A&A 674, A24 (2023)
https://doi.org/10.1051/0004-6361/202244035
© The Authors 2023

Astronomy & Astrophysics
Special issue

Gaia Data Release 3

Gaia Data Release 3

The first *Gaia* catalogue of variable AGN

Maria I. Carnerero[1], Claudia M. Raiteri[1], Lorenzo Rimoldini[2], Deborah Busonero[1], Enrico Licata[1], Nami Mowlavi[2,3], Isabelle Lecoeur-Taïbi[2], Marc Audard[2,3], Berry Holl[2,3], Panagiotis Gavras[4], Krzysztof Nienartowicz[5], Grégory Jevardat de Fombelle[2], Ruth Carballo[6], Gisella Clementini[7], Ludovic Delchambre[9], Sergei Klioner[8], Mario G. Lattanzi[1], and Laurent Eyer[3]

[1] INAF – Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese, Italy
e-mail: maria.carnerero@inaf.it, claudia.raiteri@inaf.it
[2] Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland
[3] Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland
[4] RHEA for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada 28692 Madrid, Spain
[5] Sednai Sárl, Geneva, Switzerland
[6] Dpto. de Matemática Aplicada y Ciencias de la Computación, Univ. de Cantabria, ETS Ingenieros de Caminos, Canales y Puertos, Avda. de los Castros s/n, 39005 Santander, Spain
[7] INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Gobetti 93/3, 40129 Bologna, Italy
[8] Lohrmann Observatory, Technische Universität Dresden, Mommsenstraße 13, 01062 Dresden, Germany
[9] Space sciences, Technologies & Astrophysics Research (STAR) Institute, Institute of Astrophysics and Geophysics, University of Liège, Allée du Six-Août, 17, 4000 Sart Tilman, Belgium

Received 16 May 2022 / Accepted 13 July 2022

ABSTRACT

Context. One of the novelties of *Gaia* data release 3 (DR3) with respect to the previous data releases is the publication of the multiband light curves for about 1 million active galactic nuclei (AGN) and of the values of some parameters characterising their variability properties.
Aims. The goal of this work is the creation of a catalogue of variable AGN, the selection of which is based on *Gaia* data only.
Methods. We first present the implementation of the methods used to estimate the variability parameters within the framework of a specific object study module for AGN (SOS-AGN). We then describe the selection procedure that led to the definition of the high-purity *Gaia* variable AGN sample and analyse the properties of the selected sources. We started from a sample of millions of sources, which were identified as AGN candidates using 11 different classifiers based on variability processing. Because the focus is on the variability properties, we first defined some pre-requisites in terms of number of data points in the *G* band and mandatory variability parameters. A series of filters was then applied using only *Gaia* data and the *Gaia* Celestial Reference Frame 3 (*Gaia*-CRF3) sample as a reference.
Results. The resulting *Gaia* AGN variable sample, named GLEAN, contains about 872 000 objects, more than 21 000 of which are new identifications. We checked for the presence of contaminants by cross-matching the selected sources with a variety of galaxies and stellar catalogues. The completeness of GLEAN with respect to the variable AGN in the last Sloan Digital Sky Survey (SDSS) quasar catalogue is ∼47%, while that based on the variable AGN of the *Gaia*-CRF3 sample is ∼51%. The set of filters applied to the sources selected by SOS-AGN to increase the sample purity reduced the source number by about 37%. From both a comparison with other AGN catalogues and an investigation of possible contaminants, we conclude that purity can be expected to be above 95%. The multi-wavelength properties of these sources are investigated. In particular, we estimate that ∼4% of them are radio-loud. We finally explore the possibility to evaluate the time-lags between the flux variations of the multiple images of strongly lensed quasars, and show one case.

Key words. galaxies: active – catalogs – methods: data analysis – quasars: general – gravitational lensing: strong

1. Introduction

Active galactic nuclei (AGN) are present in a variety of different types, all characterised by accretion of matter onto a supermassive black hole (SMBH) with mass greater than 1 million solar masses. A fraction of AGN are radio-loud (Jiang et al. 2007; Kratzer & Richards 2015, see discussion in Sect. 7) and exhibit two plasma jets that are launched from, or from close to, the poles of their SMBH in opposite directions. The members of a peculiar class of AGN, called blazars (including flat-spectrum radio quasars (FSRQs) and BL Lac-type objects), have one of the two jets closely aligned with the line of sight, which makes their multi-wavelength jet emission relativistically Doppler beamed (Urry & Padovani 1995).

The flux of most AGN presents variability at some level, with different variability timescales and amplitudes. The optical continuum emission from AGN is in general dominated by the thermal radiation coming from the accretion disc, and shows smooth

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is published in open access under the Subscribe to Open model. Subscribe to A&A to support open access publication.

variability on timescales of months to years. In contrast, the prevailing source of optical emission in the most active blazars is the non-thermal radiation from the relativistic jet, where Doppler beaming enhances the amplitude and reduces the timescales of variability, and even intra-night flux changes of up to several tenths of a magnitude can be observed (e.g., Raiteri et al. 2017). In objects at low redshift, the host galaxy emission can contribute significantly to, or even dominate, the optical emission, reducing the amplitude of variability.

AGN are broadly classified in two classes. In type 1 AGN, the optical spectra show broad emission lines that are produced in a nuclear zone close to the black hole with fast-moving gas clouds. These lines are not seen in the spectra of type 2 AGN, likely because of the obscuration effect of a dusty torus. Narrow emission lines then appear in both type 1 and type 2 AGN spectra, and come from an outer nuclear region, where gas clouds have smaller velocities.

Different methods have been used to identify AGN. A series of catalogues of spectroscopically confirmed quasars from the Sloan Digital Sky Survey[1](SDSS) have been published over the last two decades, from the early data release by Schneider et al. (2002), which includes 3814 quasars detected over 494 deg^2, through various releases, until the most recent one (DR16Q, Lyke et al. 2020), which contains 750 414 quasars within 9376 deg^2. The quasar selection criteria included colour indices and variability. The SDSS quasar catalogues have been used for a large variety of studies, from cosmology to the characterisation of quasar properties.

Richards et al. (2002) selected quasar candidates in the SDSS using colour indices obtained from data in the *ugriz* filters and by searching the radio counterparts of the unresolved sources in the FIRST catalogue. Richards et al. (2009) updated this latter work by also considering the UV excess and extended the analysis to high-redshift quasars. A mixed selection method, including optical colours and variability, was adopted by Eyer (2002) and Ross et al. (2012); the latter authors also used data at other wavelengths. Some authors proposed quasar-selection methods based uniquely on variability. MacLeod et al. (2011) adopted a damped random walk model to describe the temporal behaviour of quasars and to parametrize the quasar structure function. This allowed them to derive the characteristic variability timescale and a driving amplitude of short-term variations, which allow a very efficient separation of quasars from stars. Under the same assumption that the quasar temporal behaviour can be described as a damped random walk, Butler & Bloom (2011) modelled the ensemble quasar structure function as a function of magnitude. This produced metrics for evaluating the probability that a source is a quasar.

Colour indices obtained from the mid-infrared all-sky survey performed by the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010) satellite were found to be a superb tool to classify celestial objects, in particular AGN (Mateos et al. 2012; Stern et al. 2012; Assef et al. 2013, 2018; Secrest et al. 2015). Other studies combined optical and WISE data, but were limited in sky coverage until *Gaia* data became available. Yan et al. (2013) used both WISE and SDSS photometry to characterise extragalactic sources and highlighted the power of WISE to identify AGN. In particular, these authors found that strong AGN at $z \leq 3$ show $W1 - W2 > 0.8$ and $W2 < 15.2$. Type-2 AGN candidates in addition require $r - W2 > 6$.

Shu et al. (2019) cross-matched the *Gaia*-DR2 (Gaia Collaboration 2018) and unWISE (Schlafly et al. 2019) catalogues and used a random-forest classifier based on 16 features to select AGN. Schlafly et al. (2019) found that the most effective features are the $W1 - W2$ colours, the proper motion significance, and the extinction-corrected $G - W1$ colour. The authors built two catalogues: one with overall completeness of 75% (C75), including 2 734 464 sources, 2 182 193 of which constitute a 85% reliability catalogue (R85).

The MILLIQUAS catalogue (Flesch 2015) contains about 2 million AGN and high-confidence candidates from other catalogues. It was recently updated by Flesch (2021), including associations with Very Large Array Sky Survey (VLASS; Lacy et al. 2020) radio sources. Recently, Liu et al. (2022) published a catalogue of X-ray properties of AGN in the Final Equatorial-Depth Survey (eFEDS) performed by eROSITA[2].

The problem of selecting quasars at low Galactic latitudes, where extinction makes the task extremely hard, was faced by Fu et al. (2021). These authors built a catalogue of 160 946 sources at $|b| \leq 20$ deg using photometric data from Pan-STARRS1[3] (PS1) and AllWISE (Cutri et al. 2013) for classification, and *Gaia* proper motions to exclude stellar contaminants.

