% might be due to binaries or ellipsoids.

6 Approximate sky-uniform sampling is achieved by first grouping the sources in 12288 bins (level 5 HEALPix, see Górski et al. 2005) of about 3.6 deg$^2$ and then randomly drawing non-empty pixels and extracting a source at random from it.

- **RR Lyrae stars**: 9% of the random sample was visually rejected, of which 4% are faint (20.0 < $G$ < 20.7 mag). Of the sky-uniform sample, 12% was visually rejected, of which 5% are faint (19.9 < $G$ < 20.7 mag).

- **Long-period variables**: 0.1% (1) was visually rejected; it looks like a young stellar object, confirmed by its position in the Hertzsprung-Russel (HR) diagram.

- **Short-timescale variability**: 0.3% was visually rejected, and about 24% seem to exhibit short-timescale periodic variability, but this can currently not be confirmed with a high confidence.

- **Rotation modulation**: 0.3% was visually rejected.

These identified contamination rates are compatible with the cross-match sample contamination rates found in Table 2.

### 4.4.1. Cepheids and RR Lyrae stars

The SOS Cepheids and RR Lyrae module (SOS Cep and RRL, Clementini et al. 2016, 2018), confirmed, characterised, and sub-classified in type 140 784 RR Lyrae stars and 9 575 Cepheids of the candidates provided by the classifiers. The module can re-classify the type provided by the classifiers from Cepheid to RR Lyrae star, and vice versa. Amongst the RR Lyrae stars, the following number of sub-types were identified: 98 026 fundamental mode pulsators (RRAb), 40 380 first-overtone pulsators (RRc), and 2378 double-mode pulsators (RRd). Amongst the Cepheids, there were 8890 δ Cepheids, 100 anomalous Cepheids, and 585 type II Cepheids, of which 223 were classified as BL Her, 253 as W Vir, and 109 as RV Tau.

A rough estimate of completeness and contamination was obtained by comparing our samples with OGLE catalogues of RR Lyrae stars and Cepheids in the Large Magellanic Cloud (LMC). We selected an area of the LMC centred at $\alpha = 83^\circ$ and $\delta = -67.5^\circ$ and extending from 67.5$^\circ$ to 97.5$^\circ$ in right ascension and from $-63.5^\circ$ to $-73^\circ$ in declination. Completeness of our RR Lyrae sample in this area is 64%, and contamination is less than 0.1%. In the same area, the Cepheid sample is 74% complete with 100% purity. A more in-depth description and analyses of the results, as well as the origin of the Bulge completeness rates quoted in Table 2 can be found in Clementini et al. (2018). Of the 599 Cepheids and 2595 RR Lyrae stars that were identified in DR1 (Clementini et al. 2016, Eyer et al. 2017), 533 Cepheids and 2517 RR Lyrae stars are also available in this DR2 release.

### 4.4.2. Long-period variables

Long-period variables are red giants that are characterised by long periods and variability amplitudes of up to a few magnitudes. They are identified by classification as MIRA_SR. On the basis of data available for DR2, CU7 selected 151 761 LPV candidates that fulfil the selection criteria for LPVs, which included a minimum of 12 points in $G$, a $G_{BP} - G_{RP}$ colour higher than 0.5 mag, a variability amplitude (quantified with the 5%-95% trimmed range in $G$) larger than 0.2 mag, and a correlation between $G$ and $G_{BP} - G_{RP}$ higher than 0.5. Of these, 150 757 are published in the classification table, and 89 617 have SOS results. The detailed distribution into these two tables is indicated in Fig. 4. The SOS LPV module processes LPV candidates when their period is longer than 60 days. The steps applied
Table 2: Classification (absolute) completeness and contamination estimates with respect to available cross-matched reference catalogues. For more details on the validation of the classification results, we refer to [Rimoldini & et al. (2018)]

<table>
<thead>
<tr>
<th>Group</th>
<th>Class (best_class_name)</th>
<th>Catalogue (and region)</th>
<th>Completeness</th>
<th>Contamination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cepheids</td>
<td>CEP, ACEP, T2CEP</td>
<td>OGLE-IV (LMC)</td>
<td>72%</td>
<td>&gt;5%</td>
</tr>
<tr>
<td>Cepheids</td>
<td>CEP, ACEP, T2CEP</td>
<td>OGLE-IV (SMC)</td>
<td>68%</td>
<td>&gt;5%</td>
</tr>
<tr>
<td>δ Scuti and SX Phoenicis</td>
<td>DSC_XSPHE</td>
<td>Catalina [Drake et al. 2014]</td>
<td>99%</td>
<td>&gt;13%</td>
</tr>
<tr>
<td>RR Lyrae stars</td>
<td>RRAB, RRC, RRD, ARRD</td>
<td>OGLE-IV (LMC)</td>
<td>60%</td>
<td>&gt;9%</td>
</tr>
<tr>
<td>RR Lyrae stars</td>
<td>RRAB, RRC, RRD, ARRD</td>
<td>OGLE-IV (SMC)</td>
<td>64%</td>
<td>15%</td>
</tr>
<tr>
<td>RR Lyrae stars</td>
<td>RRAB, RRC, RRD, ARRD</td>
<td>OGLE-IV (Bulge)</td>
<td>49%</td>
<td>44%</td>
</tr>
<tr>
<td>RR Lyrae stars</td>
<td>RRAB, RRC, RRD, ARRD</td>
<td>Catalina [Drake et al. 2014]</td>
<td>63%</td>
<td>&gt; 5%</td>
</tr>
<tr>
<td>LPV</td>
<td>MIRA_SR</td>
<td>OGLE-III (LMC+SMC)</td>
<td>40–50%</td>
<td>&lt; 5%</td>
</tr>
<tr>
<td>LPV</td>
<td>MIRA_SR</td>
<td>Gaia DR2 with $\sigma/\sigma_0 &gt; 10$</td>
<td>–</td>
<td>~ 7%</td>
</tr>
</tbody>
</table>

Notes. (a) Estimates are restricted to sources with peak-to-peak amplitude >0.2 mag. (b) $\sigma$ and $\sigma_0$ refer to the DR2 parallax and associated error.

Table 3: SOS (absolute) completeness and contamination estimates with respect to cross-matched reference catalogues. Processing paths are based on the number of selected FoV transits, as shown in Fig. 2 and discussed in Sect. 3. See Sect. 4.4 for additional contamination rate estimates for random and sky-uniform samples. For more details on the validation of the SOS type results, see Sect. 4.4 and the references therein.

<table>
<thead>
<tr>
<th>Processing path</th>
<th>SOS Type</th>
<th>Catalogue (and region)</th>
<th>Completeness</th>
<th>Contamination</th>
</tr>
</thead>
<tbody>
<tr>
<td>geq12 &amp; geq20</td>
<td>Cepheids</td>
<td>OGLE-IV (LMC, limited region)</td>
<td>74%</td>
<td>~ 5%$^a$</td>
</tr>
<tr>
<td>geq12 &amp; geq20</td>
<td>Cepheids</td>
<td>OGLE-IV (Bulge)</td>
<td>3%</td>
<td>$^b$</td>
</tr>
<tr>
<td>geq12 &amp; geq20</td>
<td>RR Lyrae stars</td>
<td>OGLE-IV (LMC, limited region)</td>
<td>64%</td>
<td>~ 9%$^a$</td>
</tr>
<tr>
<td>geq12 &amp; geq20</td>
<td>RR Lyrae stars</td>
<td>OGLE-IV (Bulge)</td>
<td>15%</td>
<td>$^b$</td>
</tr>
<tr>
<td>geq12 &amp; geq20</td>
<td>LPV</td>
<td>OGLE-III (SMC+LMC)</td>
<td>~ 30%</td>
<td>&lt; few%</td>
</tr>
<tr>
<td>geq12 &amp; geq20</td>
<td>LPV</td>
<td>OGLE-III (SMC+LMC)</td>
<td>~ 30%</td>
<td>&lt; few%</td>
</tr>
<tr>
<td>geq20</td>
<td>Short timescale</td>
<td>OGLE-II, III, IV (MCs)</td>
<td>0.05%</td>
<td>10–20%</td>
</tr>
<tr>
<td>geq20</td>
<td>Rotational modulation</td>
<td>The Pleiades [Hartman et al. 2010]</td>
<td>0.7 – 5.0%</td>
<td>&lt; 50%$^d$</td>
</tr>
</tbody>
</table>

Notes. (a) Estimated from non-cross-match random samples inspection discussed in Sect. 4.4, extending partially beyond OGLE-IV footprint. (b) No reliable contamination rate for the Bulge has been derived, but in the Galactic disk footprint (see [Udalski, 2017]), contamination of the SOS Cepheids and RR Lyrae stars might be of the order of several tens of percent (A. Udalski, priv. comm. based on unpublished OGLE-IV data). (c) Estimates are restricted to sources with peak-to-peak amplitude >0.2 mag and periods longer than 60 days. (d) The 50% ‘contamination’ quantifies the non-occurrence in the reference catalogue; the actual contamination of grazing binaries is estimated to be <1%, see Sect. 4.4.4.

4.4.3. Short-timescale variability

Short-timescale variable sources are astrophysical objects showing any variability in their optical light-curve with a characteristic timescale of the variation shorter than 1 d ([Roelens et al. (2017)]). A variety of astronomical objects are known to exhibit such fast photometric variations, including both periodic and transient variability, involving amplitudes from a few millimagnitudes to a few magnitudes, with different variability characteristics and phenomena at the origin of variation, from pulsations to flares to eclipsing systems. The diverse variability types targeted by the short-timescale module range from short-period binary stars to pulsating white dwarfs (e.g. ZZ Ceti stars or V777 Herculis stars) and cataclysmic variables such as AM CVn stars. However, for the Gaia DR 2, the short-timescale variability processing is oriented towards periodic variability, and the exact variable type of the detected short-timescale candidates is not yet determined. Both the investigation of transient variability ([Wevers et al. 2018]) and the further classification of the candidates are foreseen for the future Gaia data releases.

As mentioned previously, the identification of the short-timescale variable candidates published in Gaia DR 2 involves particular variability detection methods that are specifically tuned to detect fast photometric variability, complementarily to the general variability detection approach (which may miss some of these events because they are so fast). These methods are applied to a preselected subset of the Gaia data set ([Eyer et al. 2018]).

Note: the determination of bolometric correction from $G_{BP}$ and $G_{RP}$ fluxes, the identification of red supergiants, and the computation of absolute magnitude. It must be noted that all LPV specific attributes published in DR2 that depend on the parallax must be recomputed based on the published parallaxes, because these LPV-specific attributes were computed based on an older version of the parallaxes that was available at the time of variability processing within the consortium. We refer to [Mowlavi et al. 2018] for additional information. The light curves of all LPV candidates, whether identified in the classification table or in the SOS table, are published in DR2.

The completeness and contamination numbers for the LPV candidates towards the Magellanic Clouds are computed relative to Mowlavi et al. (2018) for additional information. The light curves of all LPV candidates, whether identified in the classification table or in the SOS table, are published in DR2.
4.4.4. Rotational modulation

Stellar flux modulation induced by surface inhomoogeneities and rotation is searched for in a region of the HR diagram, broadly embracing stars of spectral type later than F on the main sequence. This type of variability, often indicated in the literature as BY Draconis stars (BY Dra) regardless of possible binarity, is an indication of active regions (dark spots and bright faculae on the stellar photosphere) that are produced by stellar magnetic activity. The evolution of stellar magnetic fields, similar to solar fields, produces variability phenomena on a wide range of timescales (see e.g. Lanza et al. 2004, Distefano et al. 2012, Lanza et al. 2006 and references therein); one of the most prominent phenomena is the rotational modulation itself, from which the stellar rotation period can be inferred.

The variability pipeline comprises two packages dedicated to the detection and characterisation of solar-like stars with rotational modulation. First, solar-like variable candidates are selected, and if they are confirmed, they are studied in more detail to determine stellar rotation periods and other properties. Fig. 8 shows one of the published light-curves together with the segments in which a significant (similar) period was detected.

We detected some 7×10^5 periodic variables in the pre-selected HR diagram region, which excludes areas populated by pulsating stars. The remaining expected main contaminants are eclipsing binaries and spurious detections derived from an incomplete sampling. To filter out these cases as well as possible, we applied filters that exclude sources whose period-folded light curves have a significantly uneven distribution in phase, with significant gaps, and are far from being sinusoidal (see Lanza et al. 2018 for details). The final DR2 clean sample contains 147 535 rotational modulation candidates and fills 38% of the whole sky when divided into bins of ≈0.84 deg^2 (level 6 HEALPix), see for example Fig. 4.

An estimate of the final completeness and contamination of the DR2 rotational modulation variables is hampered by the fact that the occurrence rate of the BY Dra phenomenon is largely unknown so far. It is expected that all low-mass dwarfs are magnetically active to some extent, and the rotational modulation detectability essentially depends on the instrumental sensitivity, and also on the active region distribution and the phase of the magnetic cycle at the epoch of observations. New low-mass dwarfs displaying rotational modulation are continuously detected with the ever-wider span and increasing sensitivity of modern surveys, none of which has the full sky coverage capabilities of Gaia, however. At this stage, it is possible to perform some meaningful comparison only with the Hartman et al. (2010) and Rebull et al. (2016) observations of the Pleiades, with which, nevertheless, the DR2 geo20 sky coverage still overlaps only marginally. Assuming that the Hartman et al. (2010) catalogue lists all the BY Dra in its FoV down to G ≈ 14.5, we estimate that the completeness of the BY Dra sample is 14% down to G ≈ 16.5 in the overlapping field. Assuming that this value is uniform over the whole sky, we estimate a completeness upper limit of 5%. At the other extreme, we may assume that all low-mass dwarfs are BY Dra variables. Then, comparing with all stars observed by Gaia in the same sky region and magnitude range, we estimate a lower completeness limit of 0.7%.

We may roughly estimate an upper limit for contamination by assuming that the Gaia BY Dra detections that are not present in the Hartman et al. (2010) or in the Rebull et al. (2016) sample in the common region on the sky are variables of other types. This upper limit is obviously largely overestimated given that no BY Dra catalogue can be deemed complete to date, as also testified by the fact that there are sources in the Hartman et al. (2010) sample that are not contained in the Rebull et al. (2016) sample, and vice versa.

At the other extreme, we can make an educated guess regarding the contamination of close grazing binaries that would be incorrectly classified as BY Dra because of the similarity of the light curves. If the orbital period is long, then the grazing eclipse occurs only in a small part of the orbit, and it does not look like rotation modulation. We assume an upper limit period of 10 days for this contaminating effect. The fraction of stars with such short-period companions is only about 2–5% (Raghavan et al. 2010), of which only 2–5% are grazing binaries (at most), which means 0.04–0.25% of the stars, while we did not yet take into account that for most binaries, the secondaries are substantially smaller. Even though we cannot consider the rotation modulation sample a random subset of G and later-type stars (which is the assumption in above estimate), it seems unlikely that this type of contamination would exceed 1%. The same conclusion is reached from our internal validation against known and newly identified (but not published) eclipsing binaries, based on which, we estimate an upper limit of 0.5% grazing contamination. From the estimated grazing binary contamination and non-occurrence in the reference catalogue discussed above, we estimate a contamination level ranging from < 1% to ≈50%, as listed in Table 3. More details on the SVD solar-like and SOS rotational modulation packages and results can be found in Lanza et al. (2018).

4.5. Sources with multiple SOS types

There are 80 sources in DR2 with an entry in more than one SOS table. These cases are all overlaps between short-timescale objects and RR Lyrae stars (72), Cepheids (5), and rotation mod-
ulation objects (3), all of which are justified by their overlapping type definition. No overlap is found with long-period variables.

4.6. Sources with different classification and SOS type

Differences are found between classification types (as obtained by supervised classification in the geq2 path) and the types that are eventually derived in SOS results (either from the geq12 or geq20 paths). In particular, one source was classified as DSCT_SXPH, but also appears in SOS rotational modulation. Similarly, one source was classified as MIRA_SR, was not confirmed by SOS LPV, but appears in SOS rotational modulation. Of the sources classified as a Cepheid (any type), 6 appear in more than one SOS table. One is a confirmed SOS Cep and also appears in SOS short timescale, while 3 appear as SOS rotational modulation and 2 as SOS short timescale. Of the sources classified as an RR Lyrae star (any type), 70 appear in the SOS short timescale table, one was instead confirmed as a Cepheid by SOS (shown in the top panel of Fig. 7), and many were confirmed as RR Lyrae star by SOS. Overall, a reassignment of the type from RRL into Cep by SOS occurred in 618 cases, but only in 77 cases from Cepheid to RR Lyrae star by SOS. We have decided to retain the original classification type despite the possible reassignment by the SOS Cep&RRL module or in cases where they were confirmed as another type by SOS modules.

5. Conclusions

In this second Gaia data release, a sample of 550,737 variable sources with their three-band time series and analysis results has been released to the astronomical community, showcasing samples of high-amplitude variable candidates distributed over (a large portion of) the whole sky. It demonstrates the immense potential of Gaia data to provide an unbiased all-sky photometric survey together with astrometric and spectroscopic data. Although the variability analyses in DR2 did only mildly rely on the astrometric data and on none of the spectroscopic data, this dependence will become much heavier in future data releases.

Several caveats have been outlined in this paper:

- Most provided stellar samples are rather incomplete as a result of the current state of data calibration (e.g. even well-known variables might be missing) or the limited scanning law coverage for the minimum number of selected observations,
- Known incomplete flagging of outliers in the time series,
- Existence of uncalibrated systematics or spurious calibration error signals in some of the time series,
- The probabilistic and automated nature of this work implies certain completeness and contamination rates (Tables 2 and 3), hence even well-known literature sources could be misidentified.

Overall, we estimate to have identified the following numbers of new variables: a few hundred Cepheids, a few tens of thousand RR Lyrae stars, about one hundred thousand stars with rotation modulation, several tens of thousand long-period variables, a few thousand δ Scuti and SX Phoenicis stars, and a few thousand short-timescale variables. In total, this is about half of the variable sources released in this DR2. Compared to the DR1 example, we increased the sample size about 100 times, which is expected to result in a large number of novel studies, regardless of whether they are combined with existing data sets.

The next variability release will be based on data that are better calibrated yet again and on a longer time baseline. It will also include BP/RP spectrophotometry and radial velocities from RVs, which will allow us to improve the quality and quantity of the Gaia DR3 release.

Acknowledgements.

We dedicate this paper to our dear friend and colleague Jan Cuypers (1956-2017), whose passing has left a great void in our variability group. Thank you, Jan, for your friendship and lifelong contribution to science. We would like to thank the referee, Andrej Udalski, for various suggestions and providing so far unpublished OGLE-IV estimates that improved the contents of this paper. We thank Xavier Luri, Antonella Valmari, and Annie Robin for valuable feedback during the preparation of this paper. We also acknowledge Mark Taylor for creating and implementing new requested features in the astronomy-oriented data handling and visualization software TOPCAT (Taylor 2005). This work has made use of data from the ESA space mission Gaia, processed by the Gaia Data Processing and Analysis Consortium (DPAC).

Funding for the DPAC has been provided by national institutions, some of which participate in the Gaia Multilateral Agreement, which include, for Switzerland, the Swiss State Secretariat for Education, Research and Innovation and through the ESA PRODEX program, the "Mesures d’accompagnement", the "Activités Nationales Complémentaires", the Swiss National Science Foundation, and the Early Postdoc. Mobility fellowship. Belgium, the BE,Glavian federal Science Policy Office (BELSPO) through PRODEX grants; for Italy, the Italian National Supercomputers (INAF) and the Agenzia Spaziale Italiana (ASI) through grants I/037/08/0, I/038/10, 2014-025-R.0, and 2014-025-R.1 to INAF (PI M.G. Lattanzi); for France, the Centre National d’Études Spatiales (CNES). This research has received additional funding from the Israel Centers for Research Excellence (I-CORE, grant No. 1829/12); the Hungarian Academy of Sciences through the Lendület Programme LP2014-17 and the János Bolyai Research Scholarship (L. Molnár and E. Plachy); the Hungarian National Research, Development, and Innovation Office through grants OTKA/NKFIH K-113117, K-115709, PD-116175, and PD-121203; the UNKP-17-3 program of the Ministry of Human Capacities of Hungary (Á.L.); the ESA Contract No. 400012377/17/NL/CBI; and the Polish National Science Center through ETIUDA grant 2016/20/T/ST9/00170; Charles University in Prague through PRIMUS/SCI/17 award; and from the European Research Council (ERC) under ERC StG 2014 Uni’s Horizon 2020 research and innovation programme (grant agreement No 760519: MAMSIE).

References

Lanza, A. F., Messina, S., Pagano, I., & Rodonó, M. 2006, Astronomische Nachrichten, 327, 21

Article number, page 8 of 22
Table 4: Statistical parameters for the $n$Transits:2+ classification types published in DR2. The 1%, 50% (median), and 99% percentile are provided for each type.

<table>
<thead>
<tr>
<th>$G$ band</th>
<th>Filtered number of FoV transits</th>
<th>RR Lyrae stars</th>
<th>Long-period var.</th>
<th>$\delta$ Sct and SX Phe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13</td>
<td>27</td>
<td>513.15</td>
<td>2018.52</td>
</tr>
<tr>
<td>Time duration (d)</td>
<td>15.61</td>
<td>17.60</td>
<td>13.35</td>
<td>17.75</td>
</tr>
<tr>
<td>Mean number of magnitude</td>
<td>7.48</td>
<td>15.91</td>
<td>13.81</td>
<td>18.15</td>
</tr>
<tr>
<td>Range magnitude</td>
<td>0.19</td>
<td>0.56</td>
<td>0.30</td>
<td>0.74</td>
</tr>
<tr>
<td>Standard deviation (mag)</td>
<td>0.06</td>
<td>0.18</td>
<td>0.10</td>
<td>0.23</td>
</tr>
<tr>
<td>Skewness of magnitudes</td>
<td>-1.68</td>
<td>-0.27</td>
<td>-2.14</td>
<td>-0.53</td>
</tr>
<tr>
<td>Kurtosis of magnitudes</td>
<td>-1.88</td>
<td>-1.12</td>
<td>-2.04</td>
<td>-0.90</td>
</tr>
<tr>
<td>MAD (mag)</td>
<td>0.05</td>
<td>0.18</td>
<td>0.06</td>
<td>0.20</td>
</tr>
<tr>
<td>Abbe</td>
<td>0.12</td>
<td>0.65</td>
<td>0.49</td>
<td>0.97</td>
</tr>
<tr>
<td>IQR (mag)</td>
<td>0.09</td>
<td>0.28</td>
<td>0.11</td>
<td>0.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$G_{BP}$ band</th>
<th>Filtered number of FoV transits</th>
<th>RR Lyrae stars</th>
<th>Long-period var.</th>
<th>$\delta$ Sct and SX Phe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time duration (d)</td>
<td>15.74</td>
<td>17.53</td>
<td>12.97</td>
<td>17.39</td>
</tr>
<tr>
<td>Mean number of magnitude</td>
<td>8.67</td>
<td>16.52</td>
<td>14.39</td>
<td>18.89</td>
</tr>
<tr>
<td>Range number of magnitude</td>
<td>0.29</td>
<td>0.79</td>
<td>0.29</td>
<td>1.16</td>
</tr>
<tr>
<td>Standard deviation (mag)</td>
<td>0.08</td>
<td>0.23</td>
<td>0.12</td>
<td>0.45</td>
</tr>
<tr>
<td>Skewness of magnitudes</td>
<td>-1.82</td>
<td>-0.86</td>
<td>-2.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Kurtosis of magnitudes</td>
<td>-1.82</td>
<td>-0.86</td>
<td>-2.01</td>
<td>0.02</td>
</tr>
<tr>
<td>MAD (mag)</td>
<td>0.06</td>
<td>0.22</td>
<td>0.05</td>
<td>0.29</td>
</tr>
<tr>
<td>Abbe</td>
<td>0.14</td>
<td>0.69</td>
<td>0.41</td>
<td>0.99</td>
</tr>
<tr>
<td>IQR (mag)</td>
<td>0.10</td>
<td>0.34</td>
<td>0.13</td>
<td>0.46</td>
</tr>
</tbody>
</table>

**Notes:**
- Cepheid groups: best_class_name = CEP, Acep, T2CEP
- RR Lyrae star groups: best_class_name = RRAB, RRC, RRc, ARRD
- Long-period var. is best_class_name = MIRA_SR
- $\delta$ Sct and SX Phe is: best_class_name = DSCT_SXPH

For the skewness and kurtosis, we included only stars with > 2 and > 3 FoV transits, respectively.
Table 5: Statistical parameters for the SOS types published in DR2. The 1%, 50% (median), and 99% percentile are provided for each type.

<table>
<thead>
<tr>
<th>Type</th>
<th>Filtered number of FoV transits</th>
<th>Time duration (d)</th>
<th>Mean observation time (d)</th>
<th>Min. magnitude</th>
<th>Mean magnitude</th>
<th>Max. magnitude</th>
<th>Standard deviation (mag)</th>
<th>Skewness of magnitudes</th>
<th>IQR (mag)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G band</td>
<td>12</td>
<td>118</td>
<td>27</td>
<td>28</td>
<td>667</td>
<td>70</td>
<td>64</td>
<td>43,010</td>
<td>593,78</td>
</tr>
<tr>
<td>RR Lyrae stars</td>
<td>13</td>
<td>13</td>
<td>27</td>
<td>28</td>
<td>64</td>
<td>70</td>
<td>64</td>
<td>43,010</td>
<td>593,78</td>
</tr>
</tbody>
</table>

Notes: For RR Lyrae stars, the peculiarities for G and R bands were computed for a number of FoV transits > 1, although there are stars with 0 or 1 transit remaining in the band after our cleaning.
Fig. 1: Relative source counts with respect to sources having ≥2 G-band FoV transits, illustrating the drop in relative completeness for source-sets with ≥12 and ≥20 G-band FoV transits, as applied in our processing (see Fig. 2). Regional values are listed in Table 1 and discussed in Sect. 2.1. Values for external catalogs are Gaia source counts in the approximate sky footprint, not source cross-match counts. Note that actual time series are published only for the 550K variable sources in DR2 (see Fig. 3).
Fig. 2: DR2 variability processing overview. Data were published from the highlighted yellow boxes for the validation filtered sources. There were two main tracks: one starting from \( \geq 2 \) G-band FoV transits (left), and the other from \( \geq 20 \) G-band FoV transits (right). The former resulted in the published \( n_{\text{Transits}:2^+} \) classifier in the vari_classifier_result table, and the latter resulted in the published SOS tables of vari_short_timescale and vari_rotation_modulation. The published SOS tables of vari_rrlyrae, vari_cepheid and vari_long_period_variable result from a mixed feed of the published \( n_{\text{Transits}:2^+} \) classifier (for sources with at least \( \geq 12 \) G-band FoV transits) and from the unpublished classifier of the \( \geq 20 \) track.

Fig. 3: Accounting of the 550 737 unique variable sources published in DR2. All have an entry in the Gaia archive gaia_source and vari_time_series_statistics tables. Their distributions in the additional variability tables is shown. The classification output types are mutually exclusive so that their sum matches the total number of entries in the published vari_classifier_result table. Eighty sources appear in more than one SOS table (not illustrated), which is detailed in Sect. 4.5. The overlap between similar-type classification and SOS is detailed in a Venn-like fashion showing the counts in the three possible subsets. A few hundred sources have different assigned types in classification and SOS tables (not illustrated), as explained in Sect. 4.6, hence the difference on the left between the total type count and the unique source_id count. No SOS module treated \( \delta \) Scuti & SX Phoenicis stars.
Fig. 4: Sky source densities [count deg$^{-2}$] in Galactic coordinates of the published sources in the classification table (left column, see Sect. 4.3) and SOS tables (right column, see Sect. 4.4). In the classification plots the best_class_name entries were grouped by main type, as listed in parentheses. Galactic longitude increases to the left side.

Article number, page 14 of 22
Fig. 5: Histogram counts of the median G-band magnitude distribution of the published classification classes (left, see Sect. 4.3) and SOS tables (right, see Sect. 4.4). In the classification plots the best_class_name entries were grouped by main type, as listed in parentheses. The filled bars correspond to the linear scale on the left side of the plots (which varies in scale), while the dotted line corresponds to the logarithmic scale on the right side of the plots (always in the same scale). The bin size is fixed to 0.5 mag.
Fig. 6: Example light-curves of RR Lyrae stars and LPV in the data. For each band we show the valid FoV transits, as well as variability-rejected FoV transits (number in parentheses, data plotted with crosses). In most cases, two FoV-transits occur within the 6 h spin period, causing indistinguishable overlapping points in the left diagrams. Some outliers are not flagged in the exported data (see e.g. the $G_{BP}$ light curve of the RRc and RRd example); these transits were removed in the more strict outlier rejection used in the SOS Cep and RRL module, as mentioned in Sect. 4.1.
Therefore the period for the rejection used in the SOS Cep and RRL module, as mentioned in Sect. 4.1. No period is published for the classification results.

Fig. 7: Example light-curves of the Cepheid, δ Scuti, and short-timescale variables. For each band we show the valid FoV transits, as well as variability-rejected FoV transits (number in parentheses, data plotted with crosses). In most cases, two FoV-transits occur within the 6 h spin period, causing indistinguishable overlapping points in the left diagrams. Some outliers are not flagged in the exported data (see e.g. the G_BP and G_RP light curve of the ACEP example); these transits were removed in the more strict outlier rejection used in the SOS Cep and RRL module, as mentioned in Sect. 4.1. No period is published for the classification results, therefore the period for the DSCT_SXPHE example is taken from the LINEAR survey (Palaversa et al. 2013).
Fig. 8: Example light-curves of a solar-like star with rotation modulation. Period search is performed over multiple segments, as shown in the right panels; see Sect. 4.4.4 for more details. The bottom left panel expands a high-cadence time range of the time series. The top left panel shows for each band the valid FoV transits and variability-rejected FoV transits (number in parentheses, here 0).
Fig. 9: Selected (i.e. not-rejected) number of transits in the time series of the $G$, $G_{BP}$, and $G_{RP}$ photometric bands in Galactic coordinates for all 550,737 published sources. These numbers have been aggregated in sky bins of about 0.84 deg$^2$ (level 6 HEALPix) and their mean value is shown. The colour scale is clipped at 100, although the number of transits reaches up to 242, as shown in Fig. 10. The choice of the rainbow colour scale has been motivated by the convenient colour divisions between 5, 10, 20, 30, 40, 50, and 100 transits. Galactic longitude increases to the left side.

Fig. 10: Histogram of selected (i.e. not-rejected) number of transits in time series of the $G$, $G_{BP}$, and $G_{RP}$ photometric bands for the 550,737 published variable sources. The filled bars correspond to the linear scale on the left side, while the solid lines correspond to the logarithmic scale on the right side of the plot. The bin size is 1 FoV transit.
Appendix A: Examples of Gaia archive queries

This section describes Gaia archive queries in the ADQL format that return the various selections presented in this paper. These queries can be made online in the Gaia archive at [http://gea.esac.esa.int/archive/](http://gea.esac.esa.int/archive/).

Appendix A.1: Total counts and table merging

Figure 3 total number of unique sources with variability results (bottom left unique source_id count):

```sql
select count(*) from gaiadr2.gaia_source where phot_variable_flag='VARIABLE'
```

Retrieve various astrometric and general statistical variability fields for all published variable sources in Gaia DR2:

```sql
select gs.source_id, ra, dec, parallax, parallax_error, pmra, pmdec, epoch_photometry_url,
num_selected_g_fov, mean_mag_g_fov, range_mag_g_fov, std_dev_mag_g_fov, time_duration_g_fov,
num_selected_bp, mean_mag_bp, range_mag_bp, std_dev_mag_bp, time_duration_bp,
num_selected_rp, mean_mag_rp, range_mag_rp, std_dev_mag_rp, time_duration_rp
from gaiadr2.gaia_source as gs join gaiadr2.vari_time_series_statistics as varistats using(source_id)
where phot_variable_flag='VARIABLE'
```

Appendix A.2: Creating a variable summary table

It is very desirable to have an overview of the occurrence of a given source_id in any of the vari_* tables. The following query generates such a table in which each row contains a unique variable source source_id, and each column represents one of the vari-tables. A column contains integer 1 when it contains the source_id and NULL when not. After logging into the Gaia archive a user table can be created from the query result (suggested name: gaiadr2_vari_summary), after which it appears in the user tables and can be queried like any other DR2 table.

```sql
select
gs.source_id,
,vrrl.incl as vari_rrlyrae -- 1 if RR Lyrae SOS table entry
,vcep.incl as vari_cepheid -- 1 if Cepheid SOS table entry
,vmrm.incl as vari_rotation_modulation -- 1 if rotation modulation SOS table entry
,vsts.incl as vari_short_timescale -- 1 if short timescale SOS table entry
,vlpv.incl as vari_long_period_variable -- 1 if long period variable SOS table entry
,vcr_rrl.incl as vari_classifier_result_rrl -- 1 if classif. entry of any RR Lyrae type
,vcr_cpep.incl as vari_classifier_result_cpep -- 1 if classif. entry of any Cepheid type
,vcr_dscsxp.incl as vari_classifier_result_dsct_sxphe -- 1 if classif. entry of Delta Scuti
,vcmira_sr.incl as vari_classifier_result_mira_sr -- 1 if classif. entry of Mira or semi-regular type
,1 as vari_time_series_statistics -- 1 if time series statistics entry (1 for all in table)
,1 as gaia_source_variable_flag -- 1 if gaia_source phot_variable_flag='VARIABLE' (1 for all in table)
from gaiadr2.gaia_source as gs
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_rrlyrae) as vrrl using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_cepheid) as vcep using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_rotation_modulation) as vrmr using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_short_timescale) as vsts using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_long_period_variable) as vlpv using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_classifier_result where best_class_name='RRAB' or best_class_name='RRC' or best_class_name='RRD' or best_class_name='ARRD') as vcr_rrl using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_classifier_result where best_class_name='CEP' or best_class_name='ACEP' or best_class_name='T2CEP') as vcr_cpep using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_classifier_result where best_class_name='DSCT_SXPHE') as vcr_dscsxp using(source_id)
LEFT OUTER join (select source_id, 1 as incl from gaiadr2.vari_classifier_result where best_class_name='MIRA_SR') as vcmira_sr using(source_id)
where gs.phot_variable_flag='VARIABLE'
order by gs.source_id
```

Appendix A.3: HEALPix grouped data

To reproduce sky-distribution data sets using HEALPix grouping as shown in Figures 4 and 10, the convenient gaia_healpix_index(norder, source_id) function in the archive that extracts the HEALPix pixel id that is encoded in the Gaia source_id can be used. Using norder=6 creates pixels of about 0.84 deg^2. A program like TOPCAT (Taylor 2003) can read and visualise such data set with ease.
Figure 4 top left panel ‘Classif.: RR Lyrae stars (RRAB, RRC, RRD, ARRD)’ number of sources per square degree, and also the min, average, max, and standard deviation of the vari\_classifier\_result.best\_class\_score value per HEALPix pixel:

```sql
SELECT  
  gaia\_healpix\_index(6, source\_id) AS healpix\_6,  
  count(*) / 0.83929 as sources\_per\_sq\_deg,  
  min(best\_class\_score) AS min\_best\_class\_score,  
  avg(best\_class\_score) AS avg\_best\_class\_score,  
  max(best\_class\_score) AS max\_best\_class\_score,  
  stddev(best\_class\_score) as stddec\_best\_class\_score  
FROM gaiadr2.vari\_classifier\_result WHERE  
  best\_class\_name='RRAB' or best\_class\_name='RRC' or best\_class\_name='RRD' or best\_class\_name='ARRD'  
GROUP BY healpix\_6
```

Figure 4 top right panel ‘SOS: RR Lyrae stars’: number of sources per square degree, and also the min, average, max, and standard deviation of the vari\_rrlyrae.int\_average\_g value per HEALPix pixel:

```sql
SELECT  
  gaia\_healpix\_index(6, source\_id) AS healpix\_6,  
  count(*) / 0.83929 as sources\_per\_sq\_deg,  
  min(int\_average\_g) AS min\_int\_average\_g,  
  avg(int\_average\_g) AS avg\_int\_average\_g,  
  max(int\_average\_g) AS max\_int\_average\_g,  
  stddev(int\_average\_g) as stddec\_int\_average\_g  
FROM gaiadr2.vari\_rrlyrae  
GROUP BY healpix\_6
```

Figure 10 all three panels (‘number of transits in time series of the G, G\_BP, and G\_RP photometric bands’) average, and also the min, max, and standard deviation of the values per HEALPix pixel:

```sql
SELECT  
  gaia\_healpix\_index(6, source\_id) AS healpix\_6,  
  count(*) / 0.83929 as sources\_per\_sq\_deg,  
  -- G-band FoV  
  avg(num\_selected\_g\_fov) AS avg\_num\_selected\_g\_fov,  
  min(num\_selected\_g\_fov) AS min\_num\_selected\_g\_fov,  
  max(num\_selected\_g\_fov) AS max\_num\_selected\_g\_fov,  
  stddev(num\_selected\_g\_fov) as stddec\_num\_selected\_g\_fov,  
  -- BP-band  
  avg(num\_selected\_bp) AS avg\_num\_selected\_bp,  
  min(num\_selected\_bp) AS min\_num\_selected\_bp,  
  max(num\_selected\_bp) AS max\_num\_selected\_bp,  
  stddev(num\_selected\_bp) as stddec\_num\_selected\_bp,  
  -- RP-band  
  avg(num\_selected\_rp) AS avg\_num\_selected\_rp,  
  min(num\_selected\_rp) AS min\_num\_selected\_rp,  
  max(num\_selected\_rp) AS max\_num\_selected\_rp,  
  stddev(num\_selected\_rp) as stddec\_num\_selected\_rp  
FROM gaiadr2.vari\_time\_series\_statistics  
GROUP BY healpix\_6
```

Appendix A.4: Detailed counting

This section shows how the counts presented in Figures 3 and 5 can be reproduced from the published DR2 tables. Alternatively, a summary table can first be created as described in appendix A.2 and the relevant selections on the columns involved (not shown) can be made.

Figures 3 and 5 total number of sources with classification: all / δ Scuti and SX Phoenicis / RR Lyrae stars / Cepheids / LPV (Mira and semi-regular):

```sql
SELECT count(*) from gaiadr2.vari\_classifier\_result  
SELECT count(*) from gaiadr2.vari\_classifier\_result WHERE best\_class\_name='DSCT\_SXPHE'  
SELECT count(*) from gaiadr2.vari\_classifier\_result WHERE best\_class\_name='RRAB' or best\_class\_name='RRC' or best\_class\_name='RRD' or best\_class\_name='ARRD'  
SELECT count(*) from gaiadr2.vari\_classifier\_result WHERE best\_class\_name='CEP' or best\_class\_name='ACEP' or best\_class\_name='T2CEP'  
SELECT count(*) from gaiadr2.vari\_classifier\_result WHERE best\_class\_name='MIRA\_SR'  
```

Figures 3 and 5 total number of sources with SOS: RR Lyrae / Cepheid / LPV / short timescale / rotation modulation results:
select count(*) from gaiadr2.vari_rrlyrae
select count(*) from gaiadr2.vari_cepheid
select count(*) from gaiadr2.vari_long_period_variable
select count(*) from gaiadr2.vari_short_timescale
select count(*) from gaiadr2.vari_rotation_modulation

Figure 3 number of sources in common between classification and SOS for: RR Lyrae / Cepheid / LPV in Gaia DR2, these are the numbers in the overlap blocks of the Venn-diagrams:

select count(*) from gaiadr2.vari_classifier_result as clasrrl
INNER join gaiadr2.vari_rrlyrae as sosrrl on clasrrl.source_id=sosrrl.source_id
where best_class_name='RRAB' or best_class_name='RRC' or best_class_name='RRD' or best_class_name='ARRD'
select count(*) from gaiadr2.vari_classifier_result as clascep
INNER join gaiadr2.vari_cepheid as soscep on clascep.source_id=soscep.source_id
where best_class_name='CEP' or best_class_name='ACEP' or best_class_name='T2CEP'
select count(*) from gaiadr2.vari_classifier_result as claslpv
INNER join gaiadr2.vari_long_period_variable as soslpv on claslpv.source_id=soslpv.source_id
where best_class_name='MIRA_SR'

Figure 3 number of sources that are only found in classification and not in SOS for RR Lyrae / Cepheid / LPV in Gaia DR2, these are the numbers in the left blocks of the Venn-diagrams:

select count(*) from gaiadr2.vari_classifier_result as clasrrl
LEFT OUTER join gaiadr2.vari_rrlyrae as sosrrl on clasrrl.source_id=sosrrl.source_id
where (best_class_name='RRAB' or best_class_name='RRC' or best_class_name='RRD' or best_class_name='ARRD') and sosrrl.source_id is NULL
select count(*) from gaiadr2.vari_classifier_result as clascep
LEFT OUTER join gaiadr2.vari_cepheid as soscep on clascep.source_id=soscep.source_id
where (best_class_name='CEP' or best_class_name='ACEP' or best_class_name='T2CEP') and soscep.source_id is NULL
select count(*) from gaiadr2.vari_classifier_result as claslpv
LEFT OUTER join gaiadr2.vari_long_period_variable as soslpv on claslpv.source_id=soslpv.source_id
where best_class_name='MIRA_SR' and soslpv.source_id is NULL

Figure 3 number of sources that are only found in SOS and not in classification for: RR Lyrae / Cepheid / LPV in Gaia DR2, these are the numbers in the right blocks of the Venn-diagrams:

select count(*) from gaiadr2.vari_classifier_result as clasrrl
RIGHT OUTER join gaiadr2.vari_rrlyrae as sosrrl on clasrrl.source_id=sosrrl.source_id
where (best_class_name='RRAB' and best_class_name='RRC' and best_class_name='RRD') or clasrrl.source_id is NULL
select count(*) from gaiadr2.vari_classifier_result as clascep
RIGHT OUTER join gaiadr2.vari_cepheid as soscep on clascep.source_id=soscep.source_id
where (best_class_name='CEP' and best_class_name='ACEP' and best_class_name='T2CEP') or clascep.source_id is NULL
select count(*) from gaiadr2.vari_classifier_result as claslpv
RIGHT OUTER join gaiadr2.vari_long_period_variable as soslpv on claslpv.source_id=soslpv.source_id
where best_class_name='MIRA_SR' or claslpv.source_id is NULL