A&A 648, A44 (2021)
https://doi.org/10.1051/0004-6361/202039450

Astronomy & Astrophysics

Large-amplitude variables in *Gaia* Data Release 2

Multi-band variability characterization*

N. Mowlavi[1,2], L. Rimoldini[2], D. W. Evans[3], M. Riello[3], F. De Angeli[3], L. Palaversa[4,3], M. Audard[1,2], L. Eyer[1,2], P. Garcia-Lario[5], P. Gavras[5], B. Holl[1,2], G. Jevardat de Fombelle[2,6], I. Lecœur-Taïbi[2], and K. Nienartowicz[2,7]

[1] Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland
e-mail: Nami.Mowlavi@unige.ch
[2] Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland
[3] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[4] Ruđer Bošković Institute, Bijenička Cesta 54, 10000 Zagreb, Croatia
[5] European Space Astronomy Centre (ESA/ESAC), Villanueva de la Canada, 28692 Madrid, Spain
[6] SixSq, Route de Meyrin 267, 1217 Meyrin, Switzerland
[7] Sednai Sarl, 1204 Geneva, Switzerland

Received 16 September 2020 / Accepted 20 December 2020

ABSTRACT

Context. Photometric variability is an essential feature that sheds light on the intrinsic properties of celestial variable sources, the more so when photometry is available in various bands. In this respect, the all-sky *Gaia* mission is particularly attractive as it collects, among other quantities, epoch photometry measured quasi-simultaneously in three optical bands for sources ranging from a few magnitudes to fainter than magnitude 20.

Aims. The second data release (DR2) of the mission provides mean G, $G_{\rm BP}$, and $G_{\rm RP}$ photometry for ∼1.4 billion sources, but light curves and variability properties are available for only ∼0.5 million of them. Here, we provide a census of large-amplitude variables (LAVs) with amplitudes larger than ∼0.2 mag in the G band for objects with mean brightnesses between 5.5 and 19 mag.

Methods. To achieve this, we rely on variability amplitude proxies in G, $G_{\rm BP}$, and $G_{\rm RP}$ computed from the uncertainties on the magnitudes published in DR2. We then apply successive filters to identify two subsets containing sources with reliable mean $G_{\rm BP}$ and $G_{\rm RP}$ (for studies using colours) and sources having compatible amplitude proxies in G, $G_{\rm BP}$, and $G_{\rm RP}$ (for multi-band variability studies).

Results. The full catalogue gathers 23 315 874 LAV candidates, and the two subsets with increased levels of purity contain, respectively, 1 148 861 and 618 966 sources. A multi-band variability analysis of the catalogue shows that different types of variable stars can be categorized according to their colours and blue-to-red amplitude ratios as determined from the G, $G_{\rm BP}$, and $G_{\rm RP}$ amplitude proxies. More specifically, four groups are globally identified. They include: long-period variables in a first group with amplitudes more than twice larger in the blue than in the red; hot compact variables in a second group with amplitudes smaller in the blue than in the red; classical instability strip pulsators in a third group with amplitudes larger in the blue than in the red by 50% to 80%; and other non-pulsating variables in a fourth group, mainly achromatic, but 10% of them still having 20% to 50% larger amplitudes in the blue than in the red.

Conclusions. The catalogue constitutes the first census of *Gaia* LAV candidates extracted from the public DR2 archive. The overview presented here illustrates the added value of the mission for multi-band variability studies, even at this stage when epoch photometry is not yet available for all sources.

Key words. stars: variables: general – stars: general – surveys – methods: data analysis

1. Introduction

Since the end of the 20th century, the number of known variable stars has dramatically increased, boosted by the operation of large-scale surveys in the search for dark matter, such as the MACHO (Alcock et al. 1997), EROS (Palanque-Delabrouille et al. 1998), and OGLE (Udalski et al. 1997) surveys. In the last few years, the Catalina survey reached ∼110 000 variables (Drake et al. 2017), Pan-STARRS ∼240 000 variables (Sesar et al. 2017), ATLAS ∼430 000 variables (Heinze et al. 2018), *Gaia* ∼500 000 variables (Holl et al. 2018), ASAS-SN ∼220 000 variables (Jayasinghe et al. 2020), ZTF ∼600 000 variables (Chen et al. 2020), OGLE-IV ∼1 000 000 variables[1], and the American Association of Variable Star Observers (AAVSO) lists ∼1 500 000 variables as of June 2020[2].

The *Gaia* mission offers a unique opportunity in this field. It provides astrometry, photometry, and spectro-photometry for stars all over the sky, in the wide brightness range of a few magnitudes to above 20 mag, as well as spectroscopy for the bright objects (Gaia Collaboration 2016). And, for multi-band variability studies, the mission is unique because of the

* The catalogue is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/648/A44

[1] http://ogledb.astrouw.edu.pl/~ogle/OCVS/
[2] https://www.aavso.org

availability of quasi-simultaneous photometric measurements in three bands (G, G_{BP}, and G_{RP} within 50 s, 100 s if including the radial velocity spectrometer RVS).

In *Gaia* data release 2 (DR2, Gaia Collaboration 2018a), variability amplitudes measured from epoch photometry are provided for a subset of ∼500 000 variable stars of specific variability types. This represents only a small fraction of the variables present in the public *Gaia* archive. For all other sources not included in these ∼500 000 variables, and which hence do not have published photometric time series, their variability amplitude can still be estimated using the published photometric uncertainties. This is due to the fact that these uncertainties are derived from the standard deviation of the light curves and hence include information about both measurement uncertainties and source variability. We have taken advantage of this feature to build a multi-band variability catalogue of large-amplitude (≳0.2 mag) variables (LAVs) for all objects published in *Gaia* DR2.

The variability amplitude proxies used to estimate the amplitudes in G, G_{BP}, and G_{RP} are introduced in Sect. 2. Our catalogue of *Gaia* DR2 LAVs is then presented in Sect. 3, in which three datasets (Datasets A, B, and C) are identified for different purposes. The quality of the catalogue in terms of both completeness and purity is also addressed in that section. Section 4 then illustrates the usage of the catalogue with two examples. The first demonstrates an application of the mutli-band variability amplitudes to identify different categories of variable stars, while the second presents the sample of LAVs with good parallaxes. The main body of the text ends with a summary and concluding remarks in Sect. 5.

Additional material is presented in several appendices. The extraction of LAVs from the *Gaia* archive and the removal of outliers is detailed in Appendix A. The sum of the fluxes in the blue (BP) and red (RP) spectrophotometers is compared to the flux in the main G band in Appendix B, knowing that the summed transmission curve of the two spectrophotometers is close to the transmission curve of G. An amplitude proxy for BP + RP and its relation with the individual amplitude proxies for G_{BP} and G_{RP} are derived in Appendix C. Finally, the electronic table of our *Gaia* DR2 LAV catalogue is described in Appendix D.

The notations used in this paper regarding *Gaia* fluxes and magnitudes comply with the notations adopted in Evans et al. (2018) (see also Busso et al. 2018, Sect. 5.3.5): f_G represents the epoch flux of one CCD photometric measurement in the astrometric focal plane, and f_{BP} and f_{RP} represent the wavelength-integrated epoch flux during one transit in the blue and red spectrophotometric focal planes, respectively; I_G, I_{BP}, and I_{RP} represent the (inverse-variance weighted) mean fluxes in the respective photometric bands of a given source over the 22 months of data gathered in DR2. Finally G, G_{BP}, and G_{RP} are the mean magnitudes derived from I_G, I_{BP}, and I_{RP}, respectively. Epoch magnitudes per se are not used in this paper, but when we mention it, we notated it $G(t)$.

2. The variability amplitude proxy

2.1. Definitions

For constant stars, the uncertainty $\varepsilon(I)$ on the weighted mean flux I can be estimated from the variance σ_f^2 of the N flux measurements f using $\varepsilon^2(I) = \sigma_f^2 \left(\sum_i^N w_i^2\right) / \left(\sum_i^N w_i\right)^2$, where w_i denotes the weight associated with the ith measurement[3]. Since flux weights are not published in *Gaia* DR2, we estimate σ_f assuming equally weighted measurements: $\sigma_f = \varepsilon(I)\sqrt{N}$ (this form might overestimate σ_f as the effective number of measurements is less than N if weights are unequal). To obtain a quantity independent of the flux (which depends on various parameters such as the integration time), it is convenient to express σ_f relative to the mean flux, σ_f/I. This ratio is also proportional to the standard deviation σ_m in magnitude. For $\sigma_f/I \ll 1$, we have

$$\sigma_m \approx \frac{2.5}{\ln(10)} \frac{\sigma_f}{I} \approx 1.09 \frac{\varepsilon(I)}{I} \sqrt{N}. \quad (1)$$

The value of σ_m computed from σ_f/I in this way may be underestimated for large $\varepsilon(I)/I$ variations because of the non-linear relation between flux and magnitude. In practice, however, the approximation turns out to be sufficiently accurate for our purposes, even for the large variability amplitudes considered here (see Sect. 2.2).

In *Gaia* DR2, the published mean flux uncertainty $\varepsilon(I_G)$ of a source, whether constant or variable, is computed from the standard deviation of its f_G flux curve. Therefore, based on Eq. (1), the quantity

$$A_{\mathrm{proxy},G} = \sqrt{N_G} \frac{\varepsilon(I_G)}{I_G} \quad (2)$$

can be used as a proxy for the scatter in G light curves. For constant stars, it approximates the standard deviation of G light curves due to noise and uncalibrated systematic effects (see Sect. 5.3.5 of the *Gaia* DR2 documentation in Busso et al. 2018), to a factor of 1.09 (from Eq. (1)). For variable stars, the standard deviation is larger than it would be if the star was constant because of the additional contribution from stellar variability. Therefore, the amplitude proxy reflects the variability amplitude of astrophysical origin if the latter dominates the variability recorded in the signal.

Equation (2) has already been applied to both DR1 and DR2 for the study of specific types of variable stars such as Miras in the Magellanic Clouds (MCs; Deason et al. 2017, DR1), RR Lyrae variables in the MCs (Belokurov et al. 2017, DR1) and in the Galaxy (Iorio et al. 2018, DR1), pre-main sequence (PMS) stars (Vioque et al. 2020, DR2), white dwarfs (Eyer et al. 2020, DR2), and cataclysmic variables (CVs; Abrahams et al. 2020, DR2).

Similarly to Eq. (2), we define amplitude proxies $A_{\mathrm{proxy,BP}}$ and $A_{\mathrm{proxy,RP}}$ for G_{BP} and G_{RP}, respectively, using

$$A_{\mathrm{proxy,BP}} = \sqrt{N_{BP}}\, \varepsilon(I_{BP})/I_{BP}, \quad (3)$$

$$A_{\mathrm{proxy,RP}} = \sqrt{N_{RP}}\, \varepsilon(I_{RP})/I_{RP}, \quad (4)$$

where N_{BP} and N_{RP} are the numbers of observations in G_{BP} and G_{RP}, respectively, and $\varepsilon(I_{BP})$ and $\varepsilon(I_{RP})$ are the published uncertainties on I_{BP} and I_{RP}, respectively.

[3] The expected variance of a weighted mean $\bar{x}$ is $\sigma_{\bar{x}}^2 = \sigma_x^2 V_2/V_1^2$, where σ_x^2 is the true variance of measurements x_i, $V_1 = \sum w_i$ and $V_2 = \sum w_i^2$, with w_i denoting the weights associated with measurements x_i (see Eq. (A.31) in Rimoldini 2014). When $\sigma_{\bar{x}}^2$ is estimated by this expression, the true variance σ_x^2 is unknown but it can be represented by the sample-size unbiased weighted variance $S_x^2 = [V_1^2/(V_1^2 - V_2)]\, s_x^2$, where $s_x^2 = \sum_i w_i (x_i - \bar{x})^2 / V_1$ is the biased weighted variance (Eq. (A.140) of Rimoldini 2014). It follows that $\sigma_{\bar{x}}^2 \approx [V_2/(V_1^2 - V_2)]\, s_x^2$, or $s_x^2/(N-1)$ for N observations in the unweighted limit. In our case, however, $\sigma_{\bar{x}}^2$ is published, so we can estimate the true variance σ_x^2 of the measurements from $\sigma_{\bar{x}}^2 V_1^2/V_2$, or $\sigma_{\bar{x}}^2 N$ in the unweighted scenario.

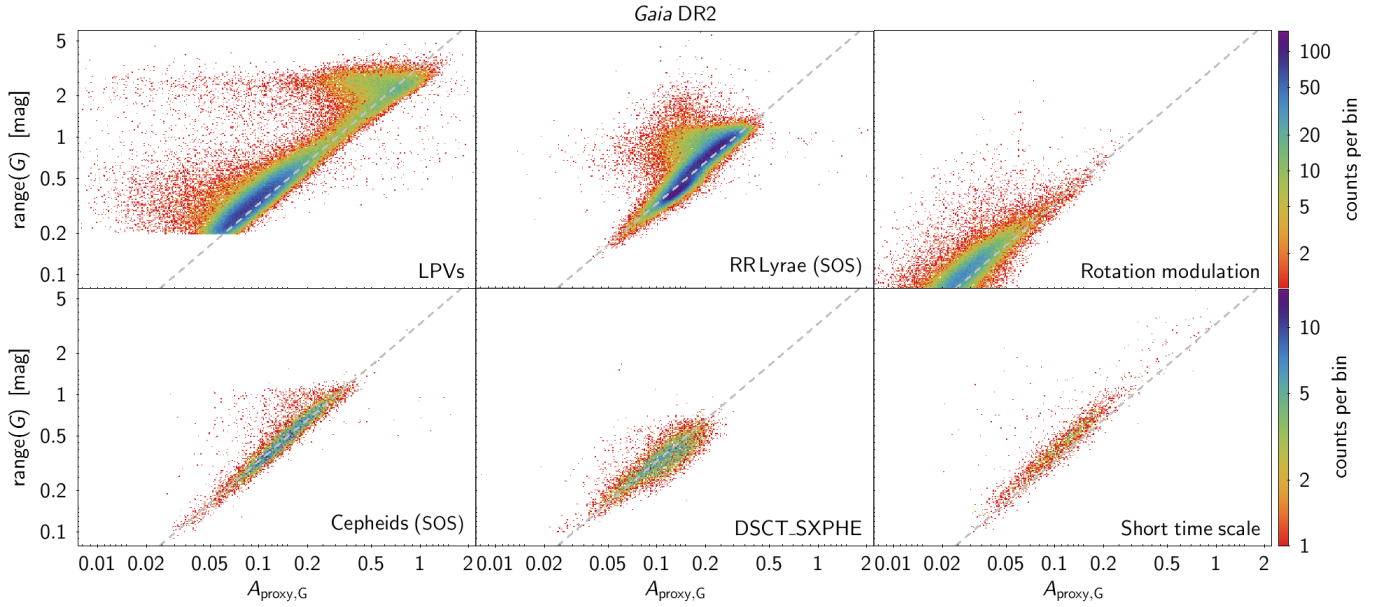

Fig. 1. Density maps of the variability range of G time series (in ordinate) versus amplitude proxy (in abscissa) of selected variable stars published in *Gaia* DR2 for the variability types indicated in the upper-right corner of each panel. The RR Lyrae and Cepheid type candidates shown in the figure are restricted to the subset provided in the `Specific Object Study` (SOS) tables of the data release (see Holl et al. 2018, and more specifically their Fig. 3). The colours of each grid cell in the maps are related to the logarithm of the density of points in the cells according to the colour scale shown on the right of each row of panels. All panels in a given row share the same density colour scale. The dashed diagonal line in each panel corresponds to $\mathrm{range}(G) = 3.3\,A_{\mathrm{proxy},G}$.

2.2. Relation between amplitude proxy and range

The relation between $A_{\mathrm{proxy},G}$ and range(G) is not unique, as it depends on light curve shape and time sampling. For a purely sinusoidal function of amplitude A (i.e. peak-to-peak amplitude $2\,A$), the standard deviation is $\sigma_{\sin} = \sqrt{\frac{1}{\pi}\int_0^{\pi} A^2 \sin^2(x)\,\mathrm{d}x} = A/\sqrt{2}$. Therefore, for a densely and evenly sampled sine light curve $G_{\sin}$, Eq. (1) leads to $\mathrm{range}(G_{\sin}) = 2\sqrt{2}\,\sigma_{\sin} \simeq 3.07\,\sqrt{N_G}\,\varepsilon(I_G)/I_G$. The proportionality constant would be different for other curve shapes. For a triangular or a sawtooth wave, for example, range $= 2\sqrt{3}\,\sigma$, and the proportionality factor would be 3.76 instead of 3.07.

The relation between $A_{\mathrm{proxy},G}$ and range(G) for *Gaia* DR2 is verified with data published in DR2[4]. This is shown in Fig. 1 for the various variability types for which time series have been published in DR2. They concern 151 761 long-period variables (LPVs), 140 784 RR Lyrae variables, 9575 Cepheids, 147 535 main-sequence (MS) variables induced by rotation modulation, 8882 δ Scuti/SX Phoenicis type candidates, and a sample of 3018 short time-scale variables.

Figure 1 shows a proportionality between $A_{\mathrm{proxy},G}$ and range(G) that is globally linear. The relation between these two quantities is, however, not uniquely defined because of at least six reasons. First, the variability proxy is based on the standard deviation of a time series, and its relation to the range depends on the light curve shape. Second, it depends on the sampling of the signal, and thus on the position in the sky because of the *Gaia* scanning law. An example of this dependence is illustrated by the tail objects in Fig. 1 departing from the diagonal line towards small $A_{\mathrm{proxy},G}$ values. This is due to a succession of measurements within a short duration relative to the typical variability time scale of the source (which is not uncommon in the *Gaia* scanning law). These measurements (often within a fraction of a day), have similar fluxes for variables with larger time scales and they bias the standard deviation (and hence $A_{\mathrm{proxy},G}$) towards small values, while the range of the full light curve remains unaffected. Third, $A_{\mathrm{proxy},G}$ is based on fluxes, which does not linearly convert to magnitudes used for the computation of range(G). Fourth, $A_{\mathrm{proxy},G}$ is derived from single CCD fluxes, while range(G) is computed from their integration per field-of-view transit. Fifth, different outlier-removal algorithms are used to disregard corrupt measurements in the per-ccd versus per-transit time series. Finally, the standard deviation used in $A_{\mathrm{proxy},G}$ is more robust against outliers than the peak-to-peak amplitude that defines range(G).

The range(G)/$A_{\mathrm{proxy},G}$ ratio is displayed in Fig. 2 for the various variability types displayed in Fig. 1. It is seen that this ratio comprises between ~3.2 for δ Sct-type variables and ~3.5 for LPVs and rotation modulation MS stars, with a value of ~3.3 for Cepheids and RR Lyrae variables. Only the small sample of short time-scale variables has a distribution peaked at a higher value around four. The relation found for LPVs is consistent with the relation $\mathrm{QR}_5(G) \simeq 3.3\,A_{\mathrm{proxy},G}$ found in Mowlavi et al. (2019) for the same set of *Gaia* DR2 LPVs, $\mathrm{QR}_5(G)$ being the 5–95% quantile range.

Given the above considerations, the relation

$$\mathrm{range}(G) \simeq 3.3\,A_{\mathrm{proxy},G} \tag{5}$$

was chosen for the LAVs studied in this paper. The proportionality factor 3.3 in Eq. (5) is of course approximate, as shown above, but it provides a useful relation to estimate the magnitude

[4] We take for range(G) the quantity `range_mag_g_fov` provided in the `vari_time_series_statistics` table in the *Gaia* archive. A more robust estimate of range(G) is available in DR2 for some variability types, such as for Cepheids and RR Lyrae for which the amplitudes determined from modelled light curves are published, while the value from the statistics table has the advantage to be computed in a similar way for all published variables.

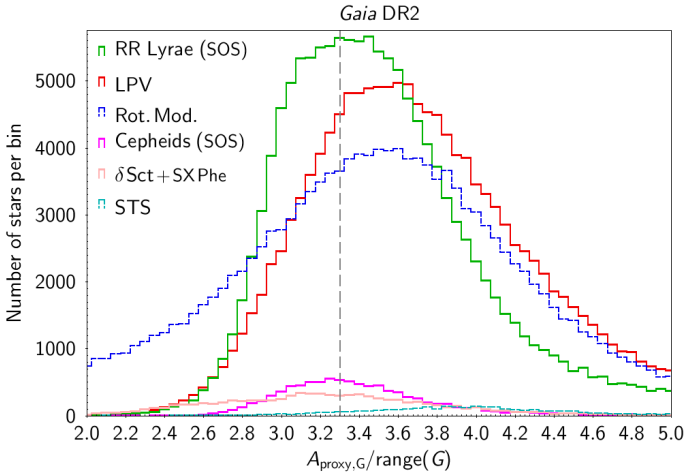

Fig. 2. Histograms of range(G)/$A_{\mathrm{proxy},G}$ ratio for the samples of various variability types shown in Fig. 1. The variability type corresponding to each histogram is written in the top left of the panel in the same colour as the histogram, in decreasing order of the histogram maximum. Pulsating stars are shown in continuous thick lines, while non pulsators, that is MS rotation modulation variables (Rot. Mod.) and short timescale variables, are shown in dashed thin lines. A dashed vertical line is plotted at range(G)/$A_{\mathrm{proxy},G}$ = 3.3.

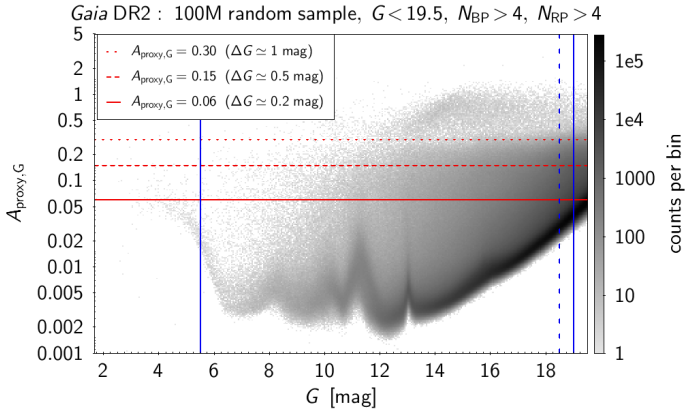

Fig. 3. Density map of the variability amplitude proxy (ordinate) versus G magnitude (abscissa) for a random sample of 100 million *Gaia* DR2 sources that have at least five measurements in G_{BP} and G_{RP}. A dotted, dashed and solid horizontal red line is plotted as eye-guides at $A_{\mathrm{proxy},G}$ = 0.30, 0.15, and 0.06, respectively. They correspond approximately to peak-to-peak amplitudes in G of 1 mag, 0.5 mag, and 0.02 mag, respectively. Vertical continuous blue lines are plotted at G = 5.5 mag and 19 mag, which define the magnitude limits of the sample studied in this paper. Additionally, a vertical dashed blue line is plotted at 18.5 mag.

variability range, which is not available in the *Gaia* DR2 archive, from the amplitude proxy.

3. The catalogue

The distribution of the amplitude proxy $A_{\mathrm{proxy},G}$ defined by Eq. (2) is shown in Fig. 3 versus magnitude G for a random sample of 100 million *Gaia* sources brighter than 19.5 mag. The lower envelope of higher density sources represents constant stars. Sources with amplitude proxies larger than the values characterizing constant stars are potentially variable. Limits at $A_{\mathrm{proxy},G}$ = 0.06, 0.15, and 0.3 are shown in the figure by solid, dashed and dotted red lines, respectively, corresponding to estimated peak-to-peak G variability amplitudes of ∼0.2, ∼0.5, and ∼1 mag, respectively. To avoid contamination by constant stars, we restrict our catalogue to sources with

$$\begin{cases} A_{\mathrm{proxy},G} > 0.06, \\ 5.5 < G/\mathrm{mag} < 19. \end{cases} \quad (6)$$

Figure 3 shows that the contribution of intrinsic stellar variability should dominate that of data noise in these parameter ranges. Caution, however, must be taken at the faintest ($G \gtrsim$ 18.5 mag) end where data noise may provide a larger contribution to $A_{\mathrm{proxy},G}$, and around G = 11 and 13 mag where the photometric data reduction pipeline changes calibration regimes (at 13 mag due to a change of window class, and at 11 mag due to gate activation, see Evans et al. 2018, in particular their Fig. 9). The second condition in Eq. (6) intends to stay clear of the faintest and brightest ends of G where noise (at the faint side) and systematics due to poor handling of saturation in DR2 (at the bright side) become significant relative to intrinsic variability.

3.1. Datasets A, B, and C

We provide three datasets, called Datasets A, B, and C[5]. Each dataset is a subset of the previous one, with Dataset A being the full catalogue of LAVs. The number of sources in each dataset, and the filtering conditions that lead to their definitions are summarized in Table 1. The datasets are characterized as follow.

Dataset A. This dataset contains all LAVs that satisfy Eq. (6) and are cleaned from sources whose light curves appear to be affected by instrumental arteafacts at specific times of the mission (filter a1 in Table 1, see Appendix A.2 for more information). Sources that may potentially contain G epoch magnitudes fainter than 20.5 mag are also excluded (filter a2 in Table 1, see Appendix A.3).

The procedure used to import the data from the *Gaia* DR2 archive and details on the filtering criteria are given in Appendix A.

Dataset B. This subset of Dataset A is to be preferentially used if reliable G_{BP} and G_{RP} magnitudes are needed (such as for colour-magnitude diagrams). The selection relies on the fact that I_{BP} + I_{RP} must be close to I_G as a result of the wavelength transmission bands of G, G_{BP}, and G_{RP} (Evans et al. 2018). A source with a larger-than-expected summed flux I_{BP}+I_{RP} relative to I_G is therefore suspected to have inconsistent G, G_{BP} and G_{RP} measurements. While unreliable BP and RP flux excesses are, in DR2, due in many cases to BP/RP integrated fluxes of poorer quality, similar problems can also affect G-band measurements. We refer to Appendix B for a discussion on this (see in particular Appendix B.3).

The BP and RP flux excess (I_{BP} + I_{RP})/I_G depends on the spectral type, and thus on G_{BP}−G_{RP} colour. We derive in Appendix B a normalized BP and RP flux excess, notated C' (Eq. (B.3)), which should be close to one at all G_{BP}−G_{RP} colours for typical stars, and apply the filtering criteria b2 and b3 listed in Table 1 to derive Dataset B. This can be done only if the source has I_{BP} and I_{RP} values in *Gaia* DR2, which imposes the additional selection criterion b1 listed in Table 1.

[5] The catalogue of LAVs is available for download, see Appendix D.

Table 1. Summary of the number of sources in Datasets A, B, and C, and of the number of sources removed by the successive filtering criteria that lead from the public *Gaia* DR2 archive (first line in the table) to each dataset.

Criterion	Nbr of sources
$5.5 < G < 19$, $A_{\mathrm{proxy},G} > 0.06$	23 830 345
(a1) In sky stripes: $A_{\mathrm{proxy},G} > 0.1$ or $G < 18.3$	−514 084
(a2) $G + 1.65\,A_{\mathrm{proxy},G} < 20.5$	−387
Dataset A	*23 315 874*
$(\varpi/\epsilon(\varpi) > 10)$	(401 480)
(b1) Has G_{BP} and G_{RP}	−5 535 102
(b2) $C' < 1.04 + 0.001\,(G_{\mathrm{BP}}-G_{\mathrm{RP}} - 1)^3$	−16 626 421
(b3) $C' > 0.9$	−5490
Dataset B	*1 148 861*
$(\varpi/\epsilon(\varpi) > 10)$	(110 521)
(c1) $A_{\mathrm{proxy},G} < 1.5\,A'_{\mathrm{proxy,BP+RP}}$	−66 364
(c2) $A_{\mathrm{proxy},G} > 0.8\,A'_{\mathrm{proxy,BP+RP}}$	−121 237
(c3) $N_{\mathrm{BP,RP}} \geq 10$	−25 755
(c4) $\lvert N_{\mathrm{BP}} - N_{\mathrm{RP}}\rvert \leq 1$	−211 607
(c5) $7.8 < N_G/N_{\mathrm{RP}} < 10.2$	−93 323
(c6) $G_{\mathrm{BP}} + 1.65\,A_{\mathrm{proxy,BP}} < 20.5$	−11 609
Dataset C	*618 966*
$(\varpi/\epsilon(\varpi) > 10)$	(85 046)

Notes. The subsets in these datasets that have parallax uncertainties better than 10% are shown in parenthesis.

Dataset C. This subset of Dataset B is to be preferentially used if reliable $A_{\mathrm{proxy,BP}}$ and $A_{\mathrm{proxy,RP}}$ are needed (such as for multi-band variability studies in G, G_{BP}, and G_{RP}). The selection relies on the fact that the variability in BP + RP must be consistent with the variability in G given the wavelength transmission bands. A variability in G that is not present in BP + RP is suspicious (note, however, that this could happen in the case of an anti-correlated variability in the blue and in the red[6]). Likewise, a variability observed in BP + RP but not in G may indicate additional noise in G_{BP} and/or G_{RP} that would make $A_{\mathrm{proxy,BP}}$ and/or $A_{\mathrm{proxy,RP}}$ unreliable (note, however, that, in such a case, $A_{\mathrm{proxy},G}$ may still be reliable).

The amplitude proxy $A_{\mathrm{proxy,BP+RP}}$ of the summed BP + RP is not available in *Gaia* DR2, and cannot be computed with the available DR2 quantities. This would require flux time series in order to evaluate the covariance term between G_{BP} and G_{RP}. Therefore, we derive in Appendix C an approximation to $A_{\mathrm{proxy,BP+RP}}$, notated $A'_{\mathrm{proxy,BP+RP}}$, that neglects the covariance term but is computable with the available DR2 data (Eq. (C.10)). The filtering conditions c1 and c2 listed in Table 1 use this quantity to select sources for Dataset C, based on the analysis performed in Appendix C on the conditions expected to be satisfied by $A'_{\mathrm{proxy,BP+RP}}$ (Eq. (C.12)). In addition, we require the sources to have at least ten measurements in G_{BP} and G_{RP} (condition c3), and to have similar numbers of field-of-view transits in G, G_{BP}, and G_{RP} (conditions c4 and c5). These extra conditions are meant to ensure similar time distributions between the three photometric time series, a condition that is essential for useful comparison of their variability properties given the large amplitudes considered here. Finally, sources that may potentially contain G_{BP} epoch magnitudes fainter than 20.5 mag are also excluded (filter c6 in Table 1). We note that the equivalent condition for G_{RP} is always satisfied.

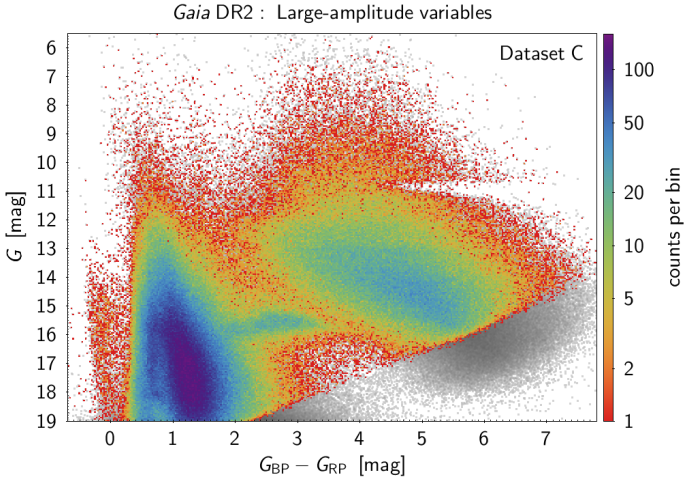

Fig. 4. Density map of the colour-magnitude diagram of Dataset C. Dataset B is plotted in grey in the background. The axes range have been limited for better visibility. The lack of very red sources at $G \simeq 11$ mag in Datasets B and C is due to limitations in the DR2 processing leading to too low BP and RP flux excesses (see text).

Colour-magnitude diagram. The colour-magnitude (CM) diagram of Dataset C is shown in Fig. 4. It reveals a lack of very red sources ($G_{\mathrm{BP}}-G_{\mathrm{RP}} \gtrsim 4.5$ mag) at $G \simeq 11$ mag. This is due to limitations in the DR2 processing, as shown in Appendix B.2, which lead to too low BP and RP flux excesses for very red stars at these magnitudes (see in particular Fig. B.11). The feature is present in Dataset B as well (shown in grey in the background of Fig. 4) as the exclusion of sources with too small BP and RP flux excesses is performed with filter b3 listed in Table 1. The excess of sources around $G \simeq 16$ mag with $G_{\mathrm{BP}}-G_{\mathrm{RP}}$ from 2 mag to 3.5 mag in Fig. 4 is linked to the population of LPVs in the Magellanic Clouds.

G-band variability. The distributions of $A_{\mathrm{proxy},G}$ versus G and versus $G_{\mathrm{BP}}-G_{\mathrm{RP}}$ are shown in Fig. 5 for the three datasets (we stress however that Dataset A should in principle not be used when $G_{\mathrm{BP}}-G_{\mathrm{RP}}$ colour is required). We highlight here two features seen in these diagrams to illustrate the pros and cons of the various datasets. The first concerns the presence of a population with very large amplitudes in all three datasets, with $A_{\mathrm{proxy},G} \gtrsim 0.3$. The great majority of them are Miras, as suggested by their red colours in the right panels of Fig. 5. They are relatively numerous in Dataset A, but their number decreases significantly in Datasets B and C. This loss of completeness from Dataset A to B and C is addressed in Sect. 3.2.

The second feature is the presence of a large number of faint LAV candidates ($G \gtrsim 18$ mag) in Dataset A (top-left panel in Fig. 5) with amplitudes close to the lower limit of $A_{\mathrm{proxy},G} = 0.06$

[6] An example of anti-correlated blue/red variability is given by Ap stars. The variability of these stars is due to the presence of spots caused by the migration of certain chemical elements modifying the opacity of the star's surface. The presence of spots may induce two effects in the star's atmosphere: a blocking effect and then a back-warming effect. For cold Ap stars (<10 000 K), the blocking effect is in the blue part of the stellar spectra, and the re-emission in the red. The variations in different filters can be anti-phased (Muciek et al. 1985), thus strongly attenuating the variations integrated in a large filter. This was highlighted for the wide HIPPARCOS H_{p} band by Eyer (1998) (Sects. 10.2.2 and 13.3). It can also be seen in the motion of variable stars in the Hertzsprung-Russell diagram (such as in Fig. 11 of Gaia Collaboration 2019), where Ap stars (α^2 Canum Venaticorum) have horizontal motions, that is noticeable variations in $G_{\mathrm{BP}}-G_{\mathrm{RP}}$ with little change in the G band.

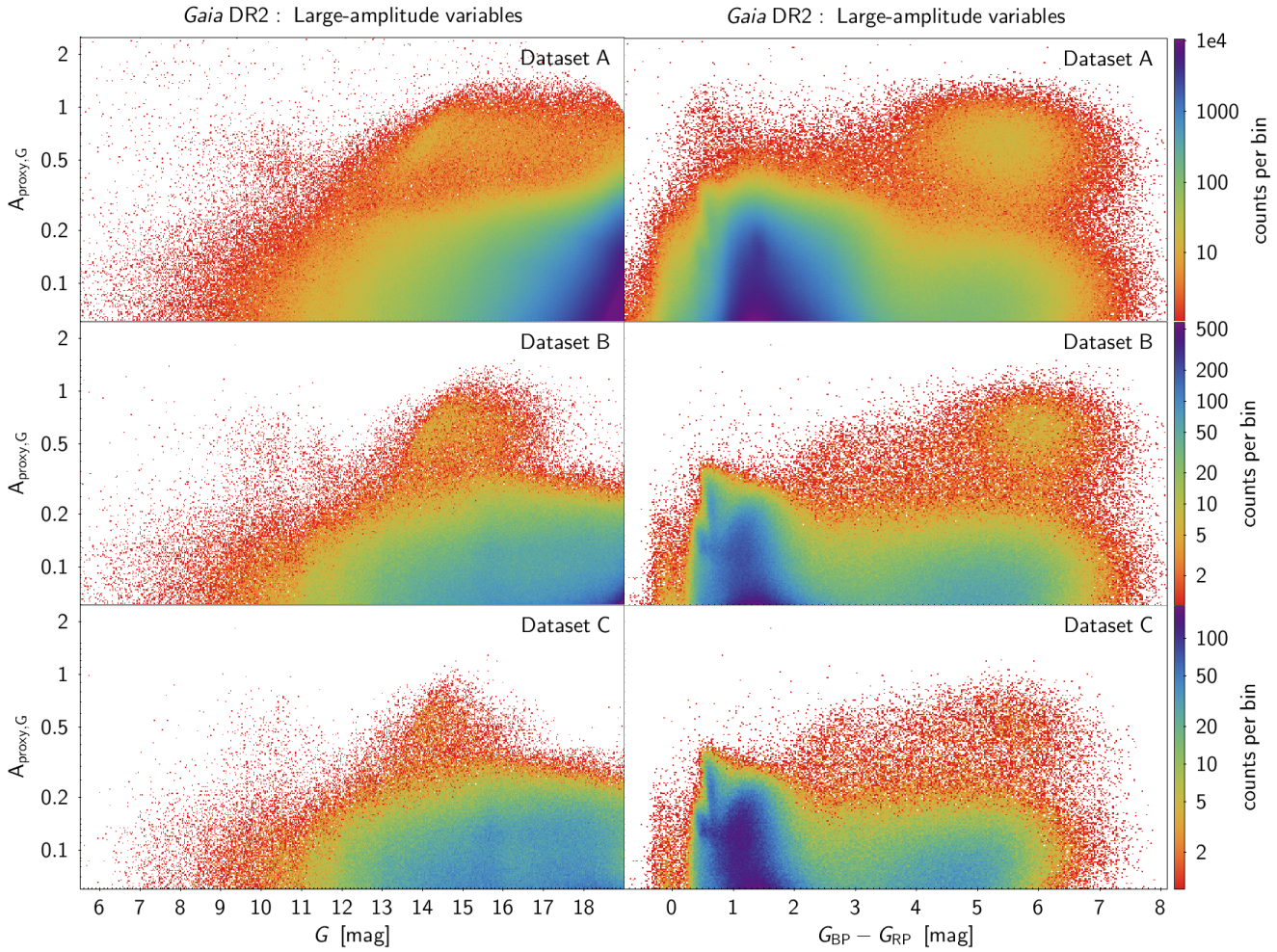

Fig. 5. Density maps of the G amplitude proxy versus G magnitude (*left panels*) and versus $G_{BP}-G_{RP}$ colour (*right panels*) for Datasets A (*top panels*) B (*middle panels*) and C (*bottom panels*). The colours of each grid cell in the maps are related to the logarithm of the density of points in the cells according to the colour scale shown on the right of each row of panels. The two panels in a given row share the same density colour scale. We note that the *top-right panel* of Dataset A should be analysed with caution, as it contains sources with unreliable G_{BP} and/or G_{RP}.

considered here. These are most probably contaminants due to increasing noise level when G approaches 19 mag, as seen in Fig. 3. This excess of faint LAVs is much smaller in Dataset B, and basically absent in Dataset C (Fig. 6). The purity of the datasets will be addressed in Sect. 3.3.

Multi-band variability. Figure 7 displays $A_{\rm proxy,RP}/A_{\rm proxy,G}$ versus $A_{\rm proxy,BP}/A_{\rm proxy,G}$. High densities of sources are observed in specific regions of the diagram, revealing distinct multi-band variability properties. The densest region contains quasi-achromatic variables with $A_{\rm proxy,G} \simeq A_{\rm proxy,BP} \simeq A_{\rm proxy,RP}$. The next two densest groups are both observed in the region where variability amplitude is larger in the blue than in the red (i.e. on the right side of the uppermost dashed line in Fig. 7). This is seen more clearly in Fig. 8 that displays $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ versus $G_{BP}-G_{RP}$, where the two groups are observed at $A_{\rm proxy,BP}/A_{\rm proxy,RP} \simeq 1.63$ and 2.2, respectively. The multi-band variability properties of Dataset C is analysed in Sect. 4.1.

3.2. Completeness

The completeness of the three datasets is difficult to assess in absolute terms. Estimates are attempted in this section based on data from the *Gaia* DR2 catalogues of variable stars (Sect. 3.2.1), from the ASAS-SN survey (Sect. 3.2.2), and from the ZTF survey (Sect. 3.2.3).

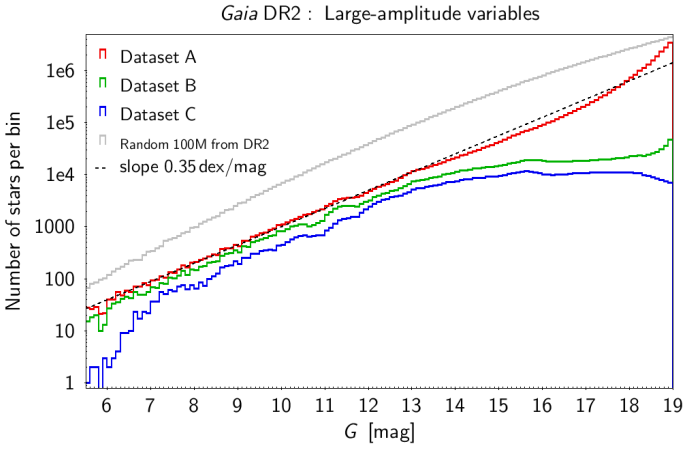

Fig. 6. G magnitude histograms for, from top to bottom lines, a sample of 100 million sources randomly taken from *Gaia* DR2 (in grey), Dataset A (in red), Dataset B (in green), and Dataset C (in blue). Bins are 0.1 mag wide. On the log scale of the ordinate, a dotted straight line with slope 0.35 dex mag^{-1} is adjusted to the histogram of Dataset A.

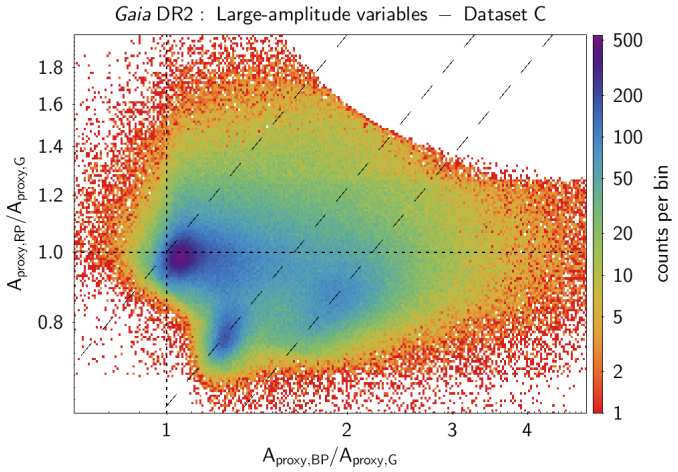

Fig. 7. $A_{\rm proxy,RP}/A_{\rm proxy,G}$ versus $A_{\rm proxy,BP}/A_{\rm proxy,G}$ for all LAV candidates in Dataset C. Dashed lines are drawn at $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ = 1, 1.63 and 2.2. Dotted lines are further added at $A_{\rm proxy,RP} = A_{\rm proxy,G}$ and $A_{\rm proxy,BP} = A_{\rm proxy,G}$ to guide the eyes. The axes ranges have been limited for better visibility.

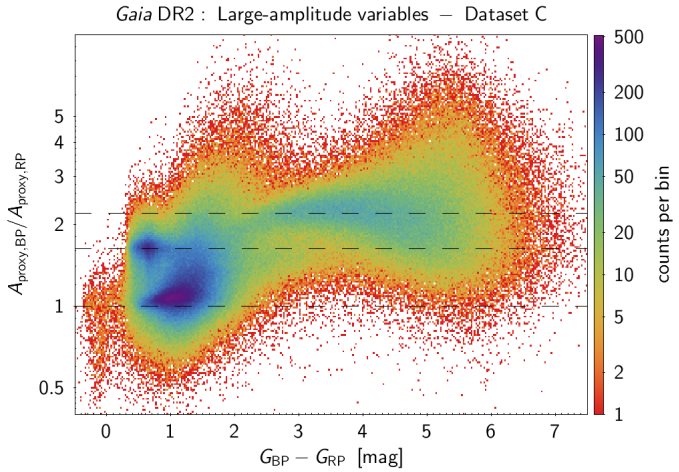

Fig. 8. Same as Fig. 7, but for $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ versus $G_{\rm BP}-G_{\rm RP}$.

3.2.1. Completeness estimate based on *Gaia* DR2 variables

The completeness relative to *Gaia* DR2 variables is given in Table 2 for the six variability types published in DR2, that is LPVs, RR Lyrae stars, Cepheids, MS rotation modulation variables, δ Scuti/SX Phoenicis stars, and short time-scale variables. For RR Lyrae and Cepheid variables, we consider both *Gaia* DR2 samples provided in the classification and Specific Object Study (SOS) tables (see Holl et al. 2018). The completeness is estimated by checking the fraction of these variables that are recovered in Datasets A, B, and C. To achieve this, we first restrict the DR2 samples to the conditions defining our datasets, that is $5.5 < G/\mathrm{mag} < 19$ and $A_{\rm proxy,G} > 0.06$. The number of variables satisfying these conditions for each variability type are given in the third column of Table 2. The fraction of these variables that are present in Datasets A, B, and C are then provided in the fourth to sixth columns, respectively, with their percentages given in the row below the numbers.

Table 2 shows that practically all variables published in DR2 are present in Dataset A. For dataset B, a difference is observed between pulsating and non-pulsating stars. Pulsating stars have completeness levels between 58% and 96% in Dataset B. This is excellent considering that Dataset B contains only ∼5% of Dataset A. For non strictly periodic stars, the completeness levels are much lower, being of 33% for the DR2 rotation modulation variables, and only 5% for short time-scale variables. We note that rotation modulation candidates are not expected to have variability amplitudes larger than ∼0.2 mag, a statement supported by the very small fraction of the DR2 rotation modulation candidates that have $A_{\rm proxy,G} > 0.06$ (2181 out of 147 535 candidates, see Table 2). The situation is different for short time-scale candidates. Contrary to the case of rotation modulation candidates, the majority of them do have large amplitudes in DR2 (2641 out of 3018 candidates, see Table 2), and these are all, except one, in Dataset A. However, only 5% of them remain in Dataset B, a reduction factor that is similar to the overall reduction from Dataset A to B (4.8%, see Table 1). We remind that the short time-scale candidates published in DR2 were identified from their variability in the G-band CCD timeseries, while we are dealing here with G-band transit photometry. That difference may explain their relative numbers in Datasets A and B.

For Dataset C, the fraction of LAVs kept from Dataset B is around 60% to 70% for all variability types, except for short time-scale candidates that have a reduction factor of ∼50% from Dataset B to C. We stress that the completeness numbers given in Table 2 are upper limits, because the catalogues of variables published in *Gaia* DR2 are themselves not complete (their completeness varies greatly with variability type and sky location, see Tables 2 and 3 of Holl et al. 2018).

3.2.2. Completeness estimate based on ASAS-SN survey

The ASAS-SN survey of variable stars (Shappee et al. 2014; Jayasinghe et al. 2018) has published 666 502 sources[7], of which 646 027 have a *Gaia* crossmatch ID identified in their catalogue. To compare with our sample of *Gaia* LAVs, we apply two filters to the initial ASAS-SN dataset. The first filter consists in considering only ASAS-SN sources that have Amp(V) variability amplitudes larger than 0.2 mag to comply with the lower G amplitude limit in our datasets. The resulting ASAS-SN sample contains 468 527 sources, of which 454 910 sources have a *Gaia* DR2 ID in their catalogue.

The second filter aims at limiting the number of ASAS-SN – *Gaia* mismatched identifications. Mismatches could result from, among other reasons, the poorer sky resolution of ASAS-SN (∼8″) compared to that of *Gaia* (∼0.4″). We therefore exclude sources that have ASAS-SN V magnitudes potentially incompatible with the G magnitudes of their *Gaia* crossmatch candidates. The compatibility must take into account the two different photometric filter responses. The distribution of $V-G$ for all ASAS-SN sources with *Gaia* DR2 IDs is shown in Fig. 9 versus $G_{\rm BP}-G_{\rm RP}$. The relation

$$V = G + \begin{cases} 0 & G_{\rm BP}-G_{\rm RP} < 0.25 \\ 0.29\,(G_{\rm BP}-G_{\rm RP}-0.25)^{1.63} & 0.25 < G_{\rm BP}-G_{\rm RP} < 3.25 \\ 0.96\,(G_{\rm BP}-G_{\rm RP}) - 1.3818 & G_{\rm BP}-G_{\rm RP} > 3.25 \end{cases} \quad (7)$$

is found to describe well the transformation from G to V as a function of $G_{\rm BP}-G_{\rm RP}$ for the LAVs. It is also compatible with the $G-V$ relation provided by Evans et al. (2018) within their colour validity range. We restrict our final ASAS-SN sample to sources that have a maximum deviation of 0.5 mag between the

[7] We use in this paper the ASAS-SN catalogue downloaded from https://asas-sn.osu.edu/variables on June 2, 2020.

Table 2. Completeness of Datasets A, B, and C with respect to the samples of variable stars published in dedicated *Gaia* DR2 catalogues of variable stars.

Variability type	Total	$5.5 < G/\mathrm{mag} < 19$, $A_{\mathrm{proxy},G} > 0.06$				Reference
	in DR2	in DR2	Dataset A	Dataset B	Dataset C	
LPV (SOS)	151 761	138 193	138 150	111 941	76 527	Mowlavi et al. (2018)
				81%	*68%*	
RR Lyr (classif)	195 780	127 123	127 103	74 189	48 891	Rimoldini et al. (2019)
				58%	*66%*	
RR Lyr (SOS)	140 784	90 024	89 976	59 824	39 977	Clementini et al. (2019)
				66%	*67%*	
Cep (classif)	8550	8367	8362	5403	3887	Rimoldini et al. (2019)
				65%	*72%*	
Cep (SOS)	9575	9197	9194	5685	4131	Clementini et al. (2019)
				62%	*73%*	
δ Sct, SX Phe (classif)	8882	3328	3310	3183	2016	Rimoldini et al. (2019)
				96%	*63%*	
Rot. modul. (SOS)	147 535	2181	2129	692	463	Lanzafame et al. (2018)
				33%	*67%*	
Short time-scale (SOS)	3018	2641	2640	129	60	Roelens et al. (2018)
				5%	*47%*	

Notes. The origin of the samples in DR2 (i.e. either classification table or SOS table, see Holl et al. 2018, in particular their Fig. 3) is indicated in parenthesis. Reference to the paper describing the specific catalogue is given in the last column. The numbers in italics denote the percentages of sources present in Datasets B and C compared to Datasets A and B, respectively.

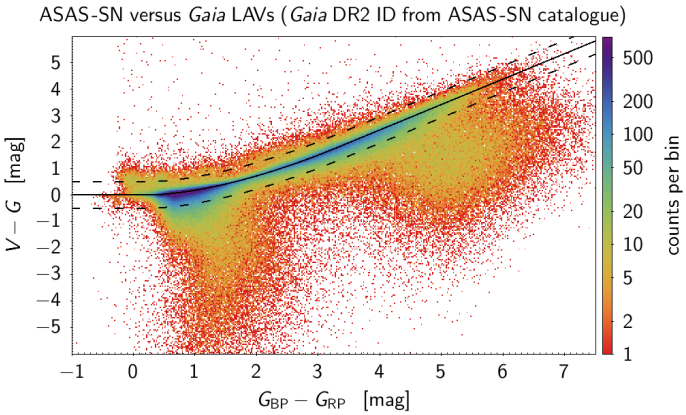

Fig. 9. Density map of the difference between mean ASAS-SN V and mean *Gaia* G magnitudes versus *Gaia* G_{BP}−G_{RP} colour for all ASAS-SN sources with a *Gaia* DR2 ID provided in the ASAS-SN catalogue. A random number between −5 and +5 mmag has been added to V to smooth out the two-digit precision magnitude published in the ASAS-SN catalogue. This is not necessary for the *Gaia* magnitudes, which are reported with four digits in that catalogue. The solid line represents Eq. (7) as a function of G_{BP}−G_{RP}, and the two dashed lines represent deviations of 0.5 mag above and below this function. The axis ranges have been limited for better visibility.

observed ZTF V and the value that would be obtained from G with relation (7) (i.e. that are located between the two dashed lines in Fig. 9).

The final 'cleaned' sample of sources satisfying both conditions, on Amp(V) and on V−G, depends on *Gaia* crossmatch identification. Using the *Gaia* DR2 IDs listed in the ASAS-SN catalogue, we get 328 249 sources, with only 59 641 of them present in our Dataset A (see Table 3). If we perform a sky crossmatch between ASAS-SN and our *Gaia* LAVs using a search cone radius of 4″, we find 198 591 crossmatches. This is about three times more than the above mentioned number of sources with a *Gaia* DR2 IDs in the ASAS-SN catalogue for this sample. The condition Amp(V) > 0.2 mag is not at the origin of this discrepancy since we get similar conclusions with Amp(V) > 0.4 mag (see Table 3). The V variability amplitudes in the final samples are also globally compatible with the values of $A_{\mathrm{proxy},G}$ of their *Gaia* crossmatched counterparts, as shown in Fig. 10 for the sample using sky crossmatches (a similar diagram is obtained using the ASAS-SN crossmatches). We thus do not know the reason for the discrepancy between the number of crossmatches reported in the ASAS-SN catalogue and the number found with a direct sky crossmatch. Therefore, we cannot use the sample of ASAS-SN LAVs to estimate the completeness of Dataset A. However, we can check the fraction of crossmatches that remain from Dataset A to B, and from B to C. It amounts to ∼90% from Dataset A to B, and to ∼75% from Dataset B to C, irrespective of the initial sample (among the four cases listed in Table 3). This shows that the majority of ASAS-SN LAVs that are present in Database A have relatively good *Gaia* multi-band photometry.

3.2.3. Completeness estimate based on ZTF survey

We consider the ZTF catalogue of periodic variables published by Chen et al. (2020) that contains 781 602 sources. We follow the same procedure as is done for ASAS-SN in Sect. 3.2.2, here applied to the r_{ZTF}-band photometry of ZTF. We first restrict the ZTF sample to sources with Amp(r_{ZTF}) > 0.2 mag. This selects 484 306 sources, of which 67% (324 646 sources) have a 2″ sky crossmatch with our Dataset A (the numbers are summarized in Table 3). We then adopt the following relation to convert from G to r_{ZTF} for our LAVs:

$$r_{\mathrm{ZTF}} = G + \begin{cases} -0.08 + 0.15673\,(G_{\mathrm{BP}} - G_{\mathrm{RP}} - 1.25)^2 & G_{\mathrm{BP}} - G_{\mathrm{RP}} < 3 \\ (2/3)\,(G_{\mathrm{BP}} - G_{\mathrm{RP}}) - 1.6 & G_{\mathrm{BP}} - G_{\mathrm{RP}} > 3 \end{cases} \quad (8)$$

Table 3. Completeness estimates of Datasets A, B, and C based on ASAS-SN (Jayasinghe et al. 2018) and ZTF (Chen et al. 2020) surveys.

Survey	Amplitude-limited sample		Crossmatch	Cleaned	Dataset A	Dataset B	Dataset C
ASAS-SN 666 502	Amp(V) > 0.2 mag:	468 527	from ASAS-SN	328 249	59 641	54 713	40 772
						92%	*75%*
			4″ XM on Dataset A	198 591	198 591	181 999	135 196
						92%	*74%*
	Amp(V) > 0.4 mag:	284 277	from ASAS-SN	195 440	43 353	39 314	28 648
						90%	*73%*
			4″ XM on Dataset A	146 287	146 287	132 518	96 293
						91%	*73%*
ZTF 781 602	Amp(r) > 0.2 mag:	484 306	2″ XM on Dataset A	311 726	311 726	225 458	146 746
						72%	*65%*
	Amp(r) > 0.4 mag:	250 821	2″ XM on Dataset A	160 602	160 602	113 288	73 673
						71%	*65%*

Notes. The numbers in italics denote the percentages of sources present in Datasets B and C compared to Datasets A and B, respectively.

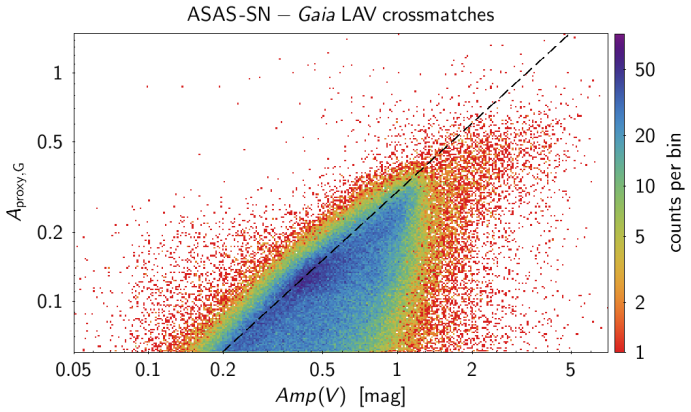

Fig. 10. Density map of the *Gaia* G amplitude proxy versus ASAS-SN V amplitude of all crossmatches found using a cone search of 4″ on the sky and that have V values between the two dashed lines in Fig. 9. A random number between −5 and +5 mmag has been added to the V amplitude to smooth out the two-digit precision numbers published in the ASAS-SN catalogue for this quantities. The dashed line shows the relation Amp(V) = 3.3 $A_{proxy,G}$. The axis ranges have been limited for better visibility.

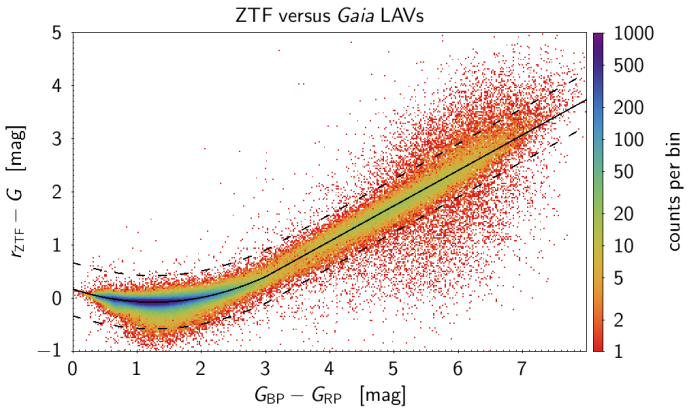

Fig. 11. Same as Fig. 9, but for ZTF r_{ZTF} magnitudes of all ZTF sources crossmatched with *Gaia* LAVs of Dataset A using a 2″ cone search on the sky. The solid line represents Eq. (8) as a function of G_{BP}−G_{RP}, and the two dashed lines represent deviations of 0.5 mag above and below this function.

which is shown by the solid line in Fig. 11. The dispersion of r_{ZTF}−G in the ZTF sample around the fiducial relation (8) is much smaller than it was the case for ASAS-SN (compare Figs. 9 and 11). We still keep a filtering condition of a maximum of 0.5 mag dispersion of r_{ZTF} with respect to $r_{ZTF}(G)$ for ZTF sources, which leads to a final ZTF sample of 311 726 sources. They are all by construction in Dataset A. 225 458 of them are present in Dataset B and 146 746 sources are in Dataset C. These reduction factors of 72% from Dataset A to B and of 65% from Dataset B to C are comparable to the factors obtained in Sect. 3.2.1 using *Gaia* DR2 variables.

The sky resolution of ZTF is very good, with 60% of the 324 646 ZTF – Dataset A crossmatches having an angular separation less than 0.1″, and 95% less than 0.2″. Given the magnitude depth of ZTF (r_{ZTF} up to 21 mag), we may expect that all ZTF sources are present in *Gaia*. The r_{ZTF} amplitude from ZTF is compared to $A_{proxy,G}$ in Fig. 12. It confirms the $A_{proxy,G}$ = 3.3 Amp(r_{ZTF}) relation established in Sect. 2.2 (Eq. (5)). It also confirms an inevitable dispersion around this relation due to survey properties (such as time sampling and photometric precision) and stellar variability and spectral properties (such as photometric filter responses). In particular, a non-negligible fraction of ZTF sources with Amp(r_{ZTF}) > 0.2 mag have $A_{proxy,G}$ < 0.06 and are missed in Dataset A (see Fig. 12). This observation enables us to estimate a completeness factor of 67% of Dataset A relative to the ZTF catalogue of periodic variables.

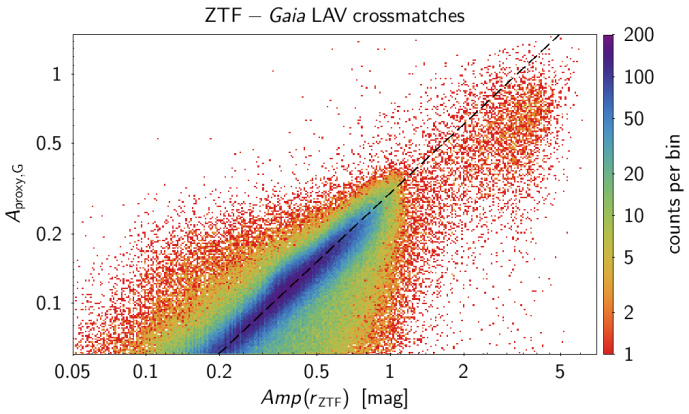

Fig. 12. Same as Fig. 10, but for *Gaia* G amplitude proxy versus ZTF r_{ZTF} amplitude of all crossmatches found using a cone search of 2″ on the sky and that have r_{ZTF} values between the two dashed lines in Fig. 11.

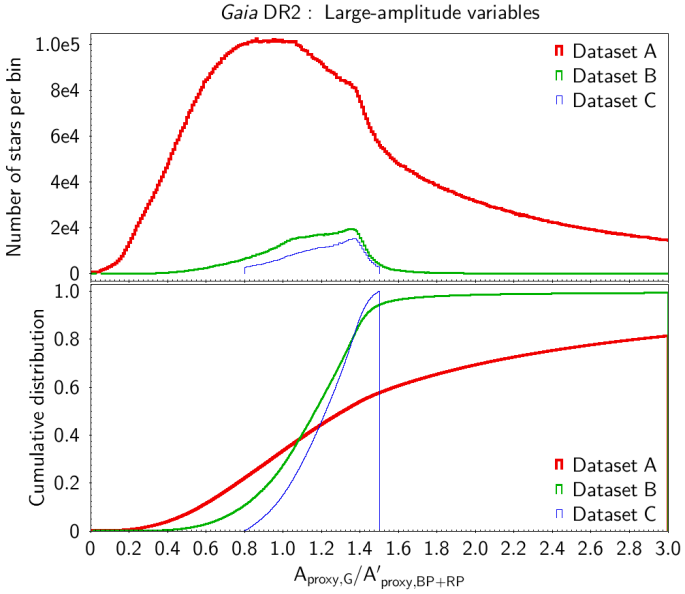

Fig. 13. Distribution of $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ in Datasets A (thick red line), B (green line) and C (thin blue line), displayed as histograms in the *top panel* and as cumulative histograms in the *bottom panel*.

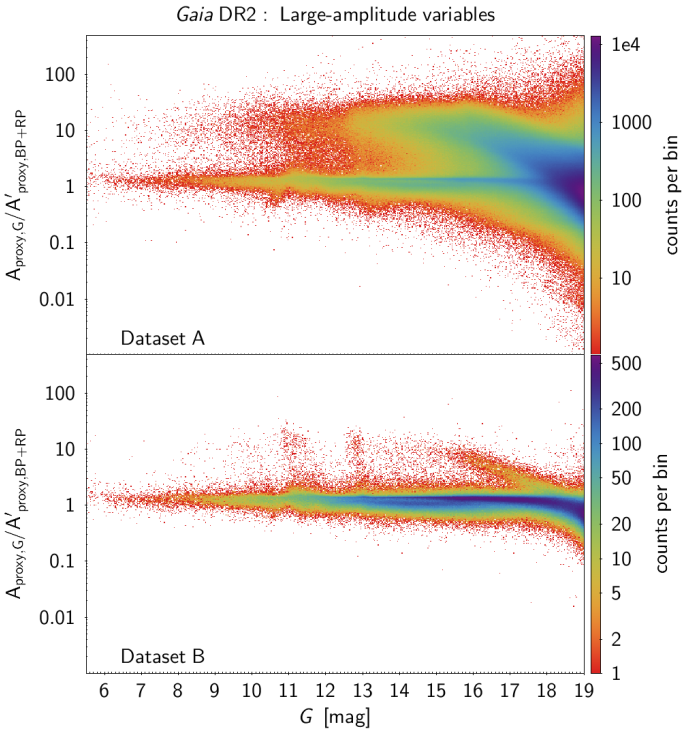

Fig. 14. Density map of the ratio $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ versus G for Dataset A (*top panel*) and B (*bottom panel*). The abscissa range has been limited for better visibility. The ordinate scales are kept identical in the two panels.

3.3. Purity

We address the question of the purity of Datasets A, B, and C in two different ways. The first method is based on the consistency of variability amplitudes in the three *Gaia* photometric bands (Sect. 3.3.1). The second method makes use of magnitude distributions (Sect. 3.3.2).

3.3.1. Purity estimate based on multi-band variability

Any variability detected in G should be present in BP + RP. Therefore, we can estimate an upper limit for the purity level by checking the consistency between G-band amplitude ($A_{\mathrm{proxy},G}$) and combined BP+RP-band amplitude ($A'_{\mathrm{proxy,BP+RP}}$). The analysis presented in Appendix C concludes that $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ should lie between ∼1 and ∼1.5, the exact value depending on variability type. However, the histograms of $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ for the three datasets, shown in Fig. 13 (top panel), reveals a large fraction of sources with ratios outside this range. This is especially true for Dataset A.

The first case to consider is $A_{\mathrm{proxy},G} < A'_{\mathrm{proxy,BP+RP}}$, that is when the variability amplitude detected in G is smaller than the one detected in BP + RP. The cumulative histogram of $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$, displayed in the bottom panel of Fig. 13, shows that about one quarter of LAV candidates in Datasets A and B have $A_{\mathrm{proxy},G} < A'_{\mathrm{proxy,BP+RP}}$, and 14% in Dataset C. This can be the case if, for example, the noise in G_{BP} and G_{RP} is larger than the noise in G due to, among other reasons, increased residual astrophysical background, fewer CCD transits for G_{BP} and G_{RP} than for G, or blending effects in *BP* and *RP* spectra. Figure 14, which displays $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ versus G for both Dataset A (top panel) and Dataset B (bottom panel), tends to support this latter explanation, as the number of cases with $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}} < 1$ increases with increasing magnitude, especially for Dataset A. We therefore cannot, in general, use the criterion $A_{\mathrm{proxy},G} < A'_{\mathrm{proxy,BP+RP}}$ to identify spurious $A_{\mathrm{proxy},G}$ values. Rather, it would point to an overestimation of $A'_{\mathrm{proxy,BP+RP}}$, and hence to unreliable $A_{\mathrm{proxy,BP}}$ and/or $A_{\mathrm{proxy,RP}}$ values. A condition based on this conclusion, but using the less restrictive condition $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}} < 0.8$, was actually used to filter out such cases in Dataset C (filter c2 in Table 1).

In the second case, when $A_{\mathrm{proxy},G} > 1.5\,A'_{\mathrm{proxy,BP+RP}}$, the amplitude is unexpectedly larger in the G band than in the BP+RP band. This would point to a spurious value of $A_{\mathrm{proxy},G}$. It represents 33% of sources in Dataset A, but only 6% in Dataset B (bottom panel of Fig. 13). These sources were removed in Dataset C (filter c1 in Table 1).

If we were to consider that $A_{\mathrm{proxy},G}$ is reliable if $1.0 < A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}} < 1.5$ (the first condition being restrictive if interpreted as being due to the unreliability of G rather than of G_{BP} and/or G_{RP}, see above), we would conclude from the above estimates that the purity level with respect to $A_{\mathrm{proxy},G}$ could be around 40% for Dataset A, 70% for Dataset B, and 85% for Dataset C. These numbers, however, must be taken with caution.

3.3.2. Purity estimate based on magnitude distribution

The G magnitude distributions of the three datasets are shown in Fig. 6. The magnitude distribution of Dataset A (red line) basically follows an exponential increase as a function of magnitude up to $G \simeq 13$ mag (a power-ten function with a slope of 0.35 dex mag^{-1} is plotted in dotted line in the figure). Above that magnitude, the slope slightly decreases up to $G \simeq 16$ mag, before strongly increasing at magnitudes above ∼17 mag. For comparison, the magnitude distribution of 100 million sources randomly selected from *Gaia* DR2 is shown in grey in Fig. 6. It reveals a continuous decrease in the dex mag^{-1} slope as a function of magnitude. If we assume a spatial distribution of LAVs in the Galaxy similar to that of all stars, the comparison of magnitude

distributions of Dataset A with the random sample suggests the presence of a significant fraction of contaminants in Dataset A at magnitudes above ∼17 mag. The number of faint contaminants is much reduced in Dataset B, whose magnitude distribution flattens above ∼16 mag (green line in Fig. 6). Yet, the increase observed at $G \gtrsim 18$ mag still indicates the presence of contaminants at the faintest end of this sample. This is no longer the case for Dataset C, whose magnitude distribution even decreases above ∼18 mag.

In conclusion, comparison of the magnitude distributions of the three datasets with that of a random *Gaia* DR2 sample suggests the presence of a non-negligible fraction of contaminants at the faint side (∼17 mag) in Dataset A. It supports a similar conclusion obtained in Sect. 3.3.1, and confirms the higher purity levels estimated for Datasets B and C. Contaminants are still expected to pollute Dataset B at magnitudes fainter than ∼18 mag, while Dataset C is the purest of the three datasets.

4. Catalogue exploration: Two examples

We provide in this section two examples that illustrate the content of the catalogue and its usage. The first case investigates the exploitation of multi-band variability amplitudes to disentangle and study different types of variable stars (Sect. 4.1). Section 4.2 then presents the sample of LAVs with parallax uncertainties better than 10%.

4.1. Multi-band variability studies

The availability of quasi-simultaneous photometric measurements in three bands confers to the *Gaia* mission an invaluable advantage for variability studies. This is obviously the case when analyzing light curves. However, it is also an advantage for studies using variability proxies because quasi-simultaneous observations lead to consistent variability amplitude proxies in the different bands. Multi-band measurements should be taken within time intervals that are short compared to the expected variability time-scale if compatible amplitudes are required in the different bands. And this time-scale can be short for such variables as EA-type eclipsing binaries with short-duration deep eclipses, flare stars, or transient objects, to cite only a few types. Therefore, in general, quasi-simultaneous photometric measurements ensure coherent variability proxies in different bands. In *Gaia*, simultaneous photometry within less than one minute is ensured in G, $G_{\rm BP}$, and $G_{\rm RP}$. Dataset C has been defined with conditions enforcing, as much as possible, similar epoch measurements in these three bands for a given source, and is therefore best suited for multi-band variability analyses. We therefore restrict our study in this section to Dataset C.

We introduced two diagrams in Sect. 3.1 that evidenced the dependence of the blue-to-red amplitude ratio ($A_{\rm proxy,BP}/A_{\rm proxy,RP}$) on stellar variability type (Figs. 7 and 8), with at least three categories of variables highlighted in the figures. Here, we further investigate this property based on literature data (Sect. 4.1.1), and identify four broad groups of variables from their multi-band variability properties (Sect. 4.1.2).

4.1.1. $A_{\rm Proxy,BP}/A_{\rm Proxy,RP}$ ratios of known variability types

We identify the variability types of the LAVs in Dataset C from three sources in the literature: the *Gaia* DR2 catalogues of variables (already used in Sect. 3.2.1), the ZTF catalogue of periodic variables (already used in Sect. 3.2.3), and the Simbad database (Wenger et al. 2000) from which crossmatches with Dataset C are extracted using a 2″ cone search on the sky. The $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ distributions of the crossmatches are shown in Fig. 15, in the top panel for *Gaia* DR2 variables, in the two middle panels for the ZTF variables, and in the bottom panel for variables extracted from Simbad among a selection of variability types. The histograms are plotted normalized to the maximum count-per-bin for better visibility in the top three panels, but kept as total count per bin for the Simbad crossmatches in the bottom panel due to the small number of sources per variability type.

Several categories of variable stars are identified from the distributions shown in Fig. 15. First, the *Gaia* DR2 samples of RR Lyrae variables (green histogram in the top panel of Fig. 15), Cepheids (filled pink), and δ Scuti/SX Phoenicis stars (filled yellow) are seen to be distributed around $A_{\rm proxy,BP}/A_{\rm proxy,RP} \simeq 1.6$. These distributions are confirmed by the ZTF samples of the same variability types, shown in the second panel from top. They represent pulsating stars in the classical instability strip (which we denote hereafter as 'classical pulsators'). Their $G_{\rm BP}-G_{\rm RP}$ colours and $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ ratios are shown in the top panel of Fig. 16.

Second, the distribution of the *Gaia* DR2 sample of LPVs, shown in red in the top panel of Fig. 15, are seen to pulsate with amplitudes about twice larger in the blue than in the red, with a peak of the distribution at $A_{\rm proxy,BP}/A_{\rm proxy,RP} \simeq 2.1$. However, a relatively large dispersion around this peak value is observed, extending from $A_{\rm proxy,BP}/A_{\rm proxy,RP} \simeq 1.2$ to above 3. The ZTF sample shown in the second panel from top distinguishes between semi-regular variables (SRVs) and Miras. Interestingly, Miras (dashed red histogram) are seen to peak at blue-to-red amplitude ratios similar to those of classical pulsators, between 1.4 and 1.7, while SRVs (solid red line) have, on the mean, ratios above 1.8. While the $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ ratios of the LPVs overlap with those of classical pulsators at ratios below ∼1.8, they are nevertheless easily identified from their red colours, as seen in Fig. 16 (top panel).

Third, some variability types are approximately achromatic with $A_{\rm proxy,BP} \simeq A_{\rm proxy,RP}$. In the ZTF sample of periodic variables, EW-type eclipsing binaries are among the most abundant ones (filled blue histogram in the third panel from top in Fig. 15). Their $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ histogram peaks at ≃1.07. EA-type eclipsing binaries (filled pink histogram) follow a distribution similar to the EW eclipsing binaries, though with a slightly wider dispersion around the peak value. These patterns are compatible with the type of binaries they represent, EA-type binaries consisting of detached systems where the two stars keep their individual characteristics in the majority of cases, while EW-type systems share a common envelope around the two stars. Other variable stars also display achromatic variability. The $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ histogram of the rotation-modulation LAV candidates published in *Gaia* DR2, shown in the top panel of Fig. 15 (blue histogram), also peaks at $A_{\rm proxy,BP}/A_{\rm proxy,RP} \simeq 1.1$ (we must however keep in mind that a fraction of these DR2 rotation-modulation candidates are misclassified, especially the ones considered in this paper, see Sect. 3.2). Achromatic variables span a wide range of $G_{\rm BP}-G_{\rm RP}$ colours, typically from ∼0.2 to ∼3 mag as seen in the middle panel of Fig. 16 for the ZTF variables and in the top panel for the DR2 rotation-modulation candidates.

Two other non-pulsating types of LAVs from the ZTF sample are shown in the third panel from top in Fig. 15. They are the RS Canum Venaticorum (RS CVn) and BY Draconis (BY Dra) variables. The $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ distribution of the RS CVn candidates peaks at values between 1.25 and 1.3 (dotted green histogram). Interestingly, the peak of the distribution is

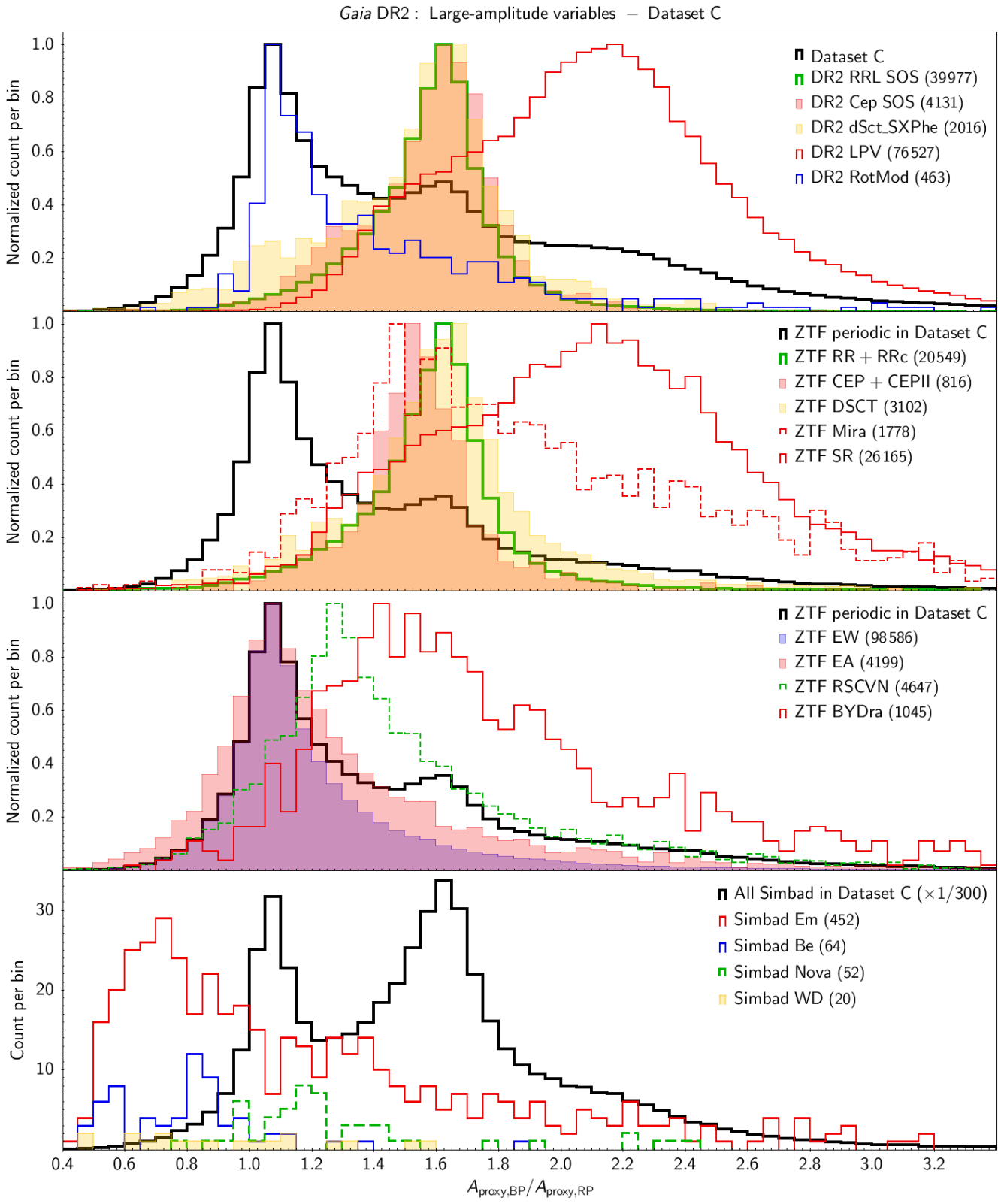

Fig. 15. Histograms of $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}}$ for various types of variable candidates identified in the catalogues of *Gaia* DR2 variables (*top panel*), in the ZTF catalogue of periodic variables (*middle panels*) and in Simbad (*bottom panel*; 'Em' are emission-line stars). Only crossmatches with Dataset C are considered. The variability type of each histogram is indicated in the upper-right corner of each panel, with the number of crossmatches available in Dataset C indicated in parenthesis next to the variability type. The thick black line in each panel represents the histogram of the full sample of crossmatches in Dataset C of the relevant catalogue. The histograms are normalized to maximum count in the *top three panels*, and the actual counts per bin in the *bottom panel*. The counts in the histogram of the full sample of Simbad crossmatches (thick black line in the *bottom panel*) have been divided by 300 for better visibility. The abscissa range has been limited for better visibility.

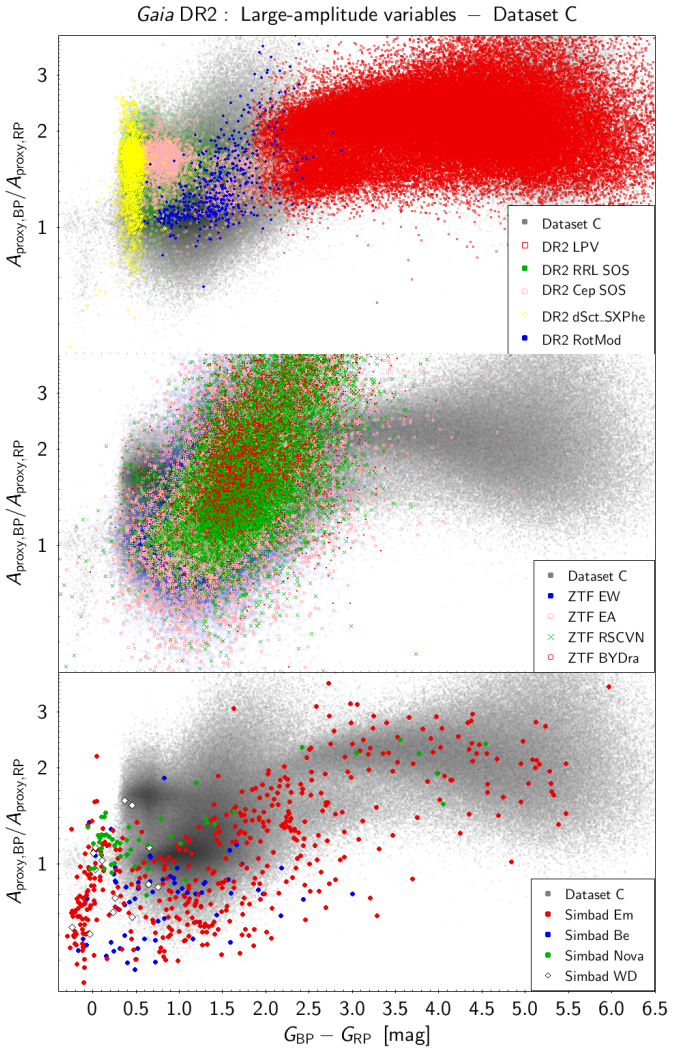

Fig. 16. Distribution in the $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ versus $G_{\rm BP}-G_{\rm RP}$ diagram of Dataset C LAVs crossmatched with some of the *Gaia* DR2 catalogues of variables (*top panel*), with the sample of non-pulsating variables from the ZTF catalogue of periodic variables (*middle panel*), and with a selection of variability types crossmatched in the Simbad databse. The colours of the symbols used to represent them are the same as the ones in Fig. 15. The background grey points represent the full Dataset C shown in Fig. 8. The axes ranges have been limited for better visibility.

relatively well defined, and its width not much larger than the widths observed for the classical pulsators shown in the top two panels. RS CVn variables are close binary systems with chromospheric activity and large spots on the stellar surface. The BY Dra candidates in the ZTF sample, on the other hand, show a much wider $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ distribution, between ∼1.2 and ∼1.9 (red histogram in the third panel from top), without a clear peak value. These variables also have active chromospheres, but are single stars, typically K- and M-type MS stars. The location of RS CVn and BY Dra variables in the $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ versus $G_{\rm BP}-G_{\rm RP}$ diagram is shown in the middle panel of Fig. 16.

Finally, Fig. 8 reveals the presence of a population of blue stars ($G_{\rm BP}-G_{\rm RP} \lesssim 0.2$ mag) with variability amplitudes larger at long than at short wavelengths ($A_{\rm proxy,BP}/A_{\rm proxy,RP} < 1$). No such specific class is found in the samples of *Gaia* DR2 variables or in the ZTF periodic variables. We therefore browsed the Simbad database for crossmatches in Dataset C that have $A_{\rm proxy,BP}/A_{\rm proxy,RP} < 1$, and report in the bottom panel of Fig. 15 some of the variability classes to which they belong. The most numerous class consists of emission-line stars, the histogram of which is shown in red (labelled 'Simbad Em') in the bottom panel of Fig. 15. Be stars (blue histogram) form another class with amplitudes larger in the red than in the blue. They are also a type of emission-line stars. Some white-dwarf (WD) variables also have $A_{\rm proxy,BP}/A_{\rm proxy,RP} < 1$ (filled yellow histogram), though their number statistics is very small (only 20 crossmatches found) and their $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ distribution extends up to 1.6. Novae do not necessarily have $A_{\rm proxy,BP}/A_{\rm proxy,RP} < 1$, though some do (dashed green histogram). The small number statistics of these Simbad crossmatchs prevents, however, to draw firm conclusions. Further insight into this group of $A_{\rm proxy,BP}/A_{\rm proxy,RP} < 1$ variables will be provided from the analysis of the sample with good parallaxes in Sect. 4.2. The colour distribution of the variables discussed here is shown in the bottom panel of Fig. 16.

The above analyses rely on variability types published in the literature, which are, however, affected by uncertainties. *Gaia* DR2 and ZTF identifications result from automated techniques, and the ones in Simbad have a wide and non-homogeneous origin. Broad distributions can thus be expected for parameters derived from the analysis of these catalogues.

The summed distribution of $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ in a given survey is very informative of the overall stellar sample. These distributions are shown with thick black lines in each panel of Fig. 15 for each of the Dataset C, ZTF and Simbad samples. The distribution of Dataset C (top panel) is very similar to that of the ZTF sample of periodic variables (middle panels), with a predominance of achromatic variables. In the ZTF sample, the main peak at $A_{\rm proxy,BP} \simeq A_{\rm proxy,RP}$ is due to eclipsing binaries. Extrapolating this result to Dataset C, we could thus expect that most of the quasi-achromatic variables in Dataset C also consist of eclipsing binaries (this will be confirmed in Sect. 4.2.3). In contrast, the summed distribution for the sample of Dataset C–Simbad crossmatches shows a predominance of $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ around 1.6, typical of classical pulsators (thick black line in the bottom panel of Fig. 15).

4.1.2. Classification of LAVs using $A_{\rm proxy,BP}/A_{\rm proxy,RP}$

Based on the results of the previous section, we schematically categorize LAVs in four groups according, mainly, to their blue-to-red $A_{\rm proxy,BP}/A_{\rm proxy,RP}$ amplitude ratio and their $G_{\rm BP}-G_{\rm RP}$ colour. The four groups and their definitions are summarized in Table 4.

Group 1. We first consider LPVs (Group 1), which are easily identified by their red colours. We require $G_{\rm BP}-G_{\rm RP} > 1.8$ mag. This, however, will also include MS red dwarf and PMS stars including young stellar objects (YSOs). To restrict Group 1 to LPVs (the other types will belong to other groups), we take advantage of their very bright intrinsic luminosities. At the large variability amplitudes considered in this paper, LPVs mainly consist of red giants on the Asymptotic Giant Branch (AGB) (as well as the even brighter, though statistically less numerous, red supergiants). Their typical absolute G magnitudes (M_G) lie between −3 and 0 mag. MS red dwarfs, on the other hand, are much fainter, with typical brightnesses of $M_G \simeq 8$ mag at $G_{\rm BP}-G_{\rm RP} = 2$ mag, and up to $M_G \simeq 11$ mag at $G_{\rm BP}-G_{\rm RP} = 3$ mag. They are thus of the order of ten magnitudes fainter than typical LPVs. At brightnesses between these two extremes, we find PMS stars, usually still several magnitudes fainter than LPVs.

Table 4. Schematic categorization of variables using colour and wavelength-dependent variability (Figs. 7 and 8).

Group 1 (mainly LPVs)
$\begin{cases} G_{\rm BP}-G_{\rm RP} > 1.8 \\ \varpi < 0.12 + \exp\left[10 + 3\,(G_{\rm BP}-G_{\rm RP}) - 1.5\,G\right] \end{cases}$ (9)
Group 2 (hot compact LAVs with $\frac{A_{\rm proxy,BP}}{A_{\rm proxy,RP}} < 0.9$)
$\begin{cases} G_{\rm BP}-G_{\rm RP} < 0.2 \\ A_{\rm proxy,BP} < 0.9\,A_{\rm proxy,RP} \\ \varpi > 0.12 + \exp\left(19 - 1.5\,G\right) \end{cases}$ (10)
Group 3 (mainly classical pulsators)
$\begin{cases} \text{Not in Group 1 or 2} \\ A_{\rm proxy,BP} > 1.4\,A_{\rm proxy,RP} \\ A_{\rm proxy,RP} < 0.85\,A_{{\rm proxy},G} \end{cases}$ (11)
Group 4 (mainly non-pulsating variables)
Not in Group 1, 2 or 3 (12)
Group 4a (mainly chromatic non-pulsating variables)
$\begin{cases} \text{In Group 4} \\ A_{\rm proxy,BP} > 1.2\,A_{\rm proxy,RP} \\ A_{\rm proxy,RP} < 0.92\,A_{{\rm proxy},G} \end{cases}$ (13)

Notes. In each group, the listed conditions must all be satisfied (`'AND'` operator).

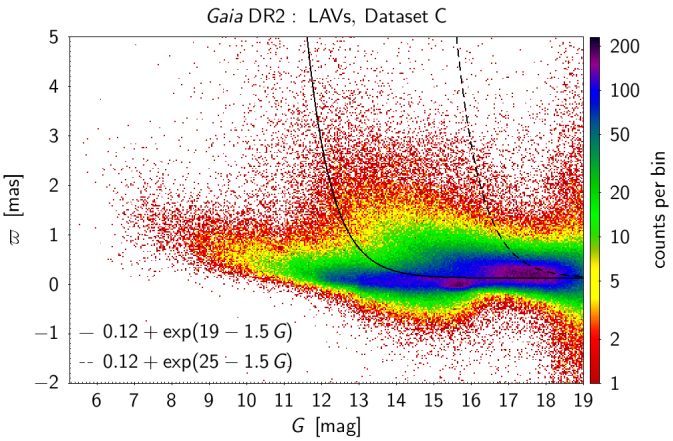

Fig. 17. Density map of the parallax versus G for Dataset C. The lines are examples of parallax limits below which stars are classified as LPVs (for $G_{\rm BP}-G_{\rm RP} = 3$ mag for the solid line and for for $G_{\rm BP}-G_{\rm RP} = 5$ mag for the dashed line). A rainbow colour-code is used for the density map to highlight the location of the densest regions with respect to the solid and dashed lines. The axes ranges have been limited for better visibility.

The much larger intrinsic luminosities of LPVs compared to red dwarfs and PMS stars translate into much smaller parallaxes at any given observed magnitude. To illustrate this, we will consider a $G = 15$ mag LPV. If its absolute G magnitude is $M_G \simeq 0$ mag, the LPV would have a parallax of $\varpi = 10^{-0.2\,(G-M_G-10)}$ mas = 0.1 mas ($d = 10$ kpc). A red clump clump star that would be reddened at a colour of $G_{\rm BP}-G_{\rm RP} = 3$ mag (typical of not too evolved LPVs as shown in Fig. 8) would have $M_G \simeq 3.5$ mag. Consequently, its parallax would be of 0.5 mas ($d = 2$ kpc) if it were seen with $G = 15$ mag. A MS red dwarf at the considered colour, on the other hand, with $M_G \simeq 11$ mag, would need to be much closer to have $G = 15$ mag, at a parallax of $\varpi \simeq 15.8$ mas ($d = 63$ pc). Therefore, a star at $G = 15$ mag and $G_{\rm BP}-G_{\rm RP} = 3$ mag has a high probability to be a LPV if its parallax is smaller than ∼0.5 mas. The upper parallax limit sketched above for a star to be a LPV decreases with increasing magnitude. It also depends on colour. An empiric exploration of LAVs in Dataset C leads to the condition $\varpi < 0.12 + \exp\left[10 + 3\,(G_{\rm BP}-G_{\rm RP}) - 1.5\,G\right]$ mas to identify LPV candidates in the sample of red LAVs (the additive factor of 0.12 mas counts for the typical *Gaia* DR2 parallax uncertainty). The final conditions for Group 1 are given by Eq. (9) in Table 4.

Conditions (9) properly select LPV candidates in the sample of Dataset C LAVs with parallax uncertainties better than 10% (see Sect. 4.2), but also in the full sample of Dataset C due to their much larger brightnesses compared to other red stars. Conditions (9) also correctly select LPVs in the Magellanic Clouds, as seen in Fig. 17 where they form the over-density of sources at $\varpi \simeq 0$ mas and $15 \lesssim G/\text{mag} \lesssim 16.2$. We note that red LAVs other than LPVs can also be present in this group, such as R CrB and RV Tauri variables.

Regarding the G-band variability amplitudes of these LPVs, most of them have $A_{{\rm proxy},G} \lesssim 0.3$ (see top panel of Fig. 18), which corresponds to peak-to-peak G amplitudes of less than ∼1 mag. Miras stand out at amplitudes larger than this value. Group 1 contains one third of all LAVs in Dataset C.

Group 2. We gather in this group blue LAVs ($G_{\rm BP}-G_{\rm RP} < 0.2$ mag) with variability amplitudes larger in the red than in the blue ($A_{\rm proxy,BP} < 0.9\,A_{\rm proxy,RP}$). We restrict to hot stars fainter than the MS, including WDs and subdwarfs, leaving hot MS LAVs such as Ae or Be stars to Groups 3 and 4 that contain MS stars. In order to do so, we use a magnitude-dependent limit on the parallax similar to the method used for LPVs in Group 1. The condition is given by Eq. (11) in Table 4. We note, however, that the absolute magnitude separation between hot subdwarfs and MS blue variables is only between ∼3 and ∼5 mag, which will necessarily lead to some confusion for sources that do not have good parallaxes. Group 2 contains only a small fraction of all LAVs, about 0.1% in Dataset C.

Group 3. The multi-band amplitude ratios of the Dataset C sample cleaned from Group 1 and 2 stars are shown in Fig. 19. The removal of Group 1 stars from the sample has adequately removed the over-density of sources that was present around $A_{\rm proxy,BP}/A_{\rm proxy,RP} = 2$ in the sull Dataset C sample (see Fig. 7). Two main groups remain clearly visible in Fig. 19, one below $A_{\rm proxy,RP}/A_{{\rm proxy},G} \simeq 0.85$, and another one above that limit. The former group contains sources that have significantly larger variability amplitudes in blue than in red, and forms our Group 3. This group is further restricted to $A_{\rm proxy,BP}/A_{\rm proxy,RP} > 1.4$, leading to the sample below the solid line in Fig. 19. The conditions for Group 3 are summarized in Table 4.

The variables in this group are mainly the classical pulsators identified in Sect. 4.1.1, with $A_{\rm proxy,BP} \simeq 1.63\,A_{\rm proxy,RP}$. Their G amplitudes are shown in the second panel from top in Fig. 18 (sources with $A_{\rm proxy,BP}/A_{\rm proxy,RP} > 1$). Group 3 contains about 10% of Dataset C LAVs.

Group 4. The fourth group contains the remaining LAVs not in Groups 1 to 3, that is at $A_{\rm proxy,RP}/A_{{\rm proxy},G}$ ratios above the solid

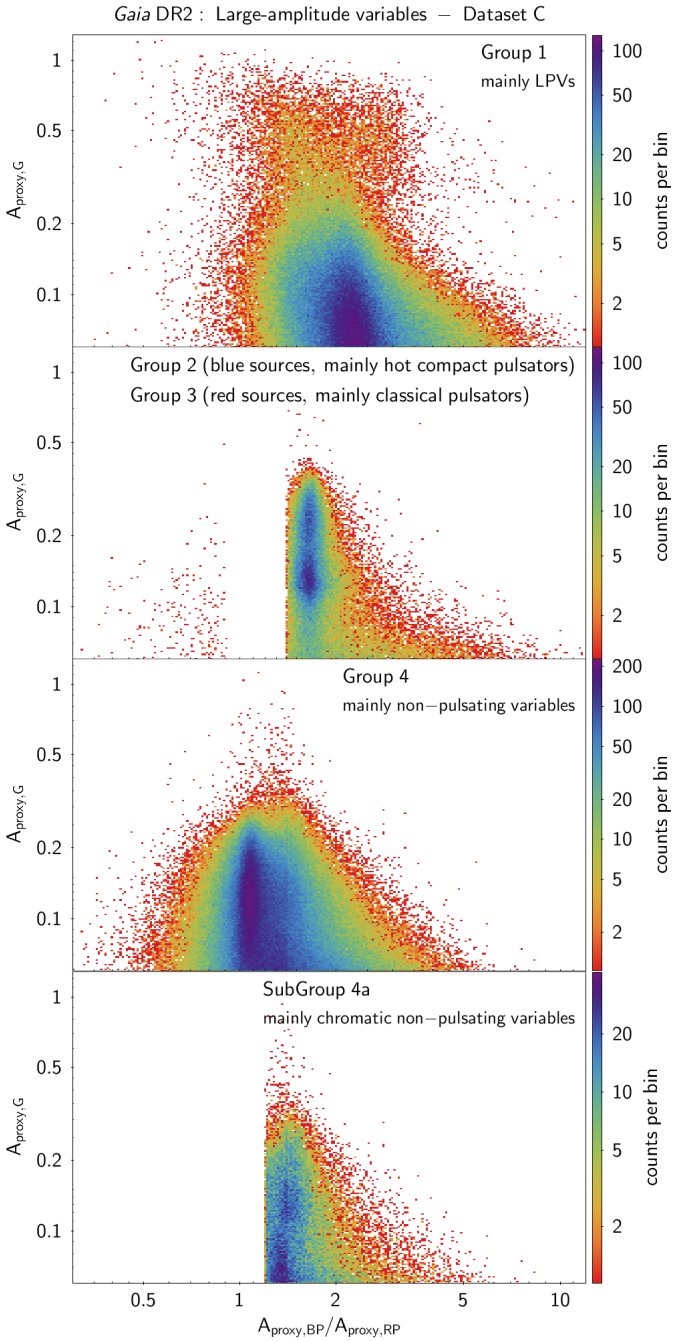

Fig. 18. Density maps of $A_{proxy,G}$ versus $A_{proxy,BP}/A_{proxy,RP}$ for the different groups in Dataset C. *Top panel*: Group 1 (mainly LPVs). *Second panel from top*: Group 2 for sources at $A_{proxy,BP}/A_{proxy,RP} < 0.9$ (mainly hot compact LAVs) and Group 3 for sources at $A_{proxy,BP}/A_{proxy,RP} > 1.4$ (mainly classical pulsators). *Third panel from top*: Group 4 (mainly non-pulsating variables). *Bottom panel*: Subgroup 4a (mainly chromatic non-pulsating variables).

line in Fig. 19 (condition (12) in Table 4). They correspond to variables with $A_{proxy,BP} \simeq A_{proxy,G} \simeq A_{proxy,RP}$. The great majority of them are known to be non-pulsating variables. In particular, they contain eclipsing binaries, as seen in the ZTF sample of periodic variables (see Sect. 4.1.1).

Figure 19 further reveals a small over-density of sources, within Group 4, close to the transition between Groups 3 and 4, at $1.2 < A_{proxy,BP}/A_{proxy,RP} < 1.4$ (between the solid and dotted lines in the figure). With amplitudes 20% to 40% larger in

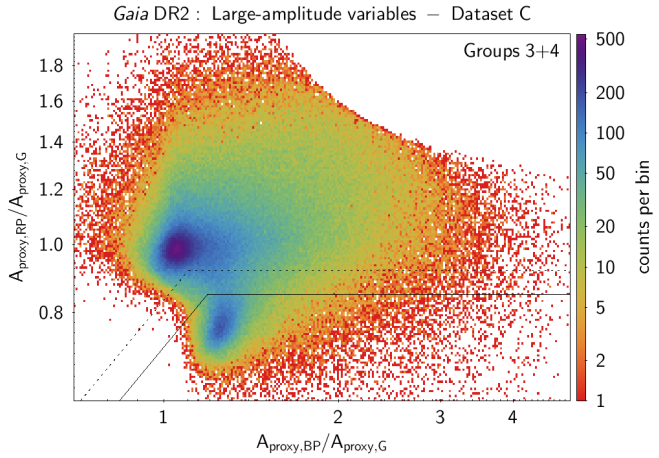

Fig. 19. Same as Fig. 7, but for Groups 3 and 4 in Dataset C. The solid-line delineates Group 3 (below the line) and Group 4 (above the line). The dashed line identifies Subgroup 4a at the small $A_{proxy,RP}/A_{proxy,G}$ side of Group 4 (see text). The diagonal lines are given by $A_{proxy,BP}/A_{proxy,RP} = 1.2$ (dashed diagonal line) and $A_{proxy,BP}/A_{proxy,RP} = 1.4$ (solid diagonal line).

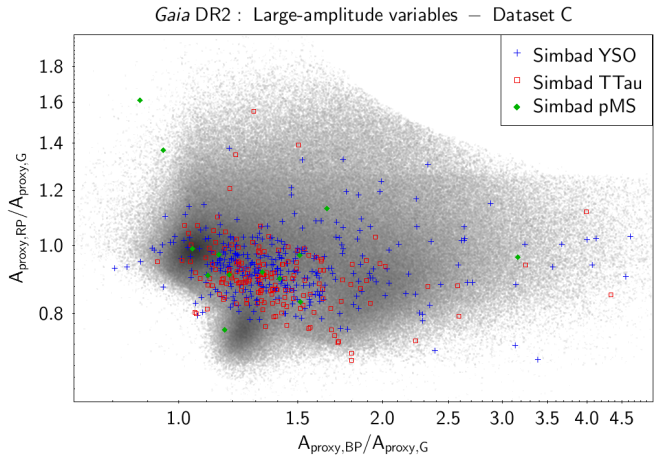

Fig. 20. Same as Fig. 7, but with Dataset C sources crossmatched with Simbad YSOs, T Tauri stars and pre-MS stars shown with the markers labelled in the upper-right inset of the figure.

G_{BP} than in G_{RP}, the variability cannot be considered achromatic. We therefore define Subgroup 4a with the conditions (13) listed in Table 4, which select the sample between the solid and dotted lines in Fig. 19. The RS CVn variables in the ZTF sample analysed in Sect. 4.1.1 typically fall in this subgroup. Pre-MS variables, YSOs and T Tauri variables are also found in this subgroup, as shown in Fig. 20 where these stars identified from the Simbad database have been plotted. The G amplitude distribution of Subgroup 4a is shown in the bottom panel of Fig. 18. It confirms the relevance of this subgroup as being distinct within Group 4, with an over-density of sources at $1.25 \lesssim A_{proxy,BP}/A_{proxy,RP} \lesssim 1.5$.

Group 4 is the most populated of the four groups, gathering 58% of all LAVs in Dataset C. Subgroup 4a contains 10% of Group 4.

4.2. The sample with parallaxes better than 10%

We present in this section the sample of LAVs with relative parallax uncertainties better than 10%. We do not impose any restriction on the number of visibility periods used in

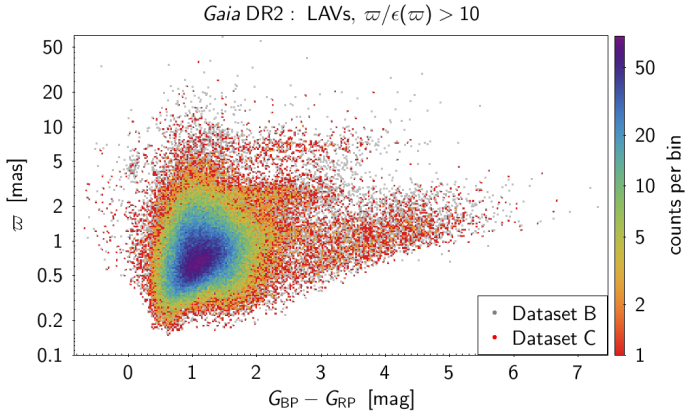

Fig. 21. Parallax versus colour for sources that have relative parallax uncertainties better than 10% in Dataset C (coloured with the density of points according to the colour scale shown on the right of the figure). Dataset B is plotted in the background in grey.

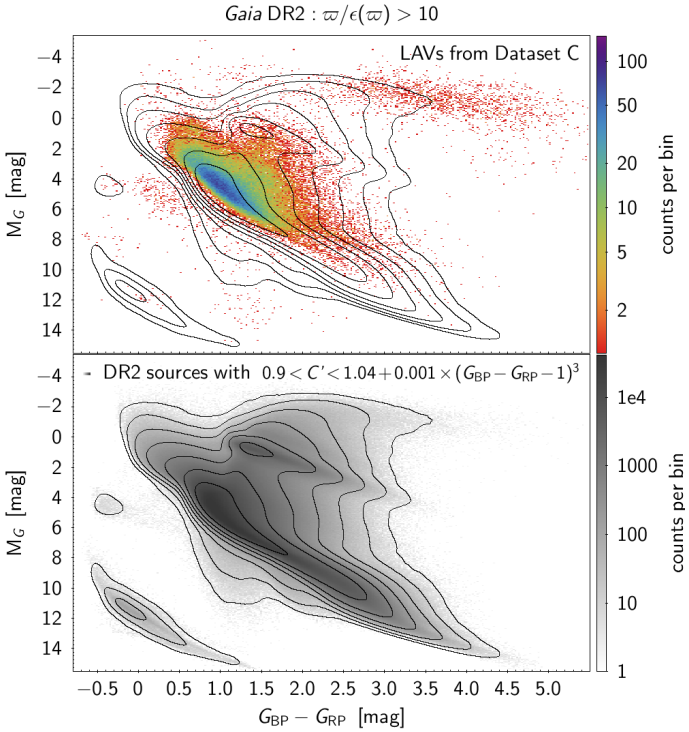

Fig. 22. Observational HRDs of *Gaia* DR2 sources that have relative parallax uncertainties better than 10%. *Top panel*: density map of LAV candidates from Dataset C, while *bottom panel*: random selection of variable+constant sources with good BP and RP flux excess. The black contour lines in both panels correspond to the density lines of the sample shown in the *bottom panel*. No correction for interstellar reddening and extinction is applied.

the derivation of the astrometric solution as less than 1% of Datasets B and C that have parallax uncertainties better than 10% have fewer than eight such periods, while 82% of them have at least ten periods. We will use either Dataset B or C, as needed. We recall from Sect. 3.1 that Dataset B is suitable when colours are required, while Dataset C is required when G_{BP} and G_{RP} variability amplitudes are used. The subsets with good parallaxes in Datasets B and C according to $\varpi/\epsilon(\varpi) > 10$ are hereafter called subsets Bgp and Cgp. The number of their sources is given in Table 1. The distribution of the parallaxes versus $G_{BP}-G_{RP}$ is shown in Fig. 21.

We first give an overview of the datasets in the observational Hertzsprung-Russell diagram (HRD) in Sect. 4.2.1. We then summarize their multi-band variability properties in Sect. 4.2.2, and finally present in Sect. 4.2.3 the distribution in this diagram of the four variability groups identified in the previous section.

4.2.1. Overview

The observational HRD of Subset Cgp is shown in Fig. 22 (top panel), to be compared with the distribution of a random sample of constant+variable DR2 sources in the bottom panel that also have parallaxes better than 10% and good BP and RP flux excess (conditions b2 and b3 in Table 1) (the reader is referred to Gaia Collaboration 2018b, for a detail presentation of this diagram). To guide the eyes, contour lines from the DR2 distribution in the bottom panel are reported on the Cgp distribution in the top panel. Absolute magnitude in this and following observational HRDs is computed with $M_G = G + 5 - 5\ \log_{10}(1000/\varpi)$. The comparison between the two panels shows a potential shortage of LAV sources from Dataset Cgp in some parts of the diagram. This can be due to a real shortage of LAVs in a specific region of the diagram, such as for stars in the red clump around $(G_{BP}-G_{RP}, M_G) \simeq (1.4, 1)$ mag. Or it can be due to smaller statistics in Subset Cgp (~85 000 sources) compared to the DR2 sample (~67 million sources), combined with the parallax-limited selection. This may explain the shortage of blue MS LAVs. However, it can also be a selection effect resulting from the filters leading to Dataset C (like filters c3 to c5 in Table 1). This is expected to be the case for the shortage of LAVs among low-mass M-type MS stars (red dwarfs at $M_G \simeq 9-14$ mag). M0–M5.5 dwarf stars are known to be photometrically variable with flare amplitudes that can reach the order of 1 mag (e.g., Günther et al. 2020), which fall in the amplitude range considered in this paper. A fraction of these stars could be detected with *Gaia* (see Distefano & Lanzafame 2020, for a candidate identified in DR2). However, the faint magnitudes of these stars combined with their red colours lead to the exclusion of the majority of them from Datasets B and C. Most of these excluded sources also have Renormalized Unit Weight Errors (RUWE) larger than 1.4 (see Fig. B.8).

Except for these faint sources, subsets Cgp and Bgp provide a reliable picture of LAVs in the sample with parallax uncertainties better than 10%. These sources reach distances up to 5 kpc at $G_{BP}-G_{RP} \simeq 0.6$ mag (see Fig. 21), while limited to ~1 kpc for the reddest and bluest LAVs in the samples.

4.2.2. Multi-band variability properties in the observational HRD

A summary picture of the variability of LAVs across the observational HRD is shown in Fig. 23, where each cell of size $[\Delta(G_{BP}-G_{RP}), \Delta M_G] = (0.045, 0.12)$ mag has been colour-coded according to either the mean value of $A_{\mathrm{proxy},G}$ (left panel, using Dataset B), or the mean value of $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}}$ (right panel, using Dataset C). The largest variability amplitudes in G (red areas in the left panel) are mainly observed for LPVs (bright red side of the diagram), Cepheids (bright side of the classical instability strip), RR Lyrae variables (in the instability strip close to the MS), eclipsing binaries (on the MS), and variables in the hot subdwarf region of the HRD (bluewards of the MS). The regions of pre-main sequence stars redwards of the MS and between the MS and WD sequence also contain cells with mean $(A_{\mathrm{proxy},G}) > 0.13$.

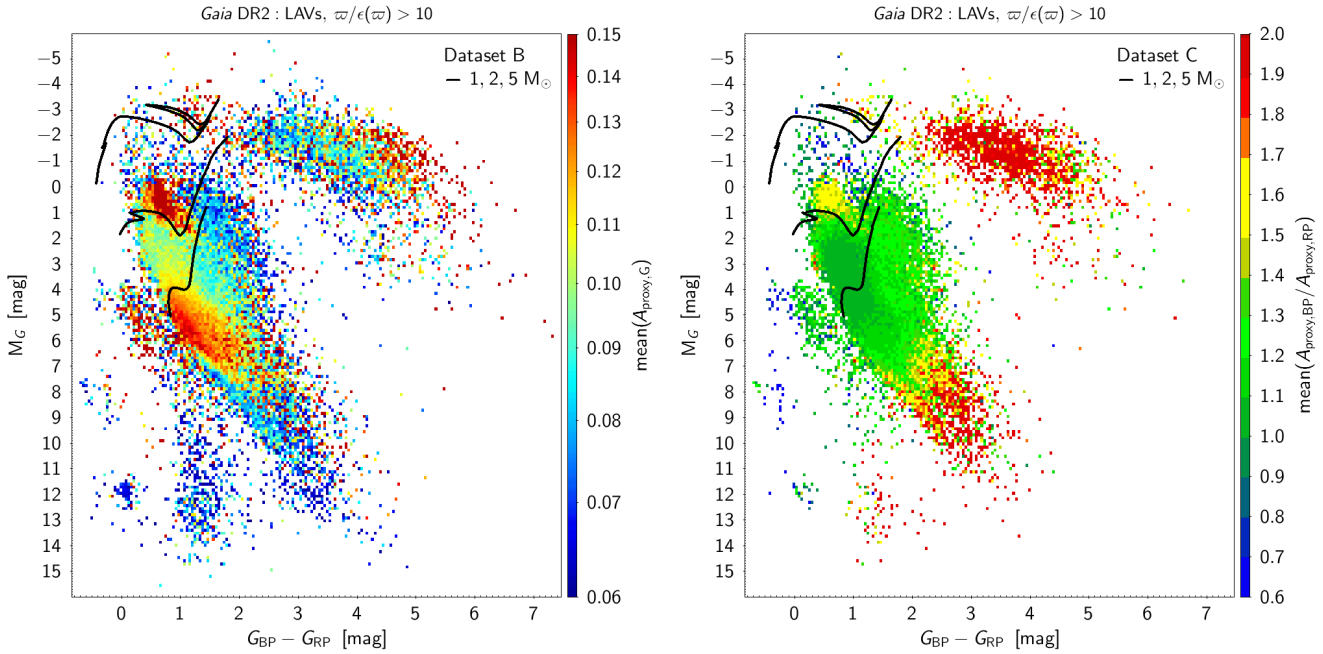

Fig. 23. Observational HRD with the mean value of $A_{\mathrm{proxy},G}$ from Subset Bgp (*left panel*) and of $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}}$ from Subset Cgp (*right panel*) for each cell of size $[\Delta(G_{\mathrm{BP}}-G_{\mathrm{RP}}), \Delta M_G] = (0.045, 0.12)$ mag, plotted in colour according to the colour-scale on the right of each panel. The thin contour lines in black correspond to the density lines of the DR2 sample of constants+variables shown in Fig. 22 (bottom panel). The thick lines correspond to evolutionary tracks of (from bottom to top) 1, 2, and 5 $M_\odot$ solar-metallicity stellar models from Ekström et al. (2012).

The left panel of Fig. 23 is advantageously put in perspective with Fig. 9 of Gaia Collaboration (2019), which plots the distribution of DR2 variability amplitudes across the observational HRD. The selection criteria of the sources displayed in that paper are, however, not the same as here, leading to different patterns when comparing the two figures. In particular, the selection used by Gaia Collaboration (2019) excludes the largest-amplitude variables with range$(G) \gtrsim 0.75$ mag (their exclusion filter $\varepsilon(I_G)/I_G > 0.02$ is equivalent, with a mean number of 130 CCD measurements, to an exclusion of sources with $A_{\mathrm{proxy},G} \gtrsim 0.23$). Another difference is the absence, in their Fig. 9, of LAV subdwarfs observed in our Fig. 23 around $(G_{\mathrm{BP}} - G_{\mathrm{RP}}, M_G) = (0.5, 5.5)$ mag. This absence in their figure is due to their additional selection criteria, mainly on astrometry.

Large-amplitude variables that have the largest blue-to-red amplitude ratios (red areas in the right panel of Fig. 23) mainly consist of LPVs. Two other areas in the observational HRD also display large mean $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}}$ ratios, one at the faint side of the MS and another one in the faint region between the MS and the WD sequence. Caution must however be taken for these faint red sources, as the $A_{\mathrm{proxy,BP}}$ values most probably result from noise in the G_{BP} light curves and are thus not reliable. The second largest $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}}$ ratios in the right panel of Fig. 23 are found for classical pulsators in the instability strip (yellow concentrations in the upper MS and Cepheid region of the diagram). White dwarfs and hot subdwarf variables, on the other hand, have the smallest blue-to-red amplitude ratios, with $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}} < 1$ (blue areas in Fig. 23, right panel). Few bright LAV candidates in the upper MS, or redwards of it, also show $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}} < 0.9$.

4.2.3. Properties of the four classification groups in the observational HRD

The LAVs in Dataset C have been categorized in Sect. 4.1.2 into four groups according, mainly, to their blue-to-red amplitude ratio and their colour. Here we further analyse their properties using the observational HRD. Their distributions in that diagram are shown in Fig. 24. Stellar evolutionary tracks at solar-metallicity are also added from Ekström et al. (2012)[8] to evaluate the stellar masses and evolutionary stages. The considered sample with good parallaxes contains 2033 LAVs in Group 1 (sources with $G_{\mathrm{BP}}-G_{\mathrm{RP}} > 1.8$ mag in the top panel of the figure), 59 in Group 2 (sources with $G_{\mathrm{BP}}-G_{\mathrm{RP}} < 0.2$ mag in the top panel), 3531 in Group 3 (second panel from top) and 79 074 in Group 4 (third panel from top) LAV candidates, with 6154 sources in Subgroup 4a (bottom panel).

Group 1 LAVs populate the red part of the HRD as expected for LPVs (top panel in Fig. 24). We note in Fig. 21 that the redder an LPV is, the less far from the Sun it can be detected in subsets Bgp and Cgp. This is due to the combined effect of redder LPVs being fainter, and of fainter stars having less precise parallaxes.

Group 2 LAVs are not numerous. Their location in the observational HRD indicates that they contain hot subdwarfs and white dwarfs (top panel in Fig. 24), in agreement with the definition of the group.

Group 3 LAVs are expected from the analyses of literature data presented in Sect. 4.1 to predominantly contain pulsating stars. This is confirmed from their distribution in the observational HRD (second panel from top in Fig. 24), where most of them are seen to gather in the region of RR Lyrae stars around $(G_{\mathrm{BP}}-G_{\mathrm{RP}}, M_G) \simeq (0.51, 0.5)$ mag. A tail extending from that region towards the faint-red side of the HRD, down to $(G_{\mathrm{BP}}-G_{\mathrm{RP}}, M_G) \simeq (1.8, 5)$ mag, is also observed, compatible with RR Lyrae stars reddened by extinction on the line of sight. Two other classical pulsators are also visible in the diagram: δ Sct stars extending below the bulk of RR Lyr stars at

[8] Downloaded from https://www.unige.ch/sciences/astro/evolution/en/database/syclist/. The transformation relations used in the website to derive photometry in the *Gaia* bands are taken from Evans et al. (2018).

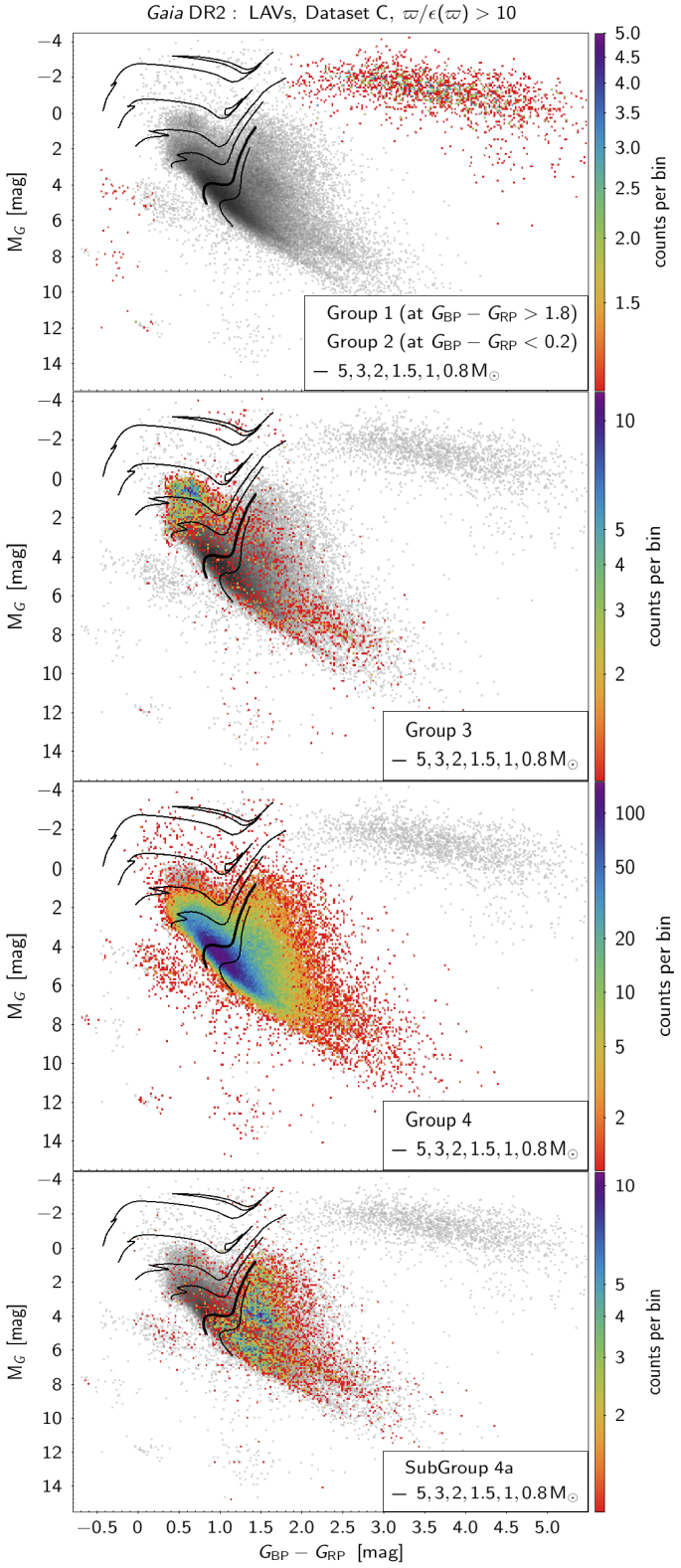

Fig. 24. Density maps of observational HRDs of Group 1 (*top panel*, sources having G_{BP}−$G_{RP} > 1.8$ mag), Group 2 (*top panel*, sources having G_{BP}−$G_{RP} < 0.2$ mag), Group 3 (*second panel from top*), Group 4 (*third from top*) and Subgroup 4a (*bottom panel*) of LAVs in dataset C with parallax uncertainties better than 10%. The background grey points show the full sample of dataset C with $\varpi/\varepsilon(\varpi) > 10$. Evolutionary tracks of (from bottom to top) 0.8, 1, 1.5, 2, 3, and 5 $M_\odot$ solar-metallicity stellar models from Ekström et al. (2012) are over-plotted in black, with the 1 $M_\odot$ track rendered in thick line. The axes ranges have been limited for better visibility.

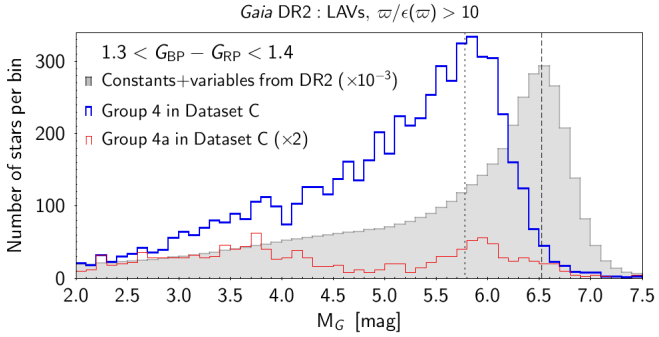

Fig. 25. Histograms of the absolute G magnitude of sources with parallax uncertainties better than 10% in the colour bin $1.3 < G_{BP}$−G_{RP}/mag < 1.4, from Group 4 (blue thick histogram) and Subgroup 4a (red thin histogram) in Dataset C, and from the constant+variable DR2 sample (filled grey histogram). Bins are 0.1 mag wide, and the numbers of sources per bin have been multiplied by two for Subgroup 3a and by 10^{-3} for the full DR2 sample. The vertical dashed line locates the absolute magnitude at maximum of the full DR2 distribution (6.525 mag). A vertical dotted line is added at an absolute magnitude 0.75 mag brighter than the dashed line. The abscissa range has been limited to highlight the distribution of main-sequence stars.

$1 \lesssim M_G/\text{mag} \lesssim 3$, and Cepheids at the bright side of the HRD at $-3 \lesssim M_G/\text{mag} \lesssim -1$. Group 3 also contains a small fraction of variables that are not classical pulsators, as witnessed by the fainter candidates present at $M_G > 5$ mag (Fig. 24, second panel from top). They amount to less than 15% of Group 3.

Group 4 was shown in Sect. 4.1 to predominantly contain non-pulsating LAVs. In particular, the analysis of ZTF periodic variables in Sect. 4.1.1 showed the quasi-achromaticity of the majority of their eclipsing binaries (see in particular Fig. 15). A query in the SIMBAD database confirms this expectation, with three quarters of Group 4 LAVs in Subset Cgp being classified as eclipsing binaries. This is also consistent with the distribution of subset Cgp in the observational HRD (top panel of Fig. 22) when compared to the distribution of constant+variable stars shown in the bottom panel of that figure. They reveal (top panel) a lack of sources close to the zero-age MS, as expected if they are composed of binary stars of similar masses (required for near-achromatic variability). Figure 25 quantifies this observation for Group 4 stars by comparing their M_G histogram in a given colour range (taking $1.3 < G_{BP}$−G_{RP}/mag < 1.4, blue histogram) with that of constant+variable DR2 stars in the same colour range (filled grey histogram). The histogram of MS dwarf stars in the latter sample peaks at $M_G = 6.525$ mag (vertical dashed line in Fig. 25), while that of Group 4 LAVs peaks at a magnitude almost 0.75 mag brighter than this value (dotted line in Fig. 25), as expected if they are composed of equal-mass eclipsing binaries. In addition to eclipsing binaries, various other types of variables are present in Group 4, as witnessed from their distributions in the observational HRD (third panel from top in Fig. 24). A comparison with Figs. 3–7 of Gaia Collaboration (2019), derived from what is known in the literature, is most instructive for their identifications.

Subgroup 4a provides additional insight on Group 4 LAVs that display chromatic variability. The ZTF sample of periodic variables already identified RS CVn and BY Dra variables among non-pulsating variables with $A_{\text{proxy,BP}}/A_{\text{proxy,RP}} > 1.2$ (Sect. 4.1.1). The distribution of Subgroup 4a in the observational HRD (bottom panel of Fig. 24) provides additional clues on the content of this subgroup. It reveals, in particular, the presence of a significant population of potential LAVs in

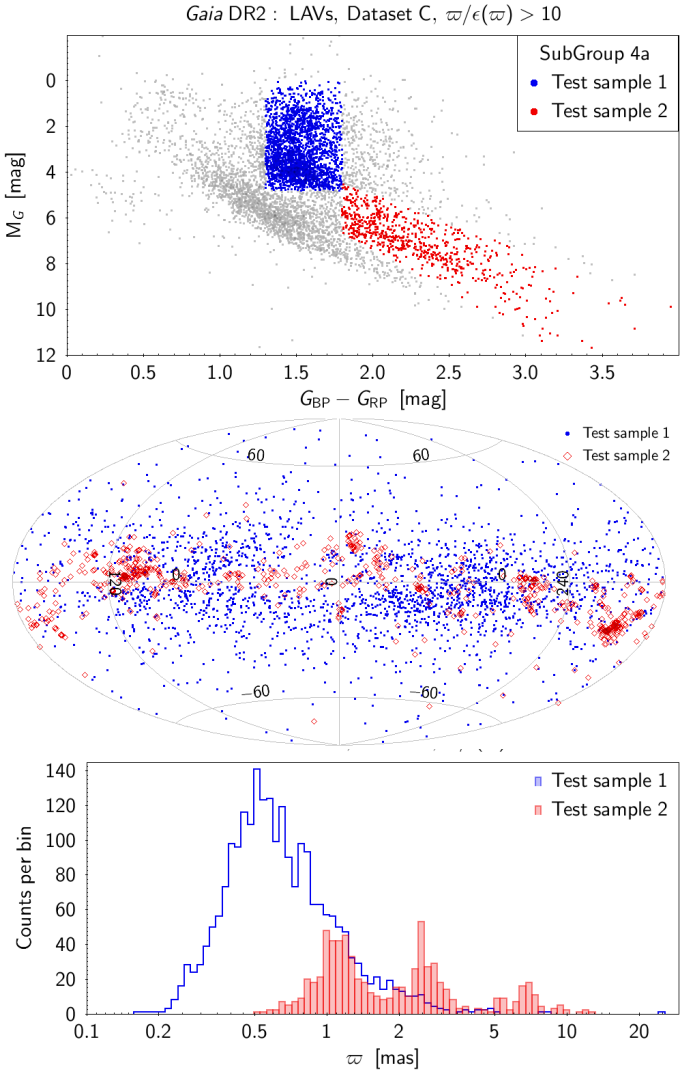

Fig. 26. Properties of two Subgroup 4a samples selected from their locations in the observational HRD as shown by the blue and red samples in the *top panel*. Their sky distributions are shown in the *middle panel*, and the histograms of their parallaxes are shown in the *bottom panel*.

the region of the diagram between the MS and the red clump (around $M_G = 4$ mag and $G_{BP}-G_{RP} = 1.5$ mag), as well as on a sequence at $G_{BP}-G_{RP} \gtrsim 1.8$ mag almost parallel to the MS and about two magnitudes brighter than it. To check their nature, we selected a sample of each of these two populations from their location in the observational HRD as shown in the top panel of Fig. 26. The first sample, shown in blue in the figure (at $G_{BP}-G_{RP} \simeq 1.5$ mag), is found to be distributed preferentially along the Galactic plane but with no obvious specific pattern (second panel from top). They lie at distances between 0.3 and 3 kpc from the Sun (bottom panel). Querying the Simbad database returns 408 crossmatches, with the four most identified types being RS CVn (46 candidates), rotational variables (29), eclipsing binaries (23), and LPVs (15). These variability types should be confirmed as we do not expect to find LPVs in this region of the diagram. We note that 212 crossmatches in this sample have unidentified or uncertain Simbad variability types. The second sample, on the other hand, shown in red in Fig. 26 (top panel), is mainly distributed on the Gould Belt with a predominance in star forming regions (middle panel). Its parallax distribution (bottom panel) confirms that its members belong to nearby star forming regions located at specific distances from the Sun. Among the 318 Simbad crossmatches of this sample, 56 are classified as YSOs, 42 as T Tauri stars, 38 as variables of unspecified type in Orion, and 17 as emission-line stars; 130 stars of the crossmatches have no or uncertain Simbad classification. Finally, Fig. 24 (bottom panel) shows that Subgroup 4a also contains a variety of other variability types, such as LAVs on the MS (compatible with most of them being eclipsing binaries as suggested by their M_G distribution shown in red in Fig. 25), as well as some hot subdwarfs and CVs.

5. Summary and conclusions

We have presented a catalogue of 23 315 874 LAVs from *Gaia* DR2 having peak-to-peak G amplitudes larger than about 0.2 mag, selected from their amplitude proxy $A_{\mathrm{proxy},G} > 0.06$ (Sect. 3). The full catalogue of sources is called Dataset A.

We identified two sub-samples, summarized in Table 1. Dataset B (~5% of Dataset A) is suitable for studies requiring G_{BP} and G_{RP} magnitudes, such as studying colour-magnitude diagrams. Dataset C (about half of Dataset B) is suited for multi-band variability studies involving the amplitude proxies $A_{\mathrm{proxy,BP}}$ and $A_{\mathrm{proxy,RP}}$ in G_{BP} and G_{RP}, respectively.

Within the magnitude and amplitude range considered in this paper, the completeness of Dataset A relative to the variables published in dedicated catalogues in *Gaia* DR2 is close to 100% (Sect. 3.2.1), while the completeness of Datasets B and C are ~70% and ~47%, respectively (Table 2). Comparison with the ZTF catalogue of periodic variables, on the other hand, suggests a completeness factor of 67% for Dataset A (Sect. 3.2.3). It also confirms the above reduction factors from Dataset A to Datasets B and C. The purity levels, on the other hand, are estimated in Sect. 3.3 to increase from less than 50% in Dataset A to ~70% in Dataset B and ~85% in Dataset C.

The power of *Gaia* to study variable stars using quasi-simultaneous multi-band photometry has been illustrated in Sect. 4 with two example cases. The first studied the blue-to-red variability amplitude ratio for different types of variable stars based on literature source identification using *Gaia* DR2 variables, ZTF, and Simbad (Fig. 15). The full Dataset C was then classified into four groups based on, mainly, $A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}}$ and $G_{BP}-G_{RP}$. The main types of variables in these groups are, schematically, LPVs in Group 1 with amplitudes more than twice larger in G_{BP} than in G_{RP}, blue compact objects in Group 2 with amplitudes smaller in G_{BP} than in G_{RP}, pulsators in the classical instability strip in Group 3 with $A_{\mathrm{proxy,BP}} \simeq 1.63\,A_{\mathrm{proxy,RP}}$, and a variety of LAVs with $A_{\mathrm{proxy,BP}} \simeq A_{\mathrm{proxy,RP}}$ in Group 4; the last group mainly consists of non-pulsating variables, but about ten percent (Subgroup 4a) have chromatic variability with $1.2 \lesssim A_{\mathrm{proxy,BP}}/A_{\mathrm{proxy,RP}} \lesssim 1.5$. The properties of these four groups have further been investigated in Sect. 4.2 using sub-samples having parallax uncertainties better than 10%, and examples of additional types of variables populating each group other than the main types just mentioned have been identified from the distributions in the observational HRD (Fig. 24) complemented with type identification using the Simbad database.

The catalogue of LAVs presented here constitutes the first *Gaia* catalogue of LAV candidates extracted from the full public DR2 archive. While it inevitably contains shortcomings inherent to an intermediate data release of such a mission, it provides the opportunity to study variable objects using the samples identified in Datasets A, B, and C, depending on the purpose of the study.

Future data releases will contain additional key *Gaia* results with the provision of data collected from both its RVS and

BP/RP spectrophotometers. These will be especially relevant for the study of LAVs considering that the spectra are taken quasi-simultaneously with the G measurements. The combined photometric+spectroscopic time series will offer unique opportunities to further characterize *Gaia* variable stars.

Acknowledgements. We thank the anonymous referee for her/his comments that led to a significant update of the paper. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This publication makes use of the Starlink Tables Infrastructure Library (Taylor 2005) to produce the figures (STILTS and Topcat). This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France.

References

Abrahams, E. S., Bloom, J. S., Mowlavi, N., et al. 2020, AAS J., submitted [arXiv:2011.12253]
Alcock, C., Allsman, R. A., Alves, D., et al. 1997, ApJ, 486, 697
Arenou, F., Luri, X., Babusiaux, C., et al. 2018, A&A, 616, A17
Belokurov, V., Erkal, D., Deason, A. J., et al. 2017, MNRAS, 466, 4711
Belokurov, V., Penoyre, Z., Oh, S., et al. 2020, MNRAS, 496, 1922
Busso, G., Cacciari, C., Carrasco, J. M., et al. 2018, Gaia DR2 Documentation Chapter 5: Photometry, 5
Chen, X., Wang, S., Deng, L., et al. 2020, ApJS, 249, 18
Clementini, G., Ripepi, V., Molinaro, R., et al. 2019, A&A, 622, A60
Deason, A. J., Belokurov, V., Erkal, D., Koposov, S. E., & Mackey, D. 2017, MNRAS, 467, 2636
Distefano, E., & Lanzafame, A. 2020, Astron. Nachr., 341, 508
Drake, A. J., Djorgovski, S. G., Catelan, M., et al. 2017, MNRAS, 469, 3688
Ekström, S., Georgy, C., Eggenberger, P., et al. 2012, A&A, 537, A146
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Eyer, L. 1998, PhD Thesis, Geneva University, Switzerland
Eyer, L., Rimoldini, L., Rohrbasser, L., et al. 2020, in Proceedings of the Conference Stars and their Variability Observed from Space, eds. C. Neiner, W. W. Weiss, D. Baade, et al., 11
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2018a, A&A, 616, A1
Gaia Collaboration (Babusiaux, C., et al.) 2018b, A&A, 616, A10
Gaia Collaboration (Eyer, L., et al.) 2019, A&A, 623, A110
Günther, M. N., Zhan, Z., Seager, S., et al. 2020, AJ, 159, 60
Heinze, A. N., Tonry, J. L., Denneau, L., et al. 2018, AJ, 156, 241
Holl, B., Audard, M., Nienartowicz, K., et al. 2018, A&A, 618, A30
Iorio, G., Belokurov, V., Erkal, D., et al. 2018, MNRAS, 474, 2142
Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2018, MNRAS, 477, 3145
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2020, MNRAS, 491, 13
Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, A&A, 523, A48
Lanzafame, A. C., Distefano, E., Messina, S., et al. 2018, A&A, 616, A16
Mowlavi, N., Lecoeur-Taïbi, I., Lebzelter, T., et al. 2018, A&A, 618, A58
Mowlavi, N., Trabucchi, M., & Lebzelter, T. 2019, https://zenodo.org/record/3269780
Muciek, M., North, P., Rufener, F., & Gertner, J. 1985, Acta Astron., 35, 377
Palanque-Delabrouille, N., Afonso, C., Albert, J. N., et al. 1998, A&A, 332, 1
Rimoldini, L. 2014, Astron. Comput., 5, 1
Rimoldini, L., Holl, B., Audard, M., et al. 2019, A&A, 625, A97
Roelens, M., Eyer, L., Mowlavi, N., et al. 2018, A&A, 620, A197
Sesar, B., Hernitschek, N., Mitrović, S., et al. 2017, AJ, 153, 204
Shappee, B., Prieto, J., Stanek, K. Z., et al. 2014, Am. Astron. Soc. Meet. Abstr., 223, 236.03
Taylor, M. B. 2005, in TOPCAT & STIL: Starlink Table/VOTable Processing Software, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Udalski, A., Kubiak, M., & Szymanski, M. 1997, Acta Astron., 47, 319
Vioque, M., Oudmaijer, R. D., Schreiner, M., et al. 2020, A&A, 638, A21
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9

Appendix A: Dataset A

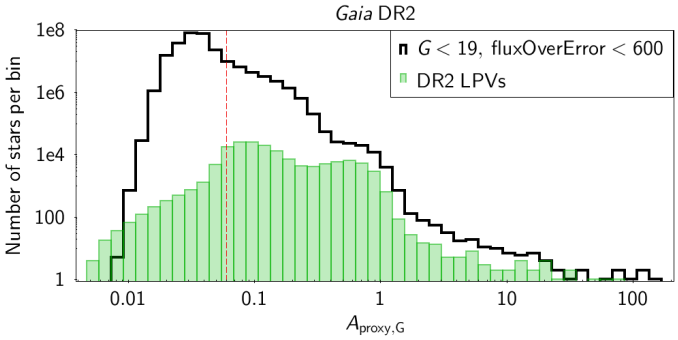

Fig. A.1. Distributions of the variability amplitude proxy of sources in the *Gaia* DR2 archive. The black histogram represents all sources brighter than $G = 19$ mag with mean G flux over error ratios $I_G/\varepsilon(I_G) < 600$. The filled green histogram represents the sources published in the specific DR2 catalogue of LPVs. The vertical red dashed line locates the $A_{\mathrm{proxy},G} = 0.06$ limit used to select large-amplitude variable candidates in this study.

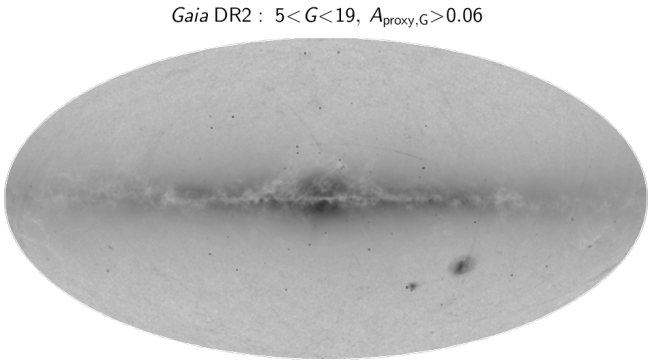

Fig. A.2. Sky distribution in Galactic coordinates of *Gaia* DR2 sources with $G < 19$ mag, $A_{\mathrm{proxy},G} > 0.06$, non-NULL parallaxes, and which have at least five measurements in both the *BP* and *RP* bands.

We detail in this appendix the procedure used to extract and filter the sample of large-amplitude variable candidates from the full *Gaia* DR2 archive. The extraction is described in Appendix A.1. Appendix A.2 then details the removal of spurious cases on specific stripes on the sky due to bad time intervals, and Appendix A.3 the removal of faint sources with large variability amplitudes. They correspond to filters a1 and a2 mentioned in Table 1 of the main body of the article. Some properties of the resulting dataset, called Dataset A, is given in Appendix A.4.

A.1. Extraction from the Gaia DR2 archive

The quantity $A_{\mathrm{proxy},G}$ is not available in the *Gaia* archive, and thus cannot be used to select sources. Instead, we use $I_G/\varepsilon(I_G)$ which is indexed in the archive. We import all sources with $G < 19$ mag and $I_G/\varepsilon(I_G) < 600$, which amount to 256 633 579 sources. The histogram of their $A_{\mathrm{proxy},G}$ is shown by the black line in Fig. A.1. Keeping only sources with $A_{\mathrm{proxy},G} > 0.06$, we are left with 23 830 862 candidates. As as comparison, the $A_{\mathrm{proxy},G}$ distribution of the LPV candidates published in DR2, which all have the 5–95% quantile range $\mathrm{QR}_5(G) > 0.2$ mag, is also shown in Fig. A.1 (green filled histrogram). It is noted that a very small (we note the logarithmic scale of the figure) fraction of these variables have $A_{\mathrm{proxy},G} < 0.06$ despite their large $\mathrm{QR}_5(G)$ amplitude, and are missed in the present sample.

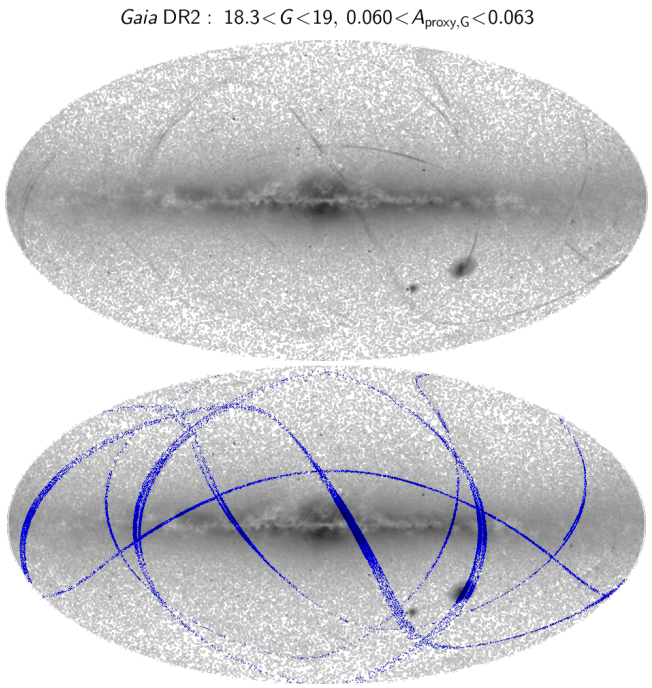

Fig. A.3. Same as Fig. A.2, but limited to the sub-sample with $0.06 < A_{\mathrm{proxy},G} < 0.063$ and $18.3 < G/\mathrm{mag} < 19$. The time-interval-limited stripes identified with the *Gaia* nominal scanning law as containing bad measurements are shown in blue in the *bottom panel*.

The above procedure using $I_G/\varepsilon(I_G) < 600$ retrieves correctly all sources with $A_{\mathrm{proxy},G} > 0.06$ if $N_G < 1297$. However, sources with $I_G/\varepsilon(I_G) > 600$ (i.e. with very small relative uncertainties on their mean G flux) may also have $A_{\mathrm{proxy},G} > 0.06$ if they have $N_G > 1297$. There are only 282 such sources, located in the north and south ecliptic poles. Their high number of observations results from the Ecliptic Pole Scanning Law used during the *Gaia* commissioning phase. They are added to the sample separately, which reaches 23 831 144 candidates.

Finally, we remove from the sample all sources brighter than 5.5 mag in G to comply with the limits taken for our catalogue (see Sect. 3 in the main body of this article). The total number of large-amplitude variable candidates with $A_{\mathrm{proxy},G} > 0.06$ in the magnitude range 5.5 mag $< G < 19$ mag amounts to 23 830 345 in this initial sample.

A.2. Filter on bad time intervals

The sky distribution of the sample constructed so far is displayed in Fig. A.2. It reveals stripes across the sky that are unphysical. The stripes are clearly visible if we display the sub-sample with $A_{\mathrm{proxy},G} < 0.063$ and $G > 18.3$ mag. This is shown in the top panel of Fig. A.3. These unphysical stripes originate from bad measurements at specific times during the mission which escaped the filters that were applied at the time of processing, sometimes coinciding with non-nominal satellite configurations but often for yet unpublished technical reasons. They provide additional noise in the time series of faint sources during those short time intervals. Given the *Gaia* scanning law, the sources affected during these time intervals are distributed on stripes in the sky.

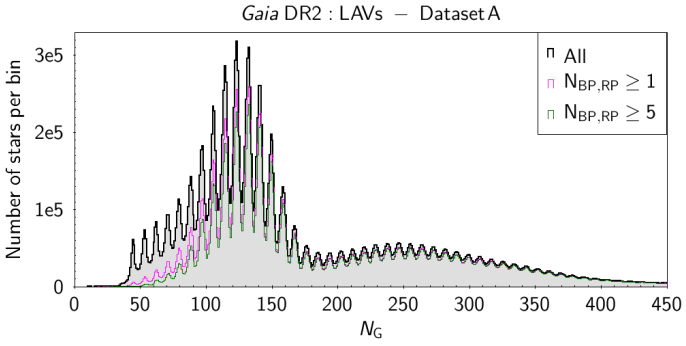

Fig. A.4. Number of CCD observations in G of Dataset A (filled grey histogram with black contour). The histogram is limited to $N_G \leq 450$ for better visibility, the maximum encountered number of CCD measurements being 2189, with a mean at 176. The thin magenta/green histograms show the number of CCD observations in G for the subsets of candidates having at least one/five measurements in both $G_{\rm BP}$ and $G_{\rm RP}$, respectively.

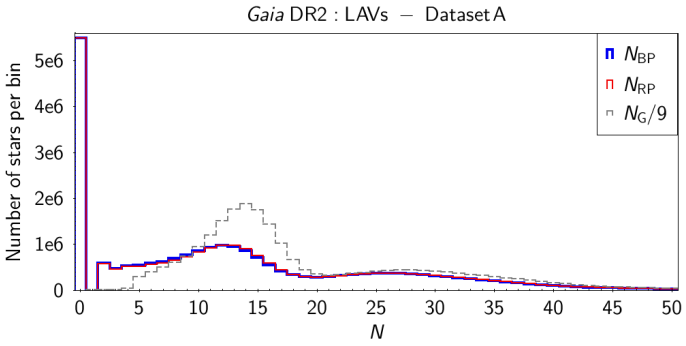

Fig. A.5. Number of observations in $G_{\rm BP}$ (solid blue) and $G_{\rm RP}$ (solid red) of Dataset A. The histograms are limited to $N_{\rm BP,RP} \leq 50$ for better visibility, the maximum encountered number of measurements in $G_{\rm BP}$ and $G_{\rm RP}$ being 235 and 235, respectively. Also shown is the distribution of $N_G/9$ in dashed grey.

Checks of the sky distributions with varying G and $A_{{\rm proxy},G}$ intervals reveal that mainly sources fainter than $G = 18.3$ mag are affected, and that the induced scatter in their light curves does not exceed $A_{{\rm proxy},G} = 0.1$ for the great majority of them. Therefore, we exclude all sources in these stripes that have $A_{{\rm proxy},G} < 0.1$ and $G > 18.3$ mag.

To identify the time ranges that are associated with the stripes, and the sources that are in the stripes, we use the *Gaia* HEALPix Time Extraction tool described in Holl et al. (in prep.). They are shown in blue in Fig. A.3, bottom panel. They contain 1 265 533 sources, of which 514 084 sources have $A_{{\rm proxy},G} < 0.1$ and $G > 18.3$ mag. The amplitude proxy of these latter sources are potentially dominated by noise rather than stellar variability, and are excluded from our sample. After this last filter on sources potentially affected in the identified bad time intervals, our sample contains 23 316 261 large-amplitude candidates.

A.3. Filter faint candidates with large variability amplitudes

Since we are dealing with large amplitude variables, we must ensure, to the best possible way, that none of the epoch measurements in the G time series becomes too faint, or else the faintest epoch measurements will be missed or have too large uncertainties, and the amplitude of the recorded light curve will be affected.

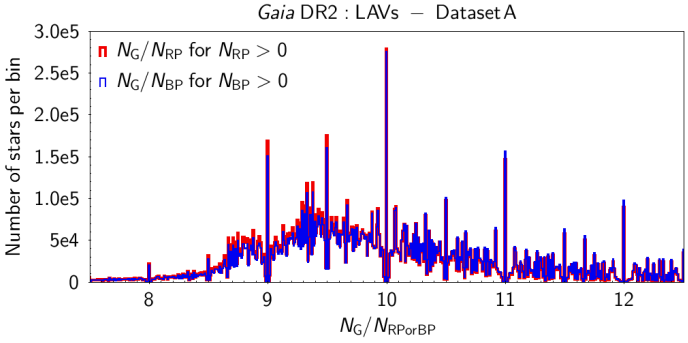

Fig. A.6. Ratio of the number of CCD observations in G to the number of transit observations in $G_{\rm RP}$ (thick red) and $G_{\rm BP}$ (thin blue) for all sources in Dataset A that have non-zero measurements in $G_{\rm RP}$ and $G_{\rm BP}$, respectively. Bins are 0.01 wide.

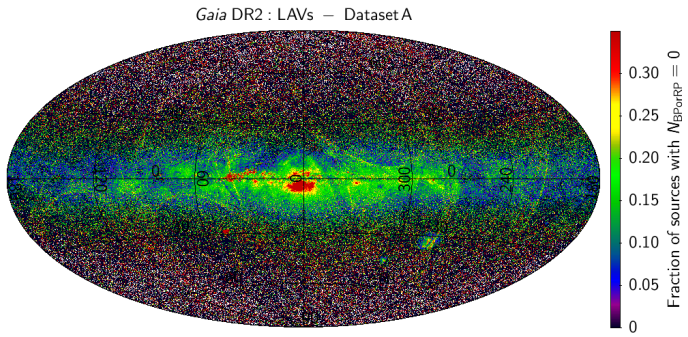

Fig. A.7. Fraction of sources from Dataset A within each level 8 HEALPix tile on the sky that have no observation in either $G_{\rm BP}$ or $G_{\rm RP}$. The fraction is displayed according to the linear colour scale shown on the right of the panel, sources with a fraction larger than 0.35 being rendered in red. The sky is represented in Galactic coordinates.

However, we do not have access to the DR2 epoch photometry for all stars. We therefore use a trick to identify (and exclude) these large-amplitude faint sources, using $A_{{\rm proxy},G}$. If we assume range$(G) \simeq 3.3\,A_{{\rm proxy},G}$ (see Eq. (5) in the main text), we estimate the faintest epoch measurement in the (not-available) time series to be equal to $G + 1.65\,A_{{\rm proxy},G}$. We then keep only sources that have

$$G + 1.65\,A_{{\rm proxy},G} < 20.5. \quad \text{(A.1)}$$

This filtering removes 387 sources from the sample determined so far, leading to a final list of 23 315 874 large-amplitude variable candidates in Dataset A.

A.4. Summary

Here we present some characteristics of Dataset A in complement to the analysis presented in the main body of the article.

(a) Number of measurements. The histograms of the number of good CCD measurements for G is shown in Fig. A.4 for Dataset A. The wiggles observed in the histogram occur at multiples of nine. They reflect the CCD distributions in the *Gaia* focal plane, each transit in the astrometric field going through nine CCDs in six cases out of seven, and through eight CCDs in one case out of seven (Gaia Collaboration 2016).

The number of observations in $G_{\rm BP}$ and $G_{\rm RP}$ are shown in Fig. A.5. About one fourth of the sources have no measurement in $G_{\rm BP}$ and $G_{\rm RP}$ (5 430 305 sources exactly, that is 23%

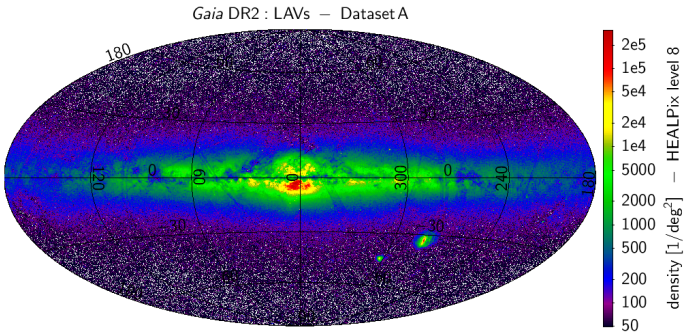

Fig. A.8. Sky density of all LAV candidates in Dataset A (Galactic coordinates). The density of sources is counted within each tile of a HEALPix level 8 sky division, and reported per unit square degree according to the logarithmic colour scale shown on the right of the panel.

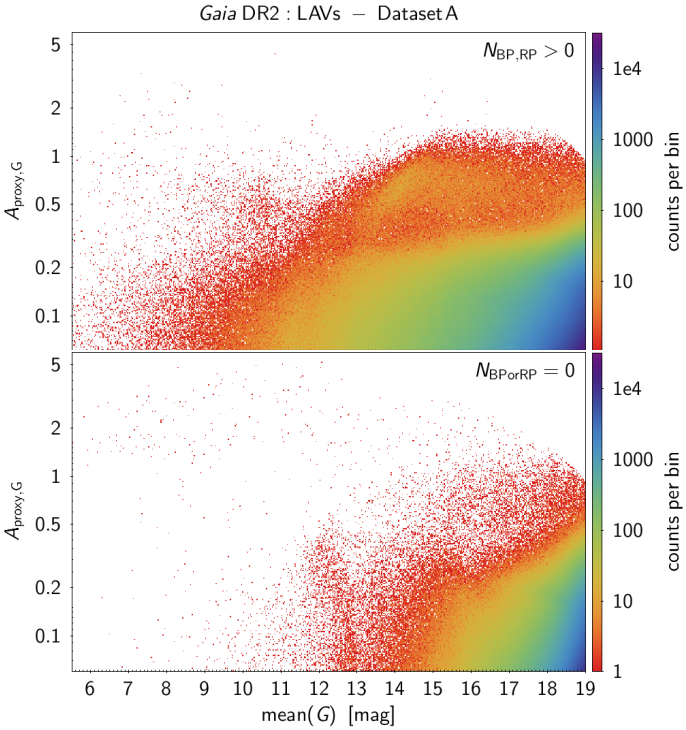

Fig. A.9. Density maps of the amplitude proxy $A_{\mathrm{proxy},G}$ versus mean G magnitude of LAV candidates in Dataset A with non-empty G_{BP} and G_{RP} time series (*upper panel*), or having at least one of the two time series empty (*lower panel*). Both panels have the same colour range, shown on the right of the each panel, giving the number of counts per bin.

of Dataset A), and another 105 025 sources lack measurement in either one of the two bands (about half for each band). Figure A.5 further shows that the number of measurements are about equal in both bands.

The number of CCD observations in G divided by nine, to get an average number of transits that is comparable to the number of observations in G_{BP} and G_{RP}, is also shown in Fig. A.5. It is seen that sources have, in general, fewer transit measurements in G_{BP} and G_{RP} than in G. This is especially true for sources with fewer than 20 transits, where the peaks of the distributions are located at lower values for G_{BP} and G_{RP} than for G. This shortness of G_{BP} and G_{RP} measurements relative to G is confirmed from the distribution of N_G/N_{BP} and N_G/N_{RP} shown in Fig. A.6. These ratios peak between 8.5 and 11 rather than between 8 and 9. This is most probably due to the window assignment and the fact that *BP* and *RP* windows are larger than G windows. This will cause more sources in dense areas not being assigned a BP/RP window. On top of this, in crowded regions, more windows will be truncated, and truncated BP/RP windows have not been included in the DR2 (neither will in DR3) processing. Dense regions on the sky are obviously expected to be most affected by these effects. Figure A.7 confirms this expectation, where the sky regions having the largest fraction of sources without G_{BP} and G_{RP} measurements are located in the densest regions of the sky (shown in Fig. A.8).

(b) Variability amplitude proxy in G. The distribution of $A_{\mathrm{proxy},G}$ versus G for Dataset A is shown in the upper panel of Fig. A.9 for sources that have measurements in G, G_{BP}, and G_{RP}. A subset of sources with $0.3 \lesssim A_{\mathrm{proxy},G} \lesssim 1$ is noticeable, reminiscent of what is expected from Miras. The distribution of sources that lack measurements in at least one of the G_{BP} or G_{RP} time series is also shown, in the bottom panel of Fig. A.9. A density excess is seen around $G \simeq 12.5$ mag for $A_{\mathrm{proxy},G} \lesssim 0.5$, and around $G \simeq 15.5$ mag for $A_{\mathrm{proxy},G} \lesssim 0.3$. These may be spurious variability detections due to photometric calibration issues.

Appendix B: *BP* and *RP* flux excess

Evans et al. (2018) defined the BP and RP flux excess C as

$$C = \frac{I_{\mathrm{BP}} + I_{\mathrm{RP}}}{I_G}. \tag{B.1}$$

Its value should be close to one given the G, G_{BP}, and G_{RP} transmission curves[9] (see Sect. 8 and Figs. 20 and 21 in Evans et al. 2018).

The value of C versus $G_{\mathrm{BP}}-G_{\mathrm{RP}}$ is shown in Fig. B.1 for all sources in Dataset A. The band of well-behaved single sources identified by Evans et al. (2018) in their Fig. 17 is well visible in Fig. B.1. We define it more precisely with the following fiducial BP and RP flux excess function:

$$C_{\mathrm{fid}} = \begin{cases} 1.2 + 0.06\,(G_{\mathrm{BP}}-G_{\mathrm{RP}} - 0.4)^2 & \text{if } (G_{\mathrm{BP}}-G_{\mathrm{RP}} < 0.6), \\ 1.2024 + 0.1\,(G_{\mathrm{BP}}-G_{\mathrm{RP}} - 0.6)^{1.27} & \text{if } (G_{\mathrm{BP}}-G_{\mathrm{RP}} > 0.6). \end{cases} \tag{B.2}$$

The function is shown by the solid line in Fig. B.1. For information, the limits $1 + 0.015 * (G_{\mathrm{BP}}-G_{\mathrm{RP}})^2 < C < 1.3 + 0.06 * (G_{\mathrm{BP}}-G_{\mathrm{RP}})^2$ proposed by Arenou et al. (2018, their Eq. (2)) to select sources with acceptable BP and RP flux excesses are shown by the dashed lines in Fig. B.1.

We now define for each source the normalized BP and RP flux excess C' by

$$C' = \frac{C}{C_{\mathrm{fid}}}. \tag{B.3}$$

The resulting diagram is shown in Fig. B.2, and the histogram of C' is plotted in Fig. B.3. A local minimum is observed in the histogram around $C' \sim 1.04$, suggesting an upper limit around

[9] A combined figure of the calibrated DR2 passbands is provided in the *Gaia* Image of the Week *IoW_20180316* published on the ESA *Gaia* web pages at https://www.cosmos.esa.int/web/gaia/iow_20180316, while the nominal pre-launch version is available in Jordi et al. (2010).

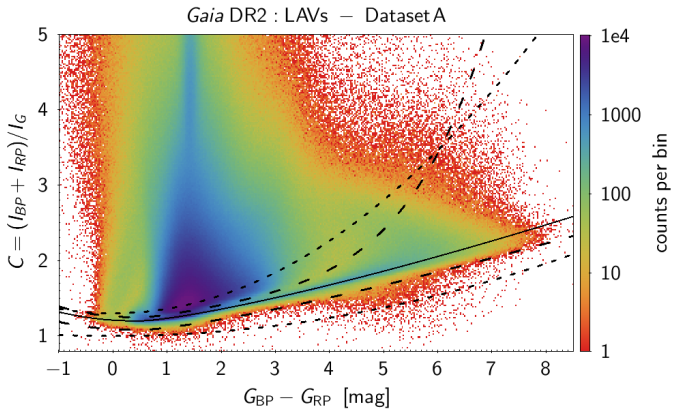

Fig. B.1. Density map of the BP and RP flux excess C (Eq. (B.1)) versus G_{BP}−G_{RP} colour for all sources in Dataset A. The solid line is the function given by Eq. (B.2), while the long-dashed thick lines are the limits given by Eq. (B.4) outside of which G_{BP} and/or G_{RP} are considered to be unreliable relative to G. For information, the limits proposed by Arenou et al. (2018) are shown in short-dashed lines.

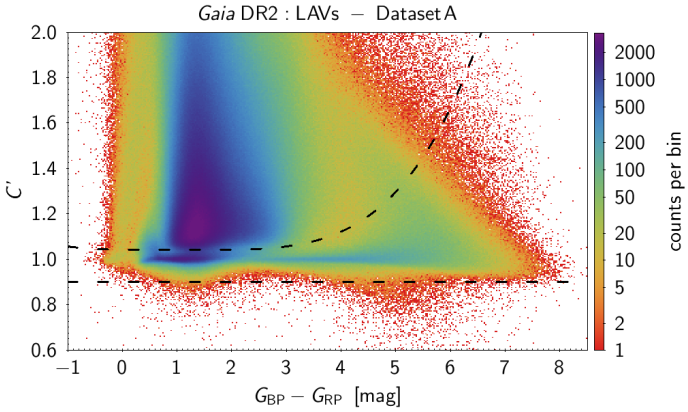

Fig. B.2. Same as Fig. B.1, but for the normalized BP and RP flux excess C' defined by Eq. (B.3). The upper dashed line is the function $1.04 + 0.001\,(G_{BP}-G_{RP}-1)^3$ above which G_{BP} and/or G_{RP} should not be reliable (see text). The lower dashed line is the lower limit at $C' = 0.9$. The ordinate is zoomed compared to Fig. B.1 for better visibility.

this value for well-behaved single sources. We therefore adopt the colour-dependent upper limit $C'_{\rm lim}$ identifying well-behaved sources as $C'_{\rm lim} = 1.04 + 0.001\,(G_{BP}-G_{RP} - 1)^3$. This limit is shown by the upper dashed line in Figs. B.1 and B.2. The characteristics of LAVs with C' values larger than this limit are analysed in the next section. Sources with too low BP and RP flux excesses are then checked in Appendix B.2.

B.1. Large BP and RP flux excesses

Several causes of the large BP and RP flux excesses have been presented in Evans et al. (2018), to which we refer. Here, we check several properties that could be specific to LAVs for the origin of the large BP and RP flux excesses. Indeed, 95% of Dataset A have too large BP and RP flux excesses. For this purpose, we compare a representative subset of these sources, called subset L, with a subset of well-behaved sources around the fiducial line, called subset F. Both subsets are restricted to sources with parallax uncertainties better than 10%. Subset L is defined with $1.11 < C' < 1.3$ and subset F with $C' < 1.02$, such that they

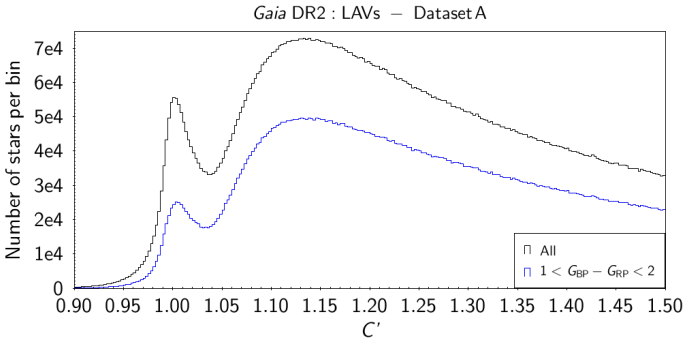

Fig. B.3. Histogram of the normalized BP and RP flux excess for all sources in Dataset A (upper black histogram) and for sources in Dataset A that have G_{BP}−G_{RP} colours between 1 mag and 2 mag (lower blue histogram). The abscissa range has been limited for better visibility.

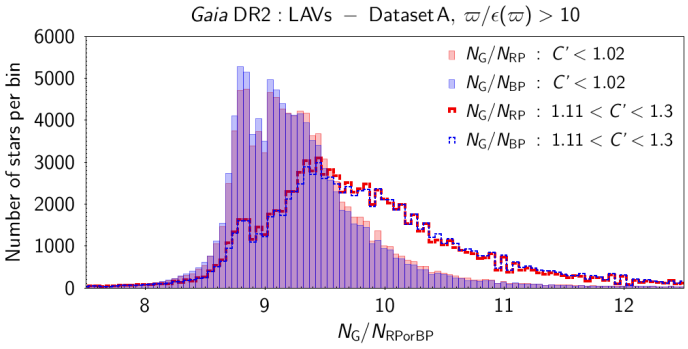

Fig. B.4. Same as Fig. A.6, but for the subsets of Dataset A that have parallax relative uncertainties better than 10% and either $C' < 1.02$ (filled histograms) or $1.11 < C' < 1.3$ (dashed lines). Histograms of G_{BP} are shown in blue and those of G_{RP} are shown in red. Bins are 0.05 wide.

have statistically comparable number of sources, of 99 833 and 97 540, respectively.

Number of G, G_{BP}, and G_{RP} measurements. Because of the large amplitudes of LAVs, non-similar time sampling in G, G_{BP}, and G_{RP} time series may lead to incompatible mean magnitudes in the three bands. The distributions of N_G/N_{BP} and N_G/N_{RP} are shown in Fig. B.4 for subsets F (filled histograms) and L (dashed-line histograms). They reveal, on the mean, slightly larger values of these ratios for subset L than for subset F, pointing to of lack of epoch measurements in G_{BP} and G_{RP} time series relative to G time series. This is most probably due to the fact that these sources are located in dense regions of the sky (see below), leading to more difficult observation conditions in *BP* and *RP* spectrophotometry than in G point-spread photometry. N_G/N_{BP} and N_G/N_{RP} are, however, only 5% to 10% larger in subset L than in subset F. It is improbable that it would be at the origin of the large BP and RP flux excesses observed in subset L.

Location in the observational HRD. The distributions of the two subsets differ in the observational HRD, as shown in Fig. B.5. Subset F populates the observational HRD in the expected regions of the diagram (top panel of Fig. B.5). In contrast, subset L is mainly located in the region of the observational HRD between the MS and WD sequence where we do not expect to have many stars (bottom panel in the figure). Subset L

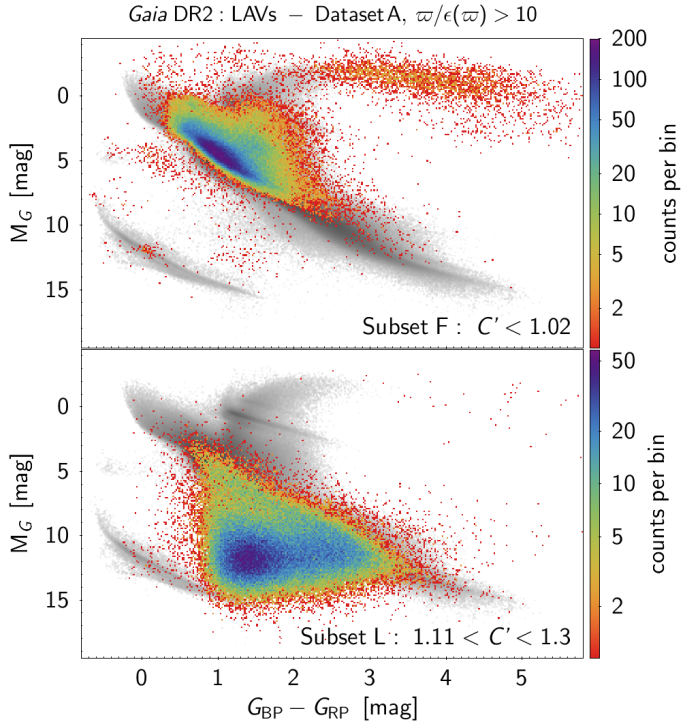

Fig. B.5. Density maps of the observational HRDs of two dataset A sub-samples with relative parallax uncertainties better than 10%: for sources with $C' < 1.02$ in the *top panel* (subset F in the text) and for sources with $1.11 < C' < 1.3$ in the *bottom panel* (subset L in the text). The density maps are plotted on top of *Gaia* DR2 sources (constant and variable) with relative parallax uncertainties better than 2.5% (light grey in the background). Density goes from low in red to high in black on a logarithmic scale.

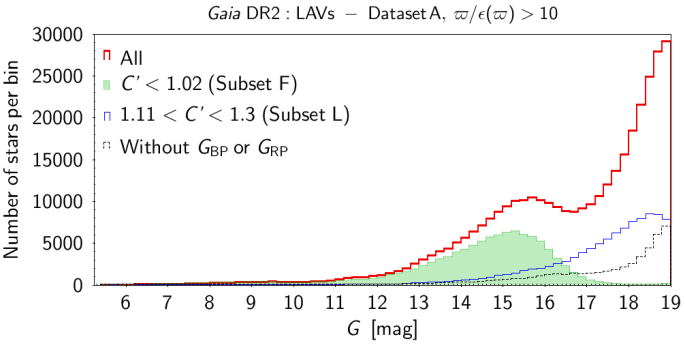

Fig. B.6. Histogram of G magnitude for LAV candidates in Database A that have parallax uncertainties better than 10% (red thickline). Subset F (see text) therein that has $C' < 1.02$ is shown by the filled green histogram, and subset L that has $1.11 < C' < 1.3$ is shown by the thin blue histogram. Also shown in black dashed histogram is the subset that has no G_{BP} and/or G_{RP} in the DR2 archive. Bins are 0.2 mag wide.

sources have also fainter apparent magnitudes, on the mean, than subset F sources (Fig. B.6), with $G \lesssim 17.5$ mag for the majority of subset L (blue histogram), while the bulk of subset F has $12 \lesssim G\,[\mathrm{mag}] \lesssim 17$ (green filled histogram).

Sky distribution. The distributions of the two subsets also differ in the sky, as shown in Fig. B.7. Subset F (top panel) is relatively homogeneously distributed on the sky, with a predominance in the Galactic plane. In contrast, subset L sources (bottom panel) are mainly clumped towards specific regions of the

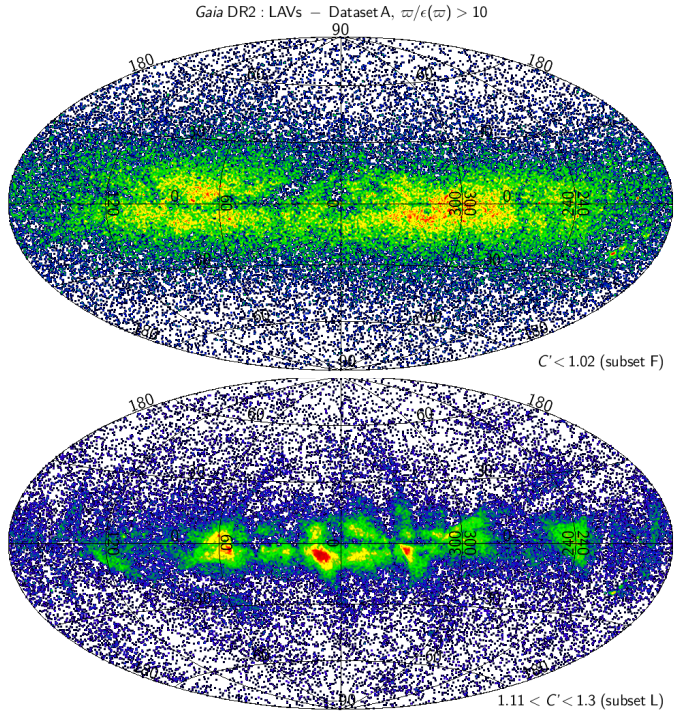

Fig. B.7. Sky density (Galactic coordinates) of LAV candidates in Database A that have parallax uncertainties better than 10% and either $C' < 1.02$ (*top panel*; subset F in the text) or $1.11 < C' < 1.3$ (*bottom panel*; subset L in the text). Density goes from low in black to high in red on a logarithmic scale.

Galactic plane, and predominantly towards the Bulge. Their sky distribution points towards regions with dense regions and with large extinction, which is compatible with them being predominantly faint.

Astrometric solution. The specific properties of subset L, that is being faint, located in dense regions of the sky, and having large BP and RP flux excesses, raise the question of the validity of their astrometric solution. One parameter to check in this respect is the Renormalized Unit Weight Excess (RUWE) from the astrometric solution. It is a goodness of fit parameter that quantifies the departure from an astrometric single star model fit. In this respect, it can be a very useful indicator of astrometric multiplicity in sources (e.g., binaries), as has been demonstrated by Belokurov et al. (2020). These authors selected a high-quality sample of *Gaia* DR2 stars located at high latitudes, with low extinction, and having low BP and RP flux excesses, among other filtering criteria. In the vast majority of cases, however, large RUWE values can also be due to various uncalibrated effects in DR2. Therefore, RUWE is commonly used in DR2 to identify inaccurate astrometric solutions, but at the expense of also removing potential binary sources. In this paper, we do not use RUWE as a criterion to filter the photometric data, given that the relation between the two is not clearly established. We explore this a little more below.

The RUWE is plotted colour-coded in the observational HRD shown in Fig. B.8 for both subsets F (upper panel) and L (bottom panel). Subset F sources (upper panel) have for the great majority of them RUWE valus below 1.4, the limit proposed on the *Gaia* DR2 known issues pages below which the astrometry is reliable. In contrast, almost all subset L sources (lower panel) that are between the MS and the WD sequence have RUWE

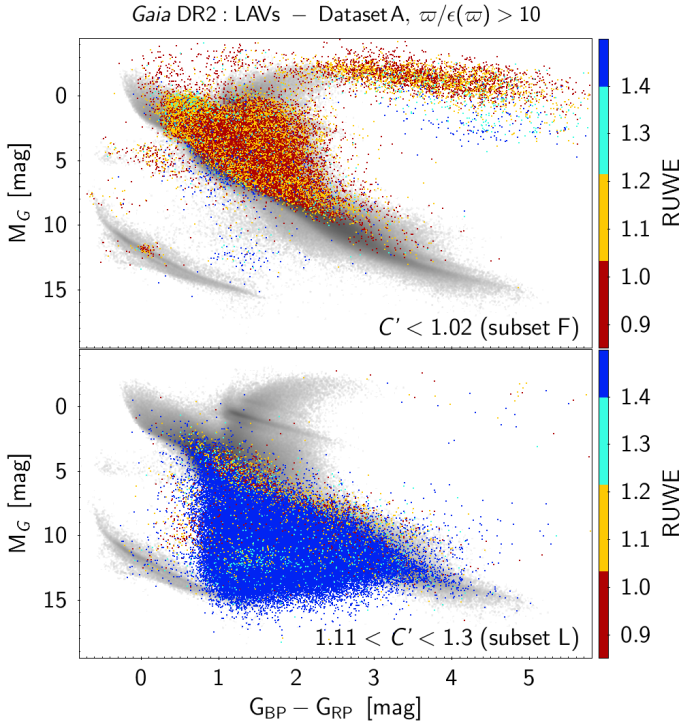

Fig. B.8. Same as Fig. B.5, but colour coded with the value of RUWE according to the colour-scale shown on the right of the figure. RUWE values larger than 1.5 are rendered in blue and values smaller than 0.85 are rendered in red.

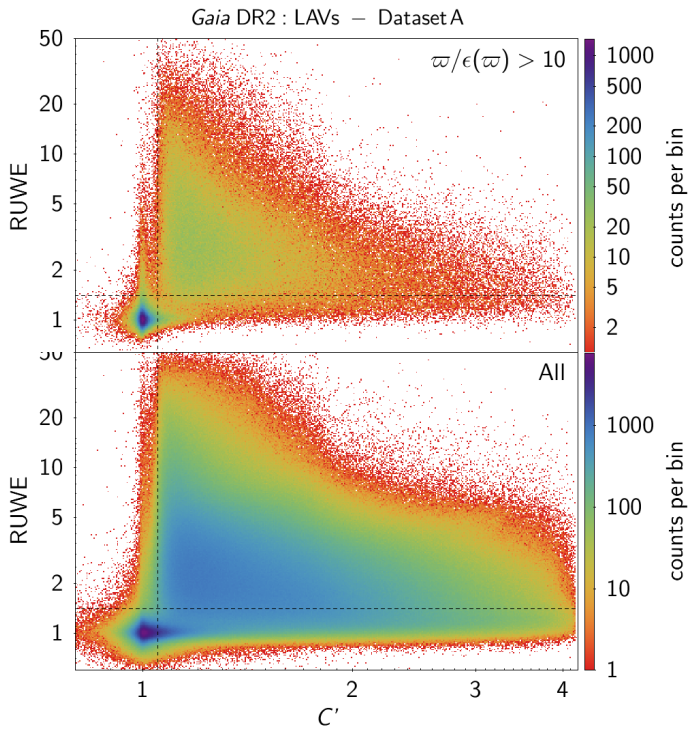

Fig. B.9. Density map of the Renormalized Unit Weight Excess versus normalized BP and RP flux excess for sources in Dataset A that have parallax uncertainties better than 10% (*top panel*). The same figure, but for all sources in Dataset A, is shown in the *bottom panel*. Vertical and horizontal dotted lines have been drawn at $C' = 1.04$ and RUWE = 1.4, respectively, as eye guides. The axes ranges have been limited for better visibility.

values above 1.4, It is interesting in this respect to note that the few subset F sources that are located in that same region of the observational HRD also have large RUWE values. They are reminiscent of the DR2 problematic astrometric cases highlighted in the known issues on the *Gaia* web pages[10].

There is actually a bi-modal distribution in the C' versus RUWE plane, as shown in Fig. B.9, with a first peak around (C', RUWE) = (1,1), and a second concentration of points in the region $C' \gtrsim 1.05$ and RUWE $\gtrsim 1.4$. The origin of this bi-modality is not known. The two quantities could be linked with each other in dense regions. However, RUWE values are also directly impacted by the BP and RP flux excess, because the normalization factor depends on the position of the colour-magnitude diagram.

It must also be noted that large RUWE values do not necessarily imply bad astrometry. Small uncertainties on sky positions may lead to large RUWE values. Caution must thus be taken when imposing a limit on RUWE, as this could remove good astrometric cases. Here, we do not impose any limit on RUWE.

Finally, we checked that there is no specific correlation between RUWE and photometric variability amplitude. This excludes an effect of photometric amplitude on astrometry.

B.2. *Low* BP *and* RP *flux excesses*

A (small) fraction of Dataset A LAVs have too low BP and RP flux excesses compared to the fiducial values, especially for red sources (see Fig. B.2), a feature not observed (at least for red

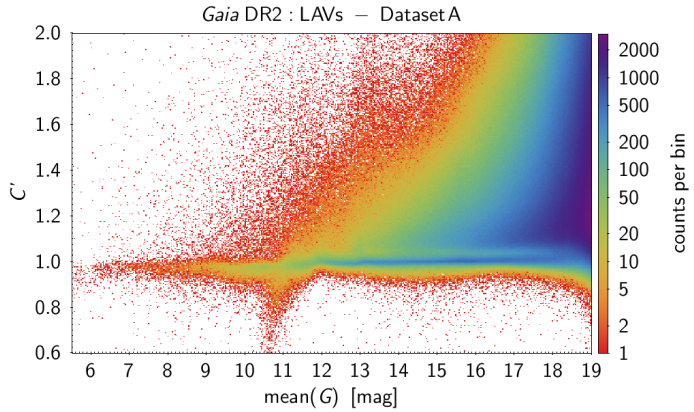

Fig. B.10. Same as Fig. B.2, but versus G magnitude.

stars) in Fig. 17 of Evans et al. (2018). It mainly occurs at magnitudes between $10 \lesssim G\,[\mathrm{mag}] \lesssim 11.5$, as shown in Fig. B.10. Figure B.11, which colour-codes the normalized BP and RP flux excess across the colour-magnitude diagram, reveals that it significantly impacts bright red stars with $G_{BP}-G_{RP} \gtrsim 4$ mag in that magnitude range.

The too-low BP and RP flux excesses imply I_G fluxes that are too high with respect to $I_{BP} + I_{RP}$. We adopt a lower limit of $C' = 0.9$ below which C' is considered unreliable. The number of sources in Dataset A with $C' < 0.9$ is only few thousand (see Table 1). However, they are at the origin of a significant shortage of the reddest LPVs at $G \simeq 11$ mag in Datasets B and C defined

[10] See presentation https://www.cosmos.esa.int/documents/29201/1770596/Lindegren_GaiaDR2_Astrometry_extended.pdf/1ebddb25-f010-6437-cb14-0e360e2d9f09 mentioned in the *Gaia* web page of known issues for DR2 https://www.cosmos.esa.int/web/gaia/dr2-known-issues

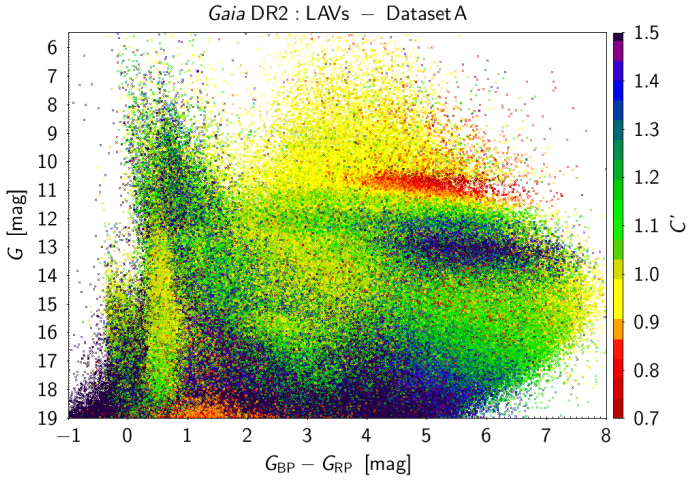

Fig. B.11. Colour-magnitude diagram of Dataset A with the normalized BP and RP flux excess colour-coded according to the colour scale shown on the right of the figure. The colour scale has been limited between 0.7 and 1.5, sources with C' values outside this range being rendered in the colour at the respective end of the scale. Sources with $C' < 0.9$ have been plotted on top of other sources for better visibility as they contain a very small number of sources compared to the size of Dataset A.

in the main text of this article, well visible in CM diagrams like in Fig. 4.

B.3. Summary

In summary, our final condition to select sources with reliable (normalized) BP and RP flux excess is

$$0.9 < C' < 1.04 + 0.001\,(G_{BP}-G_{RP}-1)^3. \tag{B.4}$$

It is used in the construction of Dataset B, which results in the selection of only 5% of LAV candidates from Dataset A to have reliable BP and RP flux excesses. For the 95% remaining candidates, the photometric values are to be taken with care, mainly, but not only (see below), due to poor G_{BP} and G_{RP} quality in DR2.

It must be noted that in the DR2 papers and documentation the BP and RP flux excess has been presented as a quality flag. This parameter, however, simply informs on the consistency between G, G_{BP}, and G_{RP} fluxes. While for DR2 it is probably true that in many cases the BP/RP integrated fluxes are of poorer quality due mostly to lower-resolution background calibration and contamination/blend cases (see known DR2 issues listed on the *Gaia* DR2 website[11]), similar problems (in particular the handling of close pairs and extended sources) will also affect G-band measurements. In the case of extended sources, for instance, the G-band measurements may show larger variations than for BP/RP due to the smaller window and different scan directions, and we know already that there were some misclassifications of extended sources as RR Lyrae in DR2. The red LAVs with too-low BP and RP flux excesses described in Appendix B.2 give another example of cases where G is in fault rather than G_{BP} or G_{RP}. The user should thus keep in mind that, in principle, outliers in the distribution of BP and RP flux excess could be due to problems in any of the bands.

[11] https://gea.esac.esa.int/archive/documentation/GDR2/index.html

Appendix C: Amplitude proxy for BP + RP

In this appendix, we derive the amplitude proxy $A_{\mathrm{proxy,BP+RP}}$ for the summed BP + RP magnitude. The derivation is based on the variances $\sigma^2(f_{BP})$ and $\sigma^2(f_{RP})$ of the f_{BP} and f_{RP} time series and on the covariance $\mathrm{Cov}(f_{BP}, f_{RP})$ between these two time series, defined by

$$\sigma^2(f_{BP}) = \frac{1}{N_{BP}} \sum_i (f_{BP,i} - I_{BP})^2\,, \tag{C.1}$$

$$\sigma^2(f_{RP}) = \frac{1}{N_{RP}} \sum_j (f_{RP,j} - I_{RP})^2\,, \tag{C.2}$$

$$\mathrm{Cov}(f_{BP}, f_{RP}) = \frac{1}{N_{RP\cap BP}} \sum_k (f_{BP,k} - I_{BP})(f_{RP,k} - I_{RP}). \tag{C.3}$$

The variance $\sigma^2(f_{BP} + f_{RP})$ of the summed flux $f_{BP} + f_{RP}$ is then given by

$$\sigma^2(f_{BP} + f_{RP}) = \sigma^2(f_{BP}) + \sigma^2(f_{RP}) + 2\,\mathrm{Cov}(f_{BP}, f_{RP}). \tag{C.4}$$

It must be noted that covariance is usually defined for random variables. In the case of f_{BP} and f_{RP}, they could be considered independent if we neglect the minimal frequency overlap between the two band passes. They can, however, be correlated due, for example, to crowdedness simultaneously affecting G_{BP} and G_{RP}, or to correlated perturbation induced by stray light. Aging of the optic/electronics and the occurence of solar storms, are other examples impacting both G_{BP} and G_{RP}. Besides, the physics behind variable stars leads, in most cases, to a coordinated variability of the flux throughout the optical spectrum. All these effects lead to a non-zero covariance, as shown later in the next section (see in particular Fig. C.1).

C.1. Definition

We define an amplitude proxy $A_{\mathrm{proxy,BP+RP}}$ for the summed BP + RP flux in the same way as we defined the amplitude proxies for f_G, f_{BP}, and f_{RP}:

$$A^2_{\mathrm{proxy,BP+RP}} = \frac{\sigma^2(f_{BP} + f_{RP})}{(I_{BP} + I_{RP})^2}, \tag{C.5}$$

which becomes, using Eq. (C.4),

$$\begin{aligned} A^2_{\mathrm{proxy,BP+RP}} = {} & \frac{I^2_{BP}}{(I_{BP} + I_{RP})^2}\, A^2_{\mathrm{proxy,BP}} \\ & + \frac{I^2_{RP}}{(I_{BP} + I_{RP})^2}\, A^2_{\mathrm{proxy,RP}} \\ & + \frac{2\,\mathrm{Cov}(f_{BP}, f_{RP})}{(I_{BP} + I_{RP})^2}. \end{aligned} \tag{C.6}$$

We now define the last term in Eq. (C.6) as a 'covariance proxy'

$$C_{\mathrm{proxy,Cov(BP,RP)}} = \frac{2\,\mathrm{Cov}(f_{BP}, f_{RP})}{(I_{BP} + I_{RP})^2}, \tag{C.7}$$

and Eq. (C.6) becomes

$$A_{\mathrm{proxy,BP+RP}} = \frac{\sqrt{I^2_{BP}\, A^2_{\mathrm{proxy,BP}} + I^2_{RP}\, A^2_{\mathrm{proxy,RP}} + (I_{BP} + I_{RP})^2\, C_{\mathrm{proxy,Cov(BP,RP)}}}}{(I_{BP} + I_{RP})}. \tag{C.8}$$

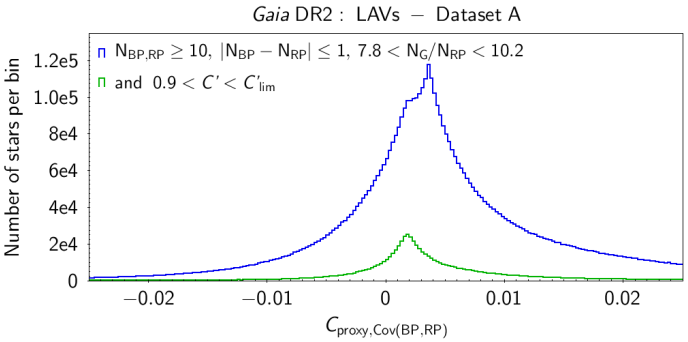

Fig. C.1. Histograms of the covariance proxy $C_{\mathrm{proxy,Cov(BP,RP)}}$ defined by Eq. (C.8) and computed from the information published in *Gaia* DR2 using Eqs. (C.8) and (C.9). The blue (upper) histogram represents the sample of Dataset A satisfying the conditions $N_{\mathrm{BP,RP}} \geq 10$, $|N_{\mathrm{BP}} - N_{\mathrm{RP}}| \leq 1$, and $7.8 < N_{\mathrm{obs}}(G)/N_{\mathrm{obs}}(RP) < 10.2$. The sub-sample therein having reliable BP and RP flux excesses according to condition (B.4) is shown by the green (lower) histogram. Bins are 2.5×10^{-4} wide.

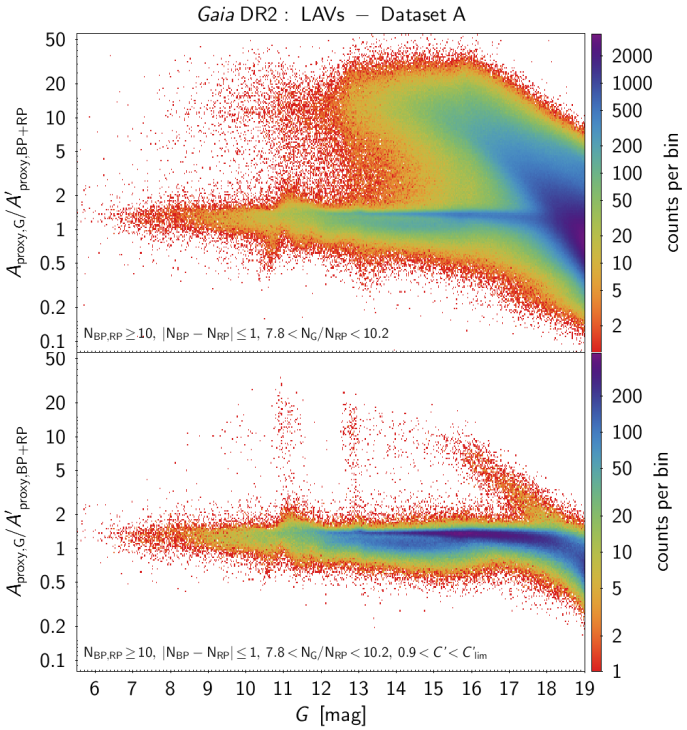

Fig. C.2. Density maps of the ratio between the variability proxy in f_G (Eq. (2)) and the variability proxy in $f_{\mathrm{BP}} + f_{\mathrm{RP}}$ neglecting the covariance between f_{BP} and f_{RP} (Eq. (C.10)), versus mean G magnitude. *Top panel*: map for all sources in Dataset A while *bottom panel*: map for the sub-sample of Dataset A that has reliable BP and RP flux excess according to Eq. (B.4). The ordinate ranges are identical in both panels and have been limited for better visibility.

C.2. Computation

An estimate of the covariance proxy $C_{\mathrm{proxy,Cov(BP,RP)}}$ (Eq. (C.7)) can be obtained under the assumption that

$$A_{\mathrm{proxy,BP+RP}} \simeq A_{\mathrm{proxy},G}. \tag{C.9}$$

This assumption is debatable for cases with bad BP and RP flux excesses, but has the advantage to allow an estimate of $C_{\mathrm{proxy,Cov(BP,RP)}}$ by combining Eqs. (C.8) and (C.9). The histogram of the resulting $C_{\mathrm{proxy,Cov(BP,RP)}}$ values is shown in Fig. C.1 for the subset of Dataset A that has similar number of transit measurements in G, G_{BP}, and G_{RP} for each source (see figure caption), a condition that is necessary for a valid comparison of the properties of f_G, f_{BP}, and f_{RP} time series given the large amplitudes considered here.

Figure C.1 shows a distribution of $C_{\mathrm{proxy,Cov(BP,RP)}}$ centred on a positive value between 0.002 and 0.004. These values are of the same order of magnitude as the values of $A^2_{\mathrm{proxy},G} > 0.0036$ considered in this study. The covariance term is thus not negligible relative to the variances. We indeed expect, for the great majority of variable stars, a concomitant increase (or decrease) in the red and blue filters.

A correct computation of the covariance between f_{BP} and f_{RP} time series should be done directly from the light curves using Eq. (C.3), rather than using Eq. (C.8) with the approximation Eq. (C.9). The flux time series, however, are not available for the majority of sources in *Gaia* DR2. We therefore consider, as an approximation, the variability proxy $A'_{\mathrm{proxy,BP+RP}}$ that neglects the covariance term. It writes

$$A'_{\mathrm{proxy,BP+RP}} = \frac{\sqrt{I^2_{\mathrm{BP}}\, A^2_{\mathrm{proxy,BP}} + I^2_{\mathrm{RP}}\, A^2_{\mathrm{proxy,RP}}}}{(I_{\mathrm{BP}} + I_{\mathrm{RP}})}. \tag{C.10}$$

We have, in general,

$$A'_{\mathrm{proxy,BP+RP}} < A_{\mathrm{proxy,BP+RP}} \tag{C.11}$$

since the covariance is, on the mean, positive.

C.3. Analysis

Since $f_G \simeq f_{\mathrm{BP}} + f_{\mathrm{RP}}$, we expect to have $A_{\mathrm{proxy},G} \simeq A_{\mathrm{proxy,BP+RP}}$. The ratio $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ is shown versus G magnitude in the top panel of Fig. C.2 for all sources in Database A. For magnitudes brighter than 18 mag, the ratio peaks at $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}} \simeq 1.35$. The fact that it is larger than one reflects the omission of the (f_{BP},f_{RP}) covariance term in the computation of $A'_{\mathrm{proxy,BP+RP}}$ (Eq. (C.10)). At the fainter side of the catalogue ($G > 18$ mag), the ratio decreases with increasing magnitude for the bulk of the data, reaching values below one. The likely cause of this effect is the residual astrophysical background, which becomes significative at faint magnitudes. The fact that only a low-resolution background calibration was used in DR2 affects the amplitude proxies. Source blending in the BP and RP spectrophotometers would also impact these proxies. As the orientation of the spectra on the CCD varies with each transit, the blending is differently affected at different times in the photometric time series.

The top panel of Fig. C.2 reveals the presence of a large number of sources having $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ ratios (much) larger than two, that is with variability amplitudes much larger in G than in the combined BP + RP band. These sources also have large (normalized) BP and RP flux excesses, as shown in the top panel of Fig. C.3, pointing to non-reliable G_{BP} and/or G_{RP} time series.

If we limit the sample to sources with reliable BP and RP flux excesses using condition (B.4), the $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}}$ versus G diagram becomes much cleaner. This is shown in the bottom panels of Figs. C.2 and C.3. The filter on C' also cleans the faint side of the diagram, where the remaining departure from $A_{\mathrm{proxy},G}/A'_{\mathrm{proxy,BP+RP}} \simeq 1$ at magnitudes fainter than 18 mag results from larger noise in BP and RP.

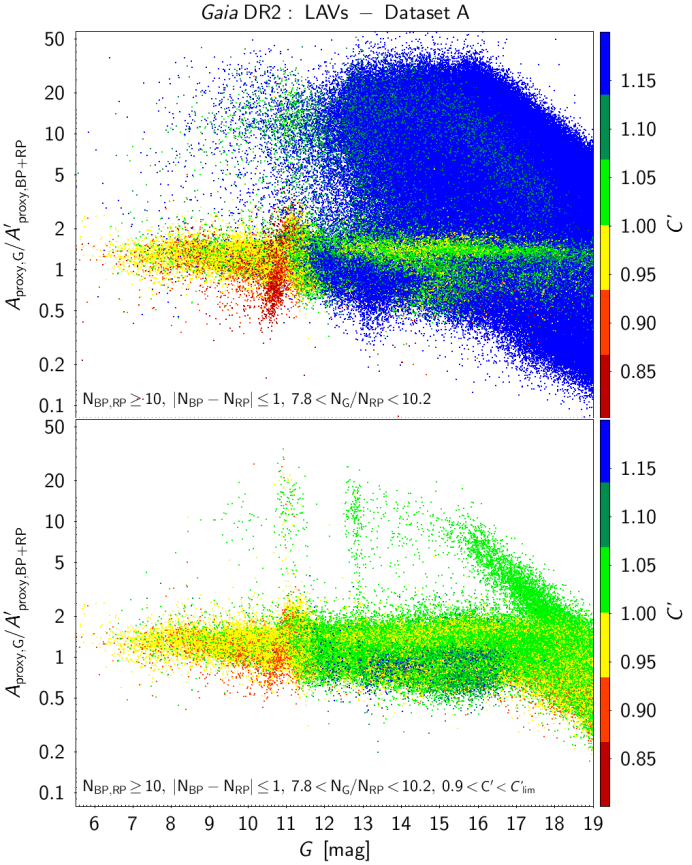

Fig. C.3. Same as Fig. C.2, but colour-coded with the value of the normalized BP and RP flux excess according to the colour scale shown on the right of the figure. The scales are identical to those in Fig. C.2.

The patterns observed at $A_{\rm proxy,G}/A'_{\rm proxy,BP+RP} \gtrsim 1.5$ in the bottom panel of Fig. C.2 are spurious. From the histograms shown in Fig. C.4, we take the limit of 1.5 above which we consider $A'_{\rm proxy,BP+RP}$ to be unreliable.

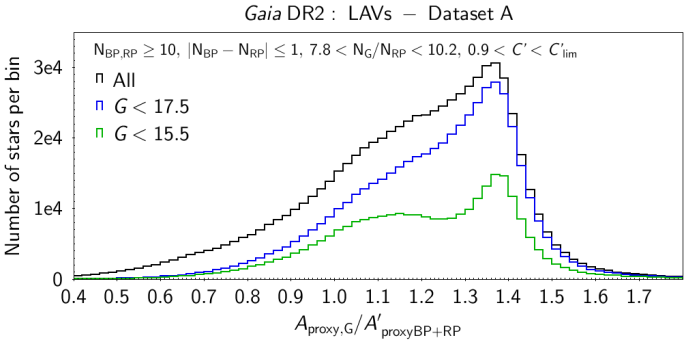

Fig. C.4. Histogram (in black) of $A_{\rm proxy,G}/A'_{\rm proxy,BP+RP}$ for the sample of Dataset A with $N_{BP,RP} \geq 10$, $|N_{BP} - N_{RP}| \leq 1$, $7.8 < N_{\rm obs}(G)/N_{\rm obs}(RP) < 10.2$, and $0.9 < C' < C'_{\rm lim}$. The blue and green histograms are for the subsets therein with $G < 17.5$ mag and $G < 15.5$ mag, respectively. The abscissa range has been limited for better visibility.

At the small side of $A_{\rm proxy,G}/A'_{\rm proxy,BP+RP}$, we note that a small ratio can indicate larger-than-expected variability in $G_{\rm BP}$ and $G_{\rm RP}$, but that the variability in G would still be good. However, if variability in $G_{\rm BP}$ and $G_{\rm RP}$ is to be studied, then a lower limit should also be applied, which we set at 0.8 This mainly impacts faint sources (see Fig. C.2, bottom panel).

Our final condition to select sources with reliable $A_{\rm proxy,BP}$ and $A_{\rm proxy,RP}$ amplitude proxies is summarize as:

$$0.8 < A_{\rm proxy,G}/A'_{\rm proxy,BP+RP} < 1.5. \quad \text{(C.12)}$$

It is used in the construction of Dataset C.

Appendix D: The electronic table

The electronic version of the catalogue is available at the CDS. The list of attributes published in the table is given in Table D.1, with reference to equations given in either the main body of the paper or in one of the Appendices.

Table D.1. Attributes published in our catalogue of *Gaia* DR2 LAVs (Dataset A).

Attribute	Notation	Description
source_id*		*Gaia* DR2 source ID
dataset_B		(boolean) Is in Dataset B
dataset_C		(boolean) Is in Dataset C
l_deg*		Galactic longitude (degree)
b_deg*		Galactic latitude (degree)
parallax_mas*		Parallax (milli arcsec)
parallax_error_mas*		Parallax uncertainty (milli arcsec)
phot_g_n_obs*	N_G	Number of points in G
phot_g_meanflux*	I_G	Mean flux in the G band (electron/s)
phot_g_meanflux_error*	$\varepsilon(I_G)$	Mean flux error in the G band (electron/s)
phot_g_mean_mag*	G	Mean G magnitude (mag)
phot_bp_n_obs*	$N_{\rm BP}$	Number of points in $G_{\rm BP}$
phot_bp_meanflux*	$I_{\rm BP}$	Mean flux in BP (electron/s)
phot_bp_meanflux_error*	$\varepsilon(I_{\rm BP})$	Mean flux error in BP (electron/s)
phot_bp_mean_mag*	$G_{\rm BP}$	Mean $G_{\rm BP}$ magnitude (mag)
phot_rp_n_obs*	$N_{\rm RP}$	Number of points in $G_{\rm RP}$
phot_rp_meanflux*	$I_{\rm RP}$	Mean flux in RP (electron/s)
phot_rp_meanflux_error*	$\varepsilon(I_{\rm RP})$	Mean flux error in RP (electron/s)
phot_rp_mean_mag*	$G_{\rm RP}$	Mean $G_{\rm RP}$ magnitude (mag)
phot_bp_rp_excess_factor*	C	BP and RP flux excess (Eq. (B.1))
phot_bp_rp_excess_factor_normalized	C'	Normalized BP and RP flux excess (Eq. (B.3))
isGoodBpRpExcessFactorNormalized		(boolean) Is BP and RP flux excess reliable?
DR2_LPV		(boolean) Is LPV candidate in DR2
DR2_RRL_SOS		(boolean) Is RR_Lyrae candidate in DR2 (sos table)
DR2_RRL_Classif		(boolean) Is RR_Lyrae candidate in DR2 (classif table)
DR2_Cep_SOS		(boolean) Is Cepheid candidate in DR2 (sos table)
DR2_Cep_Classif		(boolean) Is Cepheid candidate in DR2 (classif table)
DR2_dSct_SXPhe		(boolean) Is δ Scuti & SXPHE candidate in DR2
DR2_RotMod		(boolean) Is rotation modulation candidate in DR2
DR2_STS		(boolean) Is short time-scale candidate in DR2
amplProxyG	$A_{{\rm proxy},G}$	Amplitude proxy in G (Eq. (2))
amplProxyBP	$A_{\rm proxy,BP}$	Amplitude proxy in $G_{\rm BP}$ (Eq. (3))
amplProxyRP	$A_{\rm proxy,RP}$	Amplitude proxy in $G_{\rm RP}$ (Eq. (4))
amplProxyBPplusRPwithoutCov	$A'_{\rm proxy,BP+RP}$	Equation (C.10)
group_1_in_C		(boolean) Is in Group 1 (for Dataset C only, NULL otherwize)
group_2_in_C		(boolean) Is in Group 2 (for Dataset C only, NULL otherwize)
group_3_in_C		(boolean) Is in Group 3 (for Dataset C only, NULL otherwize)
group_4_in_C		(boolean) Is in Group 4 (for Dataset C only, NULL otherwize)
group_4a_in_C		(boolean) Is in Subgroup 4a (for Dataset C only, NULL otherwize)

Notes. The second column indicates the notation, if any, used in this paper for the attribute given in the first column. The third column provides a description of the attribute. An asterisk next to the attribute indicates that the data is a direct import from the *Gaia* DR2 archive. The catalogue is available at the CDS.