The extragalactic content of *Gaia*-DR2 was analysed by Bailer-Jones et al. (2019), who identified quasars and galaxies using *Gaia* photometric and astrometric information only. They classified 2.3 million objects as quasars, inferring that the realistic number is around 690 000. Gaia Collaboration (2023) present the extragalactic content of *Gaia*-DR3, providing catalogues of AGN (and galaxies) that were driven by completeness –but with low purity– together with the prescriptions to obtain higher purity samples.

The aim of the present paper is to first present the *Gaia* Specific Object Study package on AGN (SOS-AGN), which is part of the variability analysis pipeline discussed in Eyer et al. (2023). The package receives inputs from the classification module (see Rimoldini et al. 2023) and implements methods to estimate the variability characteristics of the candidate AGN. We then describe the procedure that led to the selection of a high-purity sample of variable AGN and analyse its properties. Because the emphasis is on variability, among the several million AGN candidates provided by the classifiers (Rimoldini et al. 2023), we consider only those sources whose G light curve contains at least 20 field-of-view (FoV) transits in the G band and for which some relevant parameters can be defined. A set of filters is then applied, which are tailored to the properties of the AGN belonging to the *Gaia*-CRF3 sample (Gaia Collaboration 2021, 2022). Some basic information on SOS-AGN and selection results can also be found in the *Gaia*-DR3 online documentation[4] (Rimoldini et al. 2022).

The paper is organised as follows. In Sect. 2, we describe the content of the SOS-AGN module, which allowed us to perform a preliminary analysis of the variability characteristics of the objects. Section 3 details the series of cuts that we applied to remove contaminants. The properties of the selected variable AGN sample are discussed in Sect. 4, while a search for possible stellar contaminants is presented in Sect. 5. The completeness and purity of our sample are addressed in Sect. 6. In Sect. 7, we look for the radio counterparts of our objects and infer the fraction of radio-loud sources, while in Sect. 8 we discuss the possibility to derive time lags from the *Gaia* light curves of the

[1] https://www.sdss.org/

[2] https://erosita.mpe.mpg.de/edr/eROSITAObservations/Catalogues/

[3] https://panstarrs.stsci.edu/

[4] https://gea.esac.esa.int/archive/documentation/GDR3/

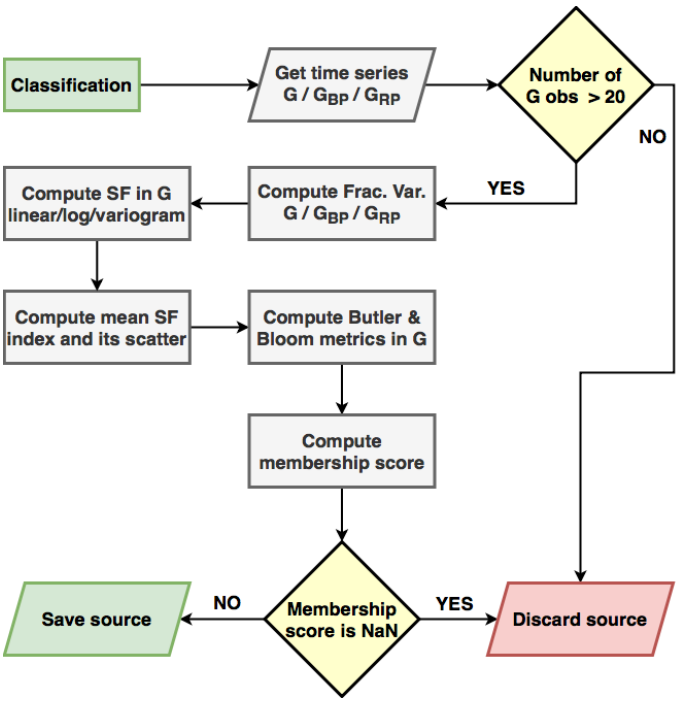


Fig. 1. Flow chart of the SOS-AGN package.

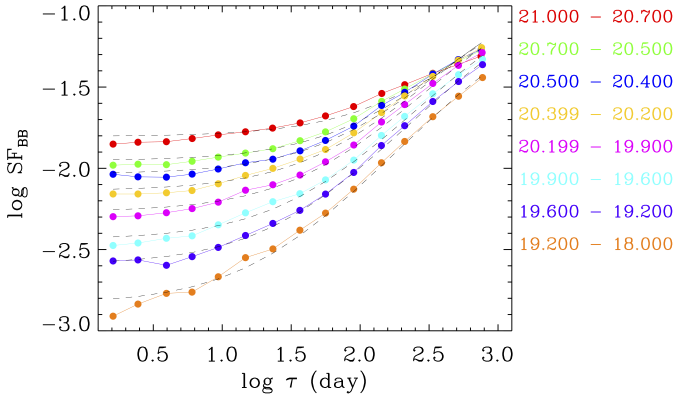


Fig. 2. Mean SF vs. time lag τ for sources with more than five FoV transits in the G band in a preliminary version of the *Gaia*-CRF3 sample, including about 1 850 000 AGN candidates. The various colours correspond to different G ranges for which the mean SFs have been estimated. Dashed lines represent the best-fit models to the mean SFs according to Eq. (3).

multiple images of gravitationally lensed quasars. A brief summary of our results and conclusions are presented in Sect. 9.

2. SOS-AGN

Our goal is to select a sample of variable AGN candidates that is as pure as possible. For this, SOS-AGN was implemented in the *Gaia* DR3 variability pipeline (Eyer et al. 2023), which depends on the upstream modules of general variability detection (GVD) and classification. GVD pre-selected the 25% most variable objects per magnitude interval in the G band. These variables were then classified using supervised methods. The training representatives of AGN originated mainly from *Gaia*-CRF3, given its high purity and all-sky distribution. The brightest known AGN were included in the training set to improve the chances of detecting rare bright AGN. All the AGN sources used for training satisfied the variability threshold of GVD. The filters applied to AGN classification results were similar to those used in SOS-AGN (as described in Sect. 3), although with generally more permissive thresholds.

The flow chart of the SOS-AGN processing is shown in Fig. 1. The first requirement for the sources to be considered was the presence of at least 20 FoV transits in the G band light curve. We then defined some 'mandatory' metrics, the values of which are listed in the *Gaia*-DR3 `vari_agn` table. By mandatory, we mean that if the parameter does not produce a real value, then the object is discarded.

The first mandatory parameter was the fractional variability (Vaughan et al. 2003) in the G band, named `fractional_variability_g` in the `vari_agn` table. The flux in the G band was calculated as $F = 10^{-0.4\,(G-ZP_G)}$, where G (`median_mag_g_fov` in the `vari_summary` table, here and thereafter) is the derived time-series median, and $ZP_G \sim 25.7$ is the zero point in the G band in the Vega system (Riello et al. 2021).

To mitigate the effect of outliers, we modified the standard fractional variability definition, adopting for the flux statistics the median instead of the mean, and the median absolute deviation (MAD) instead of the standard deviation:

$$\texttt{fractional_variability_g} = \frac{\sqrt{\mathrm{MAD}^2(F) - \langle\sigma_F^2\rangle}}{\mathrm{median}(F)}, \qquad (1)$$

where $< \sigma_F^2 >$ is the mean of the squared flux uncertainties. We note that because the standard deviation is approximately $1.5 \times$ MAD, the above definition leads to lower fractional variability values than in the classical case. In contrast, the photometric uncertainties are somewhat underestimated (Evans et al. 2023), which has the opposite effect.

The second and third mandatory parameters were the index of the structure function (SF), `structure_function_index`, and its scatter `structure_function_index_scatter`. The slope of the SF in the log(SF) versus log τ diagram is a powerful parameter for selecting AGN. There are many implementations of the SF in the literature; we adopted the classical algorithm developed by Simonetti et al. (1985):

$$\mathrm{SF}_{\mathrm{Sim}}(\tau) = \langle[\mathrm{mag}(t) - \mathrm{mag}(t+\tau)]^2\rangle, \qquad (2)$$

where τ is the time lag. The slope of the SF depends on the variability behaviour of the source and on other important physical parameters, such as redshift. AGN are known to show long-term variability, with SF slopes typically larger than 0.1 (e.g., Eyer 2002; Sumi et al. 2005).

To implement an automatic estimate of the SF slope for every single source, we must take into account the fact that, as log τ increases, log(SF) ideally presents first a plateau, which depends on the noise, then an almost linear increase, and finally another plateau, where the second break point indicates a characteristic timescale (e.g., Hughes et al. 1992). To calibrate the first break, we built SFs for the ~1 850 000 sources in a preliminary version of the *Gaia*-CRF3 sample, divided into magnitude bins. Figure 2 shows the results. As the magnitude increases, the break point shifts towards longer τ values. The second break was set where log(SF) reaches its maximum value. For each source, the SF behaviour between the two breaks was then fit with a least-square linear regression, after discarding data points with large uncertainties. Because of the dependence of the first break on magnitude mentioned above, the first break was set according to the source average magnitude, while the second break was

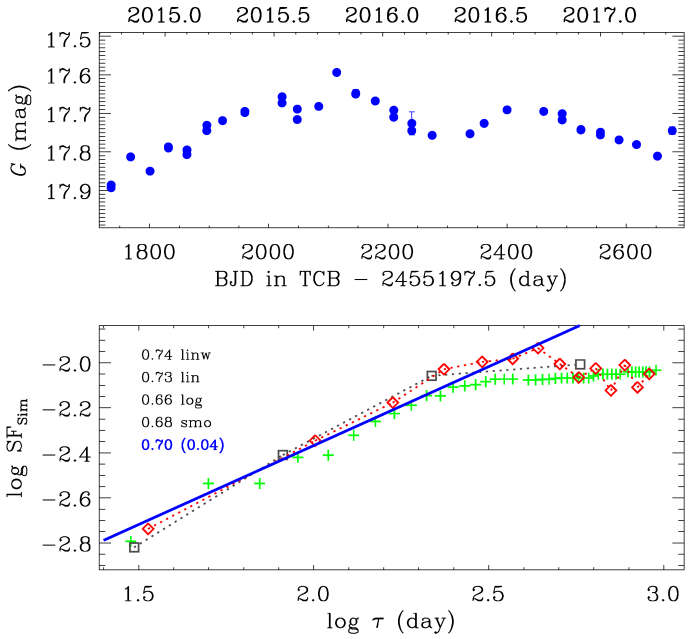


Fig. 3. *Top*: G-band light curve of a representative variable AGN (*Gaia* DR3 source_id=4768409993534612992). *Bottom*: SF obtained with linear (red diamonds) and logarithmic (grey squares) time-lag sampling, and with the smoothed variogram (green plus signs). The legend lists the SF slopes obtained with the four methods described in the text (black) and the final slope (blue) with its standard deviation in brackets. A line with this slope is drawn in blue.

defined by its log(SF) maximum. For each source, the linear fit was performed in four different ways, whose results were finally averaged. We considered both linear and logarithmic τ bins, and in the linear case we estimated the slope by also weighting for the number of data points in each bin. In addition to these three methods, we estimated the slope through linear regression of a smoothed variogram (Eyer & Genton 1999). In the *Gaia*-DR3 `vari_agn` table, `structure_function_index` represents the average value and `structure_function_index_scatter` the standard deviation of these four estimates. The bottom panel of Fig. 3 shows an example of SF slope determination using the four methods above.

The fourth and fifth mandatory parameters are the `qso_variability` and `non_qso_variability` metrics (in the `vari_agn` table) introduced by Butler & Bloom (2011). As mentioned in the introduction, Butler & Bloom (2011) developed a method to select quasars from their light curve behaviour in a single photometric band. This is based on a damped-random walk modelling of the SF. The parametrization of the SF made by the above authors using the g-band SDSS light curves of the quasars in the Stripe 82 sky region had to be adapted to the *Gaia* data. We used the expression:

$$\mathrm{SF}_{\mathrm{BB}}(\tau) = \eta^2 + \sigma^2\,[1 - \exp(-\tau/\tau_0)], \tag{3}$$

where the term η^2 accounts for the noise and τ_0 was fixed to 1000 days in agreement with the results of Butler & Bloom (2011). The application of this model to the mean SFs in selected magnitude ranges (see Fig. 2), whose average value is $\langle G\rangle$, allowed us to obtain the best-fit parameters η^2 and σ^2 for each magnitude bin. The trends of these parameters versus magnitude

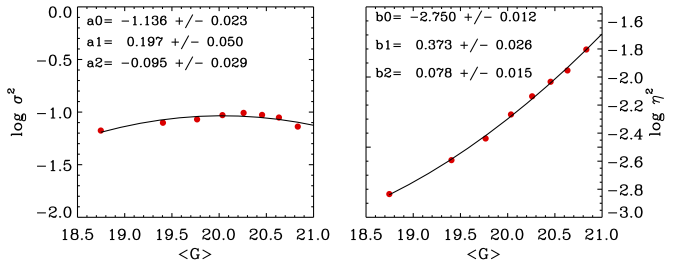


Fig. 4. Results of the quadratic fits to the quantities $\log\sigma^2$ and $\log\eta^2$ defining the SF in Eq. (3). Each data point corresponds to a mag range in Fig. 2. The best-fit values of the parameters a_i and b_i in Eqs. (4) and (5) are listed in the legends.

(see Fig. 4) were then fitted by quadratic relations of the form:

$$\log\sigma^2 = a_0 + a_1\,(\langle G\rangle - 19) + a_2\,(\langle G\rangle - 19)^2, \tag{4}$$

$$\log\eta^2 = b_0 + b_1\,(\langle G\rangle - 19) + b_2\,(\langle G\rangle - 19)^2, \tag{5}$$

to obtain the best-fit coefficients a_i and b_i that were required to calculate the `qso_variability` and `non_qso_variability` metrics for every source[5].

Finally, we defined a membership score, named `vari_agn_membership_score` and published exclusively in the `qso_candidates` table, which was calculated from the inverse of the Mahalanobis distance D based on five parameters (`fractional_variability_g`, `structure_function_index`, `qso_variability`, `non_qso_variability`, and `abbe_mag_g_fov`, where the latter is a parameter in the `vari_summary` table; see Sect. 3.3), and then rescaled by a Gaussian to return values between 0 and 1:

$$\texttt{vari_agn_membershipscore} = \exp[-D^2/(2\,\rho^2)]. \tag{6}$$

The square of the Mahalanobis distance in Eq. (6) was computed as $D^2 = (\mathbf{x} - \mathbf{m})^T\,C^{-1}\,(\mathbf{x} - \mathbf{m})$, where, for a given source, $\mathbf{x}$ is the vector of the observed values of the five parameters listed above, while the vector $\mathbf{m}$ of the mean values and the covariance matrix C are based on the observational data for a sample of *Gaia*-CRF3 objects that were detected as variable by the GVD module. The parameter ρ was set to 2.7 to have more than 90% of CRF3 sources with score larger than 0.5. Figure 5 shows the results of a check of the `vari_agn_membershipscore` values on three classes of sources: AGN in the *Gaia*-CRF3 sample, galaxies (Krone-Martins et al., in prep.), and variable stars (Gavras et al. 2023). As can be seen, the distribution of scores for *Gaia*-CRF3 sources is distinct from that of the other two classes of objects.

3. Selection procedure

Samples of variable AGN candidates with corresponding probabilities were provided by 11 classifiers in the *Gaia* DR3 variability pipeline based on different prescriptions. For details, see Rimoldini et al. (2023). About 34 million sources from different classifiers met the criteria defined by the SOS-AGN module. However, to reduce the sample to the most reliable candidates, for each classifier we compared the probabilities of the AGN candidates with those of the *Gaia*-CRF3 sources therein and set

[5] The `qso_variability` and `non_qso_variability` metrics actually represent the logarithm of the corresponding Butler & Bloom (2011) metrics.

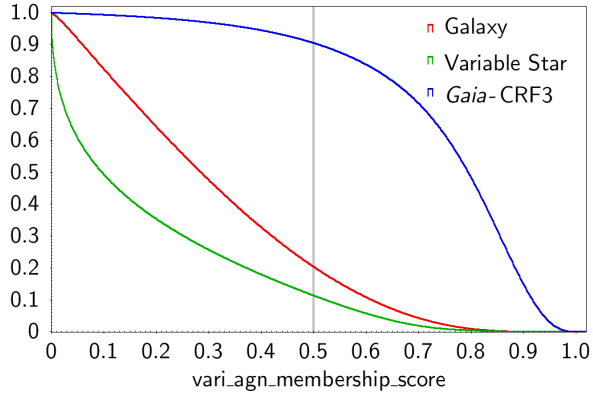


Fig. 5. Normalised reverse cumulative distribution of the membership score (1-CDF(score)) for one million AGN from *Gaia*-CRF3 (blue), 0.8 million galaxies (red; Krone-Martins et al., in prep.), and 3 million variable stars (green; Gavras et al. 2023) from the literature. More than 90% of the AGN have values greater than 0.5.

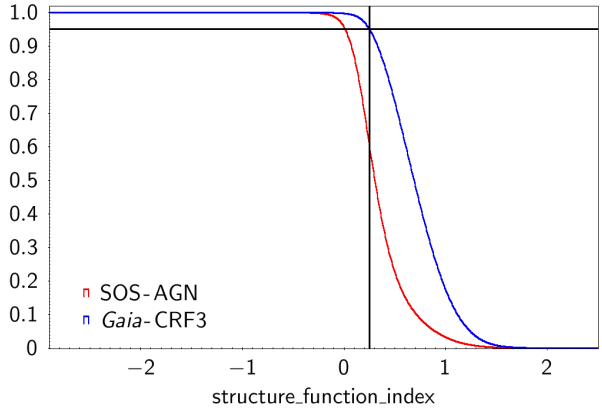


Fig. 6. Normalised reverse cumulative distribution of the `structure_function_index` for the ∼10 million variable AGN candidates (red) and for the *Gaia*-CRF3 sources (blue). The vertical line indicates the threshold of 0.25, that is, the minimum value of the index required to pass the selection.

minimum probability thresholds so that no more than 5% of the CRF3 sources were lost. This resulted in a sample of 10 million sources with more than 20 FoV transits in their *G* band light curves, and for which the mandatory parameters described in Sect. 2 are defined. Among them, 1.1 million are included in the *Gaia*-CRF3 sample.

The selection procedure continued with the application of a sequence of filters tailored to the *Gaia*-CRF3 sources. The goal was to obtain a variable AGN sample that is as pure as possible, with the minimum loss of *Gaia*-CRF3 objects. In the following, we describe the subsequent filters which were adopted to remove contaminants. We stress that the same names are used to denote both the initial sample and the subsamples that are derived from it as a result of the various steps in the selection procedure. As an example, the term '*Gaia*-CRF3' indicates both the original sample and the various ensembles of sources belonging to it that survive the subsequent selection cuts.

3.1. Structure function Index

Following the considerations in Sect. 2, we decided to keep candidates that satisfied the condition `structure_function_index` > 0.25, where the `structure_function_index` is the slope of $\log \mathrm{SF}_{\mathrm{Sim}}$ (see Eq. (2)) versus $\log \tau$. As shown in Fig. 6, in this way we lost about 40% of the dubious variable AGN candidates, but only 5% of *Gaia*-CRF3 sources, leaving 1 million *Gaia*-CRF3 objects and 6.2 million candidates.

3.2. QSO versus non-QSO statistics

The Butler & Bloom (2011) metrics were used for further cuts, defined by the region in the `qso_variability` versus `non_qso_variability` space expected to host the vast majority of AGN. The *Gaia*-CRF3 sources confirmed the locations of `qso_variability` around zero and of `non_qso_variability` above zero. We defined the following cuts, where some margin was left to minimise the loss of bona fide CRF3 sources (see Fig. 7):

`qso_variability` > −1.05
`qso_variability` < 0.6
`non_qso_variability` > 0
`non_qso_variability` > −0.7×`qso_variability`−0.33
`non_qso_variability` > 0.5 × `qso_variability`.

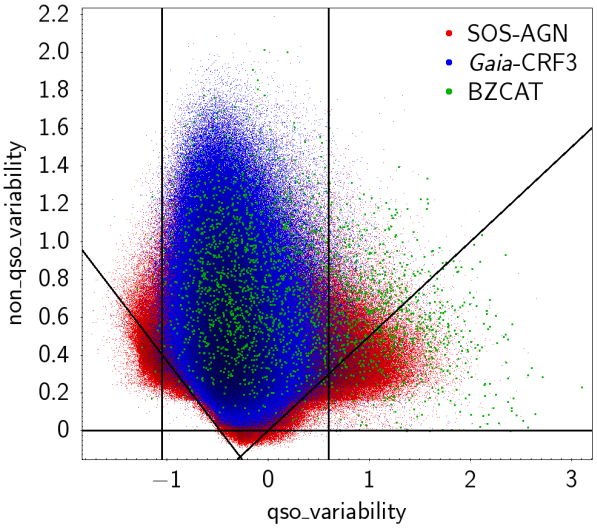


Fig. 7. Butler & Bloom (2011) metrics (or rather their logarithm) `non_qso_variability` versus `qso_variability` plot, showing the position of the variable AGN candidates (red dots), distinguishing those in the *Gaia*-CRF3 sample (blue dots), and the blazars in the BZCAT5 catalogue (green dots). The lines highlight the cuts performed to remove contaminants.

In Fig. 7, we highlight the sources included in the fifth edition of the Roma-BZCAT blazar catalogue (BZCAT5; Massaro et al. 2015). Their distribution in `qso_variability` is wider than that of typical AGN.

After the above selections, we were left with ∼6 million candidates, while the number of *Gaia*-CRF3 sources remained around 1 million. This filter removed many blazars (about 35%), whose `qso_variability` values extend to larger values than those of AGN in the *Gaia*-CRF3.

3.3. Further filtering

Constraints were set on the `abbe_mag_g_fov` (from the `vari_summary` table) and renormalised unit weight error (`ruwe`, in the `gaia_source` table) parameters. The `abbe_mag_g_fov` is defined as half of the ratio of the mean square difference between

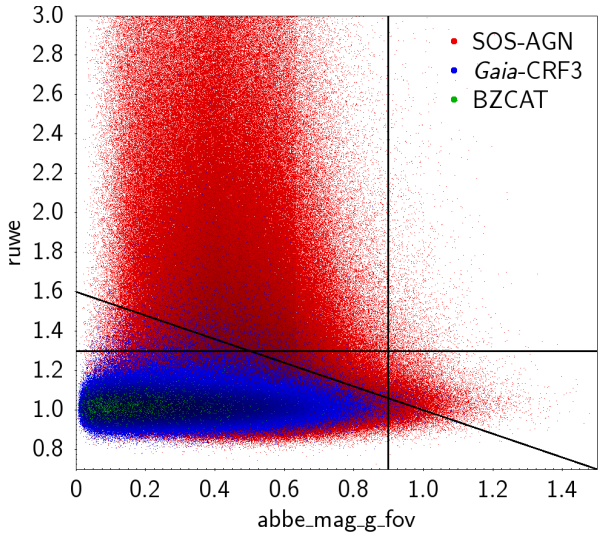


Fig. 8. As Fig. 7, but for `ruwe` versus `abbe_mag_g_fov`.

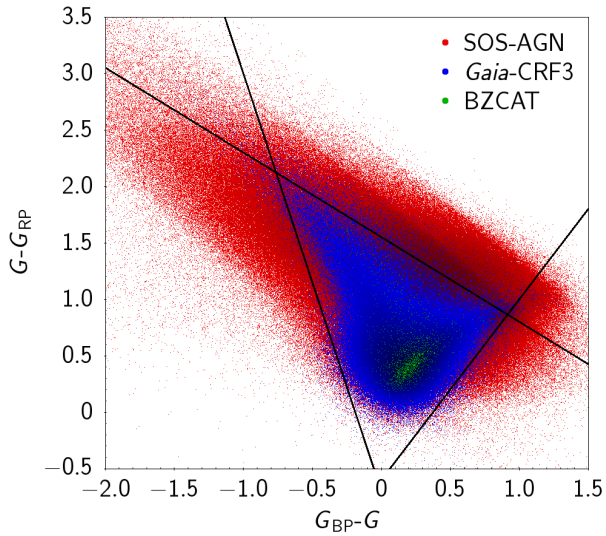


Fig. 9. As Fig. 7, but for $G - G_{\rm RP}$ versus $G_{\rm BP} - G$.

consecutive data points in the G band light curve to its variance (small values correspond to time series that are smooth in time). The `ruwe` parameter gives an estimate of the suitability of the single-star astrometric model for a given source (values close to one indicate a good agreement). The AGN light curves generally exhibit long-term variations, which are often sufficiently resolved by *Gaia*'s sampling to cause a tendency towards small values of `abbe_mag_g_fov`. Moreover, most AGN appear as astrometrically stable point sources, and therefore are usually associated with `ruwe` values close to one. The *Gaia*-CRF3 sources confirm such expectations as they populate a compact strip in the `ruwe` versus `abbe_mag_g_fov` space. Thus, the selection region was defined as follows (see Fig. 8):

`ruwe` < 1.3
`ruwe` $< -0.6 \times$ `abbe_mag_g_fov` $+ 1.6$
`abbe_mag_g_fov` < 0.9.

These 2D cuts decreased the SOS-AGN sample to around 4.8 million AGN candidates (still ∼1 million in *Gaia*-CRF3). This filter has only a minor effect on blazars, reducing them by about 1.4%.

Optical colour indices have proved to be important for quasar selection, even if not decisive in general. We used *Gaia* colours derived from time-series medians, which are found in the `vari_summary` table. We filtered the sources in the $G_{\rm BP} - G$ (`median_mag_bp` − `median_mag_g_fov`) versus $G - G_{\rm RP}$ (`median_mag_g_fov` − `median_mag_rp`) region enclosed within the following conditions (see Fig. 9):

$G - G_{\rm RP} > -3.7 \times G_{\rm BP} - G - 0.7$
$G - G_{\rm RP} < -0.75 \times G_{\rm BP} - G + 1.55$
$G - G_{\rm RP} > 1.6 \times G_{\rm BP} - G - 0.6$.

This led to around 3.9 million candidates (still ∼1 million in *Gaia*-CRF3).

Extragalactic sources should ideally have null (statistically insignificant) parallax and proper motions. Therefore, these astrometric parameters (included in the `gaia_source` table) can efficiently help to remove Galactic contaminants. To take uncertainties into account, the corresponding cut was made on the ratio of these parameters to their errors. Such ratios are expected to follow a normal distribution with unit variance and zero mean. A permissive condition kept candidates within 5-sigma (see also Gaia Collaboration 2022). For parallax:

$$|(\texttt{parallax} + 0.017)/\texttt{parallax_error}| < 5, \quad (7)$$

where the addition of 0.017 mas to `parallax` takes into account the global parallax zero point of *Gaia* EDR3 (Lindegren et al. 2021a,b). For proper motion (`pm`):

$$\texttt{pm} = \sqrt{\frac{\alpha^2 + \beta^2 - 2\,\alpha\,\beta\,\gamma}{1 - \gamma^2}} < 5,$$

where the `pm` components along the equatorial coordinates, their uncertainties, and correlation are taken into account as follows: α = `pmra/pmra_error`, β = `pmdec/pmdec_error`, and γ = `pmra_pmdec_corr`. After this selection, the number of SOS-AGN candidates was 1.6 million.

To reduce AGN misclassification in crowded stellar fields, for example, in the Galactic plane and Magellanic Clouds, we set a constraint on the environment of each candidate, limiting the maximum number density of sources within 100 arcsec to 0.004 arcsec^{-2}. About 1.2 million sources passed this requirement. Artificial variability is produced by the scan angle variations for extended objects (see Holl et al. 2023), as may happen for detectable AGN host galaxies.

We then set an upper limit to the Spearman correlation between the G-band time series and the model of the image parameter determination (IPD) $r_{\rm ipd}$ (at scan angles corresponding to the time series observations), which quantifies the amount of scan-angle-dependent signal in the photometric time series (see Holl et al. 2023, for details). The constraint $r_{\rm ipd} < 0.8$ removed only about 2000 sources.

A final cut on the GVD variability probability was also made to further increase the sample purity, in view of the fact that part of the variations might be due to a spurious signal when the host galaxy is detectable. The final list of variable AGN candidates contains 872 228 sources, 150 017 of which are not included in *Gaia*-CRF3. The final list also contains almost 3000 objects that did not pass the selection procedure because of peculiar properties (like blazars, lensed AGN, and the brightest known AGN), but were added to the final sample for their interest.

In the rest of this article, we refer to the whole set of selected *Gaia* variable AGN as the GLEAN (*Gaia* variabLE AgN) sample, and to those objects that are in the GLEAN sample but not in the *Gaia*-CRF3 one as the CANOE (CANdidates tO Explore) sample. The CANOE objects represent an interesting AGN candidate subsample to be explored in view of a possible future addition to the *Gaia*-CRF3.

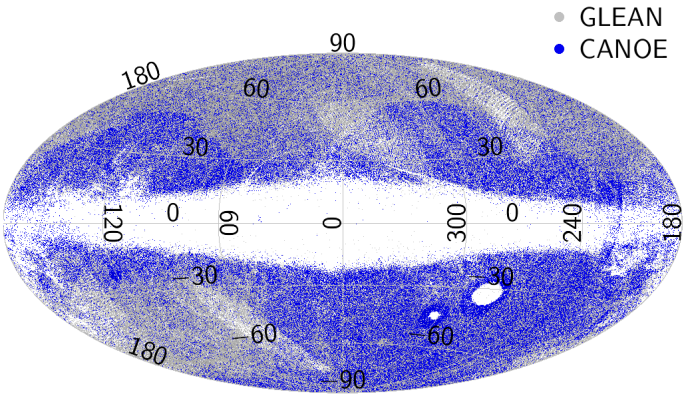


Fig. 10. Sky distribution of the sources in the GLEAN (grey) and CANOE (blue) samples in Galactic coordinates. Mostly because of the environment filter, the Galactic Plane and Magellanic Clouds are almost empty. Some scanning law footprints are still visible.

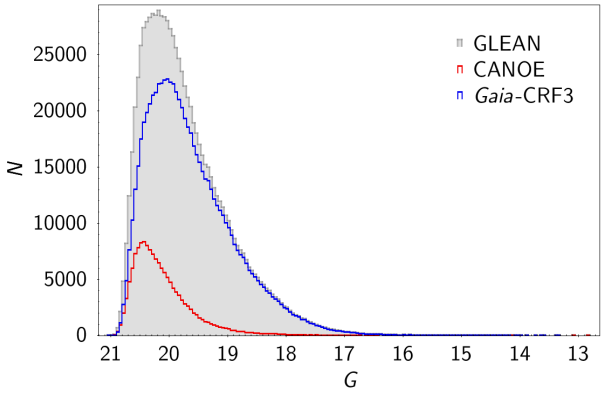


Fig. 11. Magnitude distribution (`median_mag_g_fov`) of all sources in the GLEAN (grey), CANOE (red), and *Gaia*-CRF3 (blue) samples, in bins of 0.05 mag.

4. The *Gaia* variable AGN sample

The sky distribution of the sources in the GLEAN sample is shown in Fig. 10. The Galactic Plane and Magellanic Clouds are almost empty, as expected because of the filters applied, in particular that on the environment. However, there is still an excess of AGN around the Magellanic Clouds, which may indicate some stellar contamination, or that these regions are still partially unexplored from an extragalactic point of view.

The G magnitude distribution (`median_mag_g_fov`) is plotted in Fig. 11 for the complete GLEAN sample, and for those sources in the sample that belong to the CANOE and *Gaia*-CRF3 subsamples. The distribution of the CANOE sources peaks at a fainter magnitude than that of the CRF3 objects.

One of the main novelties of *Gaia* DR3 is the publication of the light curves for the AGN selected in this paper and in the paper by Rimoldini et al. (2023). Figures 12–15 display the *Gaia* multiband light curves of four representative sources: a BL Lac-type object, a flat-spectrum radio quasar (FSRQ), a Seyfert galaxy, and a radio-quiet quasar. The figures also show the SDSS spectra (Abolfathi et al. 2018) and the *Gaia* passbands in order to highlight the spectral coverage of the *Gaia* filters. For three of these sources, *Gaia* low-resolution spectra are available in DR3 and are shown in the same figures. Details on their calibration can be found in Carrasco et al. (2021), De Angeli et al. (2023), and Montegriffo et al. (2023). We underline that *Gaia* spectroscopic information has not been used in the selection of variable AGN candidates performed in this paper.

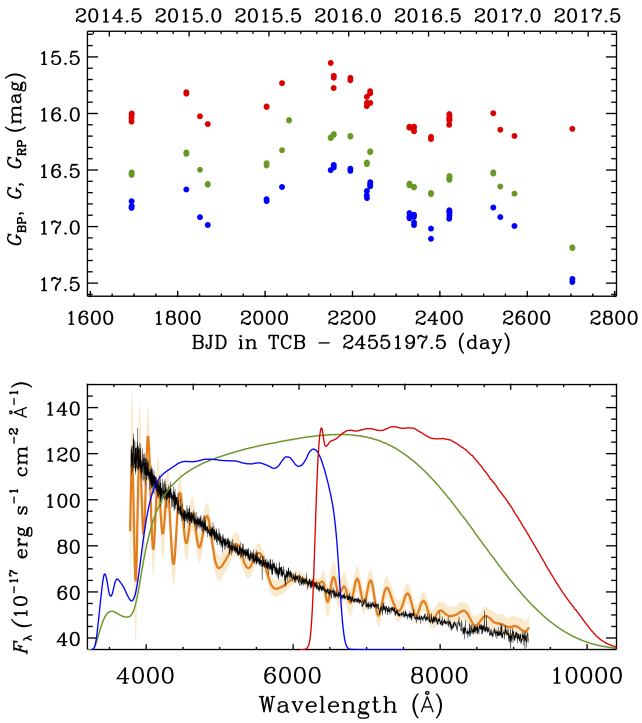


Fig. 12. *Top*: G (green), G_{RP} (red), and G_{BP} (blue) light curves of the BL Lac-type source 5BZBJ0035+1515 (*Gaia* DR3 source_id: 2780475069095852672). *Bottom*: SDSS spectrum (black), *Gaia* low-resolution spectrum (orange) with its uncertainty (shadowed orange region), and *Gaia* passbands.

The light curves of the BL Lac-type object (Fig. 12) show more than 1 mag variability in the G band; the SDSS spectrum is featureless, confirming that the dominant contribution is synchrotron emission from the jet. A rapid flare characterises the light curve of the FSRQ (Fig. 13) at the beginning of the *Gaia* monitoring, with a brightness decrease of about 2 mag, followed by a slow brightness increase. The SDSS spectrum includes the main emission lines usually present in quasar spectra, redshifted to $z \sim 0.414$. This indicates a strong emission contribution from the broad line region in addition to that of the jet. The light curves of the Seyfert galaxy (Fig. 14) show smooth variability, with maximum amplitude of about 0.7 mag in the G band and some dispersion of the data points acquired on the same Julian day, especially in the G_{BP} band. Because of the source faintness, the SDSS spectrum is somewhat noisy, but clearly shows the typical features of a narrow-line Seyfert 1 galaxy redshifted to $z \sim 0.118$. Smooth variability (with some noise) also characterises the light curves of the radio-quiet quasar (Fig. 15); the SDSS spectrum shows emission lines, in particular a prominent broad Mg II $\lambda\lambda$2796, 2803, redshifted to $z \sim 0.761$.

The amount of variability can be described by the `fractional_variability_g` parameter (see Sect. 2). Figure 16 shows its distribution for the GLEAN, CANOE, and *Gaia*-CRF3 sources: the peak value for the three samples is similar, and indicates variability at a level of 7%–8%. Only a small minority of objects have values larger than 20%. These results appear to be in agreement with those obtained by Berghea et al. (2021), who analysed the optical variability properties of 2863 sources belonging to the radio International Celestial Reference Frame 3 (ICRF3) with Pan-STARRS DR2 data.

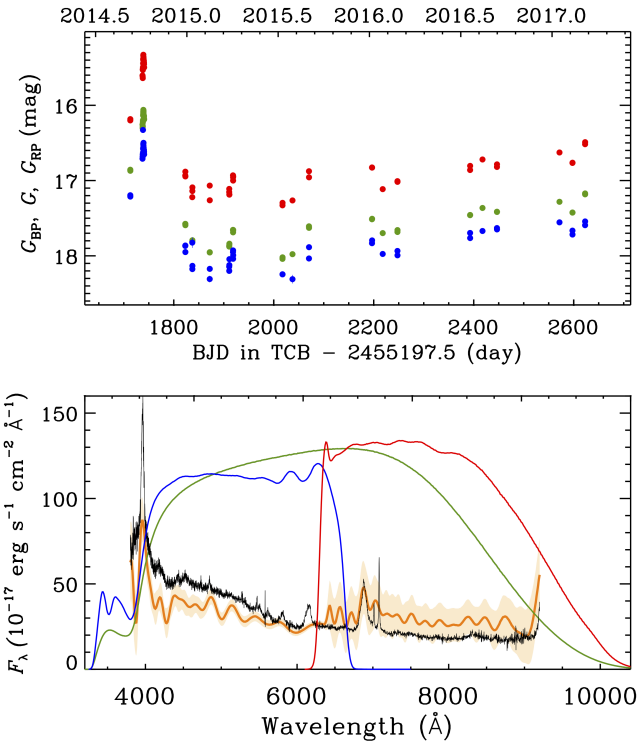


Fig. 13. As in Fig. 12, but for the FSRQ 5BZQJ1549+0237 (*Gaia* DR3 source_id: 4423448219003043968).

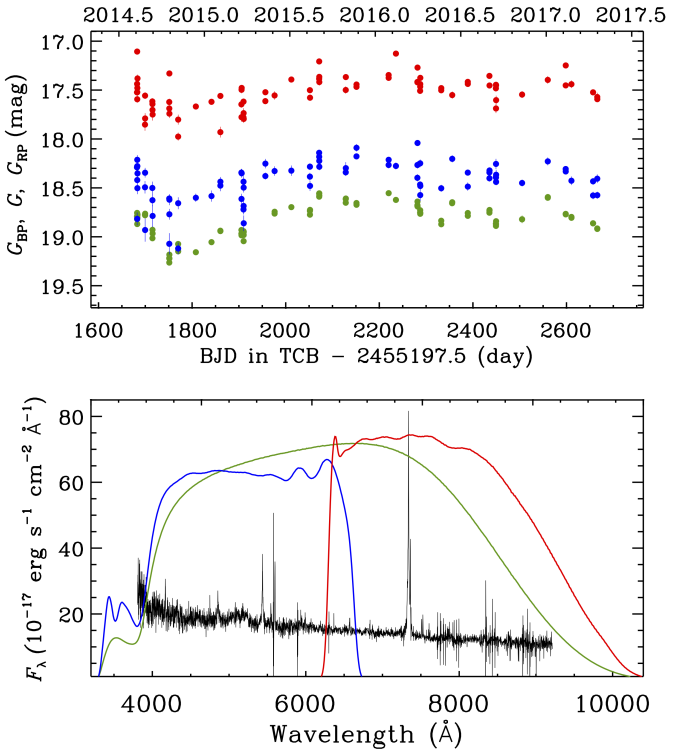


Fig. 14. As in Fig. 12, but for the narrow-line Seyfert 1 galaxy WISEA J133928.49+403229.9 (*Gaia* DR3 source_id: 1500096699133497600).

These latter authors found that the distribution of variability amplitudes is strongly skewed towards small values and peaks at about 0.1 mag.

We searched for infrared (IR) counterparts of our candidates in the AllWISE archive[6]. We considered a 3 arcsec search radius and stipulated a signal-to-noise ratio (S/N) of greater than 3 in the $W1$, $W2$, and $W3$ bands, obtaining 569 530 matches (53 144 of which are CANOE sources).

Figure 17 shows the location of the GLEAN sources in the WISE colour–colour diagram $W1-W2$ versus $W2-W3$, which is known to be a powerful tool to classify sources (see Sect. 1). The variable AGN candidates lie in the region where quasars and other types of AGN (e.g., blazars) are expected to be, confirming our selection. In particular, the CANOE sources are distributed in a somewhat smaller zone, suggesting that our selection procedure was very stringent, in line with the high-purity requirement. The 'blazar strip' (Massaro et al. 2012; Raiteri et al. 2014), connecting the locus of quasars with that of early-type galaxies and mostly populated by BL Lac objects, is clearly traced by sources belonging to the BZCAT5 catalogue. The plot also includes stellar objects of different types, which largely separate from the AGN candidates.

There is a fraction of AGN candidates (less than 12% of GLEAN and 51% of CANOE sources) with $W1-W2$ of less than 0.8, the threshold above which a genuine AGN should lie according to Stern et al. (2012). These sources are mostly faint objects, as shown in Fig. 18; about 92% of the GLEAN and 94% of the CANOE objects with $W1-W2<0.8$ have $G>19$. Moreover, as noted above, also many blazars, especially BL Lac objects, have $W1-W2<0.8$.

[6] https://irsa.ipac.caltech.edu/Missions/wise.html

Figure 19 shows the colour–colour plot $J-H$ versus $H-K$ of the 11 215 GLEAN sources (2514 in CANOE) with a near-IR counterpart in the Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) catalogue. These counterparts were obtained with a search radius of 3 arcsec and stipulating a S/N > 10. We notice a blob of CANOE sources with small values of both $H-K$ and $J-H$. Most of these bluer sources are faint in the *Gaia* G band (see Fig. 20).

There are 729 CANOE sources with $G>18$ and $H-K<0.3$ that have a WISE counterpart. Most of them lie in a thick strip in the WISE colour–colour diagram (Fig. 17), partly overlapping with the 'blazar strip', partly with the region populated by elliptical and spiral galaxies, and partly with stellar sources. This may mean that, notwithstanding all the filters we adopted, our sample is still contaminated by a small fraction of galaxies and stars. Overlaps with galaxies can be expected, as we can have very weak AGN drowned in galaxies. A search of the 729 sources in the *Gaia* DR3 catalogue of galaxies, containing more than 4.8 million sources (`galaxy_candidates` table), yielded 32 matches only. Moreover, 22 out of 729 sources have a radio counterpart, favouring an extragalactic nature.

The presence of a minor fraction of stellar contaminants is also suggested by the distributions of the *Gaia* astrometric parameters shown in Fig. 21. A small number excess characterises the tails (especially the low-side one) of the proper motion distributions of the GLEAN and CANOE samples (and of the newly identified AGN; see Sect. 6) with respect to the *Gaia*-CRF3 sample (see also Gaia Collaboration 2022; Liao et al. 2021).

We finally mention that the cross-match between the GLEAN sample and the *Gaia* DR3 `galaxy_candidates` table produces 16 854 overlaps. This is not surprising, as the host galaxy of many nearby AGN is expected to be detectable.

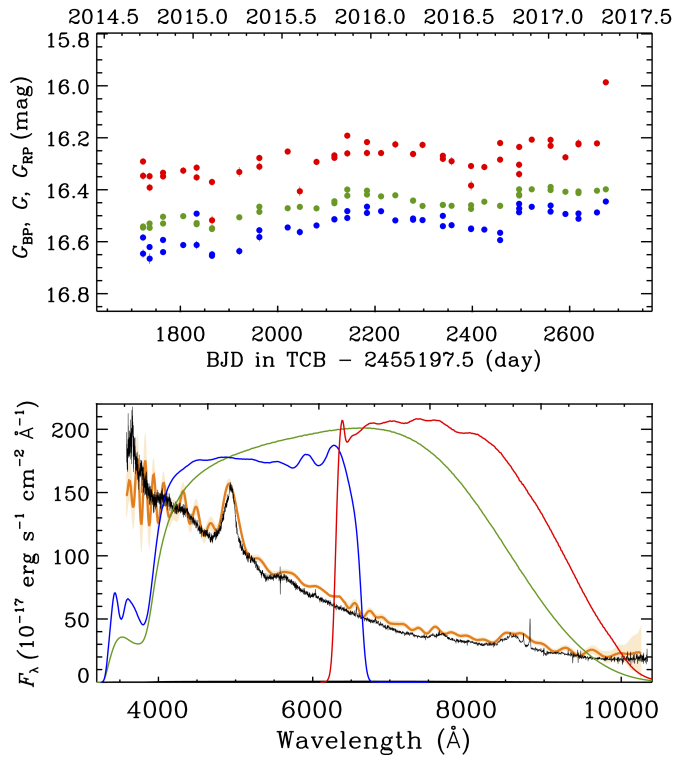


Fig. 15. As in Fig. 12, but for the radio-quiet quasar FBQS J163709.3+414030 (*Gaia* DR3 source_id: 1356927713819217664).

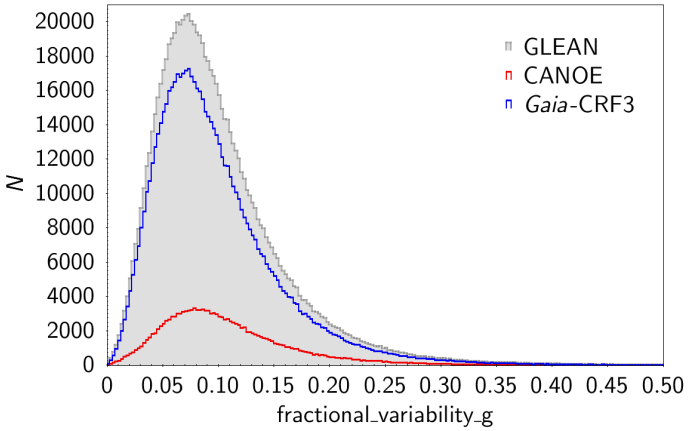


Fig. 16. Distribution of the `fractional_variability_g` parameter for the GLEAN, CANOE, and *Gaia*-CRF3 samples (bin width = 0.0025). The peaks indicate variability at a level of 7%–8%.

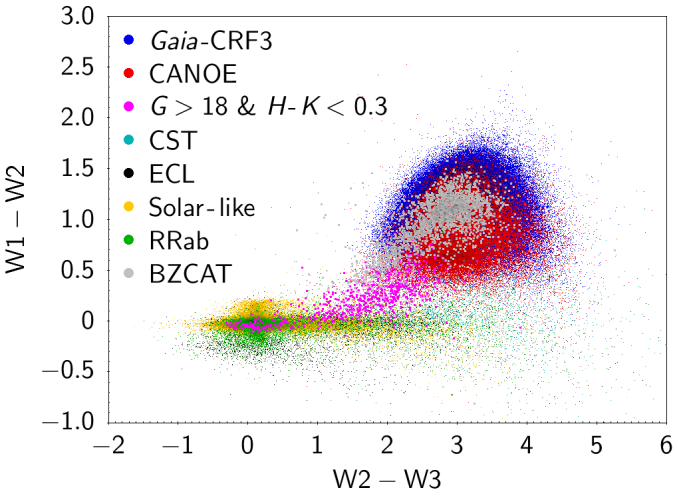


Fig. 17. WISE colour–colour diagram. The variable AGN candidates included in the *Gaia*-CRF3 sample are marked in blue, while the CANOE objects are in red. Blazars are shown with larger (grey) symbols to highlight the 'blazar strip', which extends from the quasar locus to the early-type-galaxy region. Different types of stellar objects are also shown (see Gavras et al. 2023): constant stars (CST), eclipsing binaries (ECL), solar-like stars (with spots and flares), and ab-type RR Lyrae stars (RRab). Sources with $G > 18$ and $H - K < 0.3$ are discussed in the text.

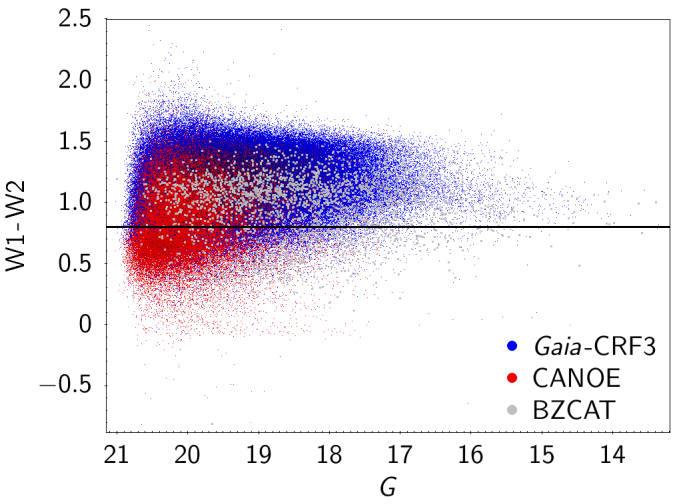


Fig. 18. WISE colour index $W1 - W2$ versus *Gaia* G-band magnitude. The horizontal line indicates the threshold $W1 - W2 = 0.8$ above which a source is expected to be a genuine quasar. Most of the sources below this line are very faint objects.

5. Check for stellar contaminants

To assess the possible presence of stellar contaminants, we cross-matched the GLEAN sample with various catalogues. We find that about 12 156 sources (only 1 896 in the CANOE sample) are included in the *Gaia* DR2 catalogue of white dwarfs (WDs) from Gentile Fusillo et al. (2019). However, verifying the P_{WD} parameter, which gives the probability of being a WD, reveals that about 95% of the sources have $P_{WD} < 0.1$ (see Fig. 22), very far from the request $P_{WD} > 0.75$ adopted in this study for high-confidence WD candidates. There are only 24 objects (9 in CANOE) with $P_{WD} > 0.75$. The inspection of the light curves of the nine CANOE objects with P_{WD} greater than 0.75 reveals long-term variability compatible with an AGN behaviour.

The cross-match with the more recent catalogue of white dwarfs in *Gaia* EDR3 by Fusillo et al. (2021) leads to 55 common objects. Only 10 of these (1 in CANOE) have $P_{WD} > 0.75$, and their corresponding light curves are compatible with an AGN behaviour.

The cross-match between the GLEAN sources and the PS1 sample of RR Lyrae stars (Sesar et al. 2017) yielded 1 319 common objects (about 388 in CANOE). The distribution of their RRab and RRc classification scores is plotted in Fig. 23 and indicates that most sources have a low probability of being an RR Lyrae star. However, there are 21 objects with $\text{score}_{3,ab} > 0.8$ and 40 objects with $\text{score}_{3,c} > 0.55$, which are the limits

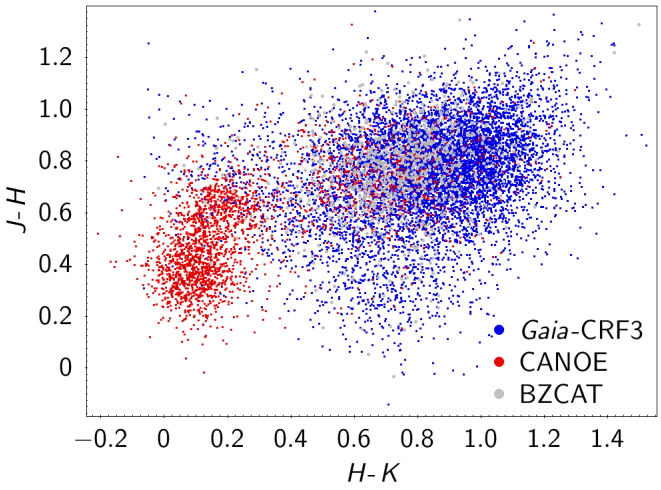


Fig. 19. Colour–colour plot of the 11 215 GLEAN sources with a near-IR counterpart in the 2MASS catalogue.

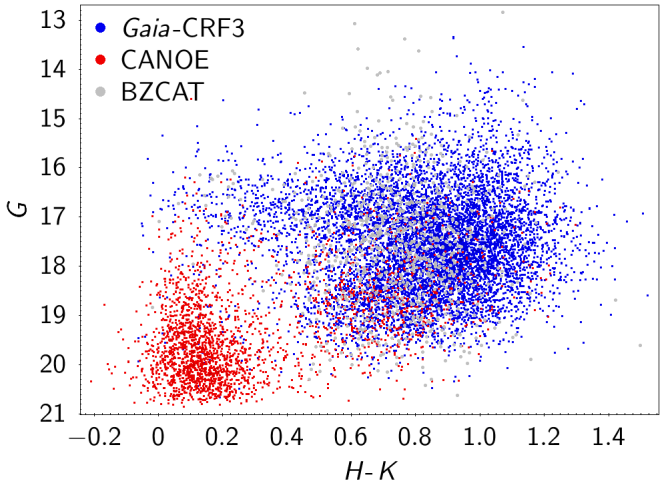


Fig. 20. *Gaia G* magnitude versus the $H - K$ 2MASS colour index.

indicating a high probability of being a RR Lyrae star. All these 61 objects belong to the *Gaia*-CRF3 sample and most of them have variability trends in agreement with those of AGN.

We also found 385 sources in GLEAN that are classified as young stellar objects (YSOs) in the All-Sky Automated Survey for Supernovae (ASAS-SN) catalogue of variable stars (Jayasinghe et al. 2020), but only 4 with a high probability (greater than 0.75) of being a YSO. Further cross-matching with other catalogues of variable stars yielded no significant overlap.

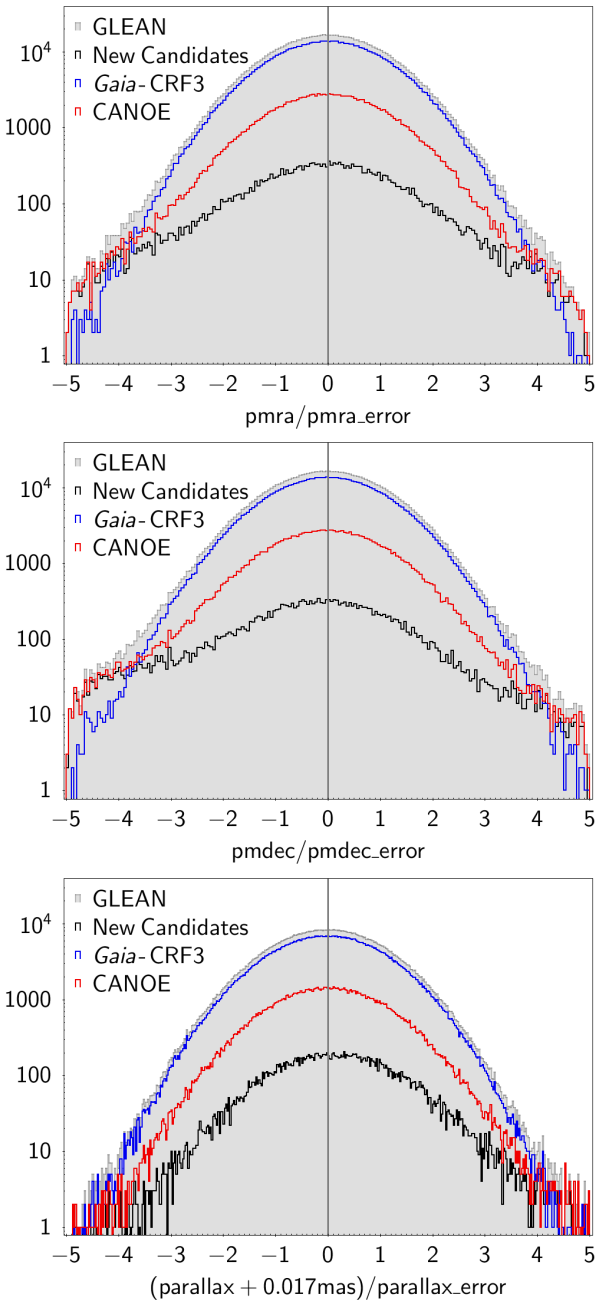


Fig. 21. Distributions of proper motion in right ascension (*top*), proper motion in declination (*middle*), and parallax (*bottom*) for the various variable AGN samples discussed in the paper.

6. Completeness and purity

We estimate the completeness and purity of the GLEAN sample we have selected, taking into account that this is not a general AGN sample, but a sample of AGN that are observed to be variable. The application of the GVD module to *Gaia*-CRF3 showed that 88% of AGN are detected as variable in *Gaia*-DR3. This is in reasonable agreement with the results of Sesar et al. (2007), who reported that ≳90% of the quasars in the Stripe 82 sky region with multiple photometric observations by the SDSS are variable at the 0.03 mag level.

We first calculated the GLEAN sample completeness with respect to the SDSS DR16Q v4 catalogue (Lyke et al. 2020), which is 99.8% complete and has only 0.3%–1.3% contamination. Because the SDSS covers only part of the sky, we selected a wide sampled region within +10 deg < dec < +50 deg and 130 deg < ra < 220 deg. We found 224 752 DR16Q sources in this area, 145 669 of which have a *Gaia* counterpart in the catalogue, and 151 915 in the *Gaia*-DR3. In line with the GVD result mentioned above, we assume that 88% of them are variable, which equates to 133 685 objects. In the same sky region, we find 62 696 sources belonging to the GLEAN sample. Therefore, we can estimate a 47% completeness of the GLEAN sample when taking the DR16Q catalogue as reference. Vice versa, there are 38 650 GLEAN (5205 CANOE) sources in the same sky region that are not included in the DR16Q catalogue.

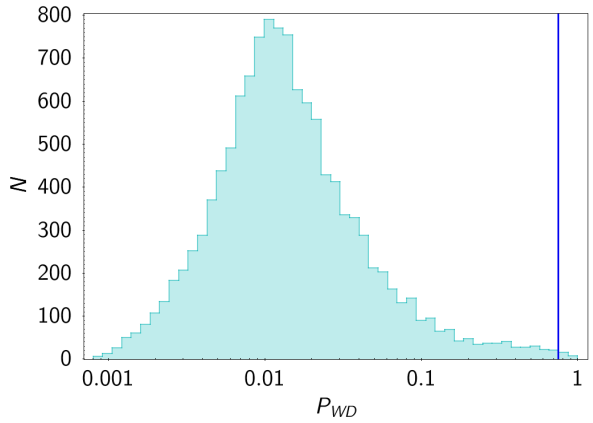


Fig. 22. Distribution of $P_{\rm WD}$ for the ∼12 000 sources of the GLEAN sample with a counterpart in the *Gaia* DR2 catalogue of white dwarfs by Gentile Fusillo et al. (2019). The vertical line highlights $P_{\rm WD}$ = 0.75, which is the threshold above which a source can be considered a high-confidence WD in the paper.

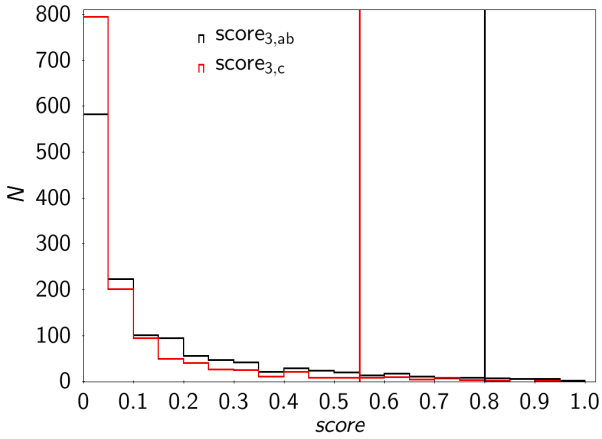


Fig. 23. Distribution of the RRab and RRc classification scores, $score_{3,ab}$ (black line), and $score_{3,c}$ (red line) for the 1319 sources in the GLEAN sample with a counterpart in the catalogue of RR Lyrae stars by Sesar et al. (2017). Vertical lines indicate the limits $score_{3,ab}$ = 0.8 and $score_{3,c}$ = 0.55, above which sources have a high probability of being RR Lyrae stars.

We then estimated the completeness with respect to the *Gaia*-CRF3 sample, which we have used as a reference for the selection procedure, assuming that it contains genuine AGN. More precisely, the contamination of the *Gaia*-CRF3 sample is expected to be at most 2% (Gaia Collaboration 2022). As before, we assume that the percentage of variable objects is 88% of the whole sample, and so we can consider that among the 1 614 173 sources in *Gaia*-CRF3, 1 420 472 are variable. On the other hand, the number of *Gaia*-CRF3 sources that survived the selection procedure and are present in the GLEAN sample is 722 211. Therefore, we can estimate a completeness of 51%. We analysed the variation of completeness with the G magnitude. Figure 24 shows the ratio between the number of *Gaia*-CRF3 sources that survived the selection procedure and the number of variable sources in the *Gaia*-CRF3 sample per magnitude bins. This reveals that the completeness of the final sample is above 90% for the sources brighter than about G=16 and then decreases in an irregular way with increasing magnitude. It is still about 50% at G = 20–20.5, and then falls rapidly.

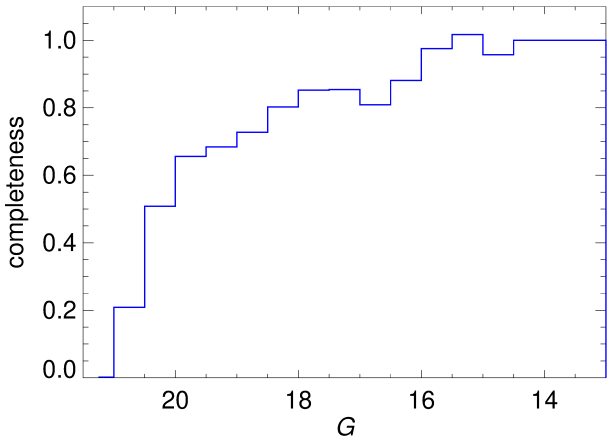


Fig. 24. GLEAN sample completeness estimated with respect to the *Gaia*-CRF3 sample versus G magnitude.

In addition, we estimated the percentage of sources that survived the series of cuts described in Sect. 3 both in the case of the *Gaia*-CRF3 catalogue and for several large external AGN catalogues (with more than 10 000 sources). The results are reported in Table 1. Columns indicate the catalogue name, the number of sources $N_{\rm cat}$ in each catalogue, the number of matches between the catalogue sources and the initial 34 million variable sources selected by the SOS-AGN module $N_{\rm match,ini}$, the number of matches with the GLEAN sample $N_{\rm match,fin}$, the ratio $N_{\rm match,fin}/N_{\rm match,ini}$, and the number of matches with the CANOE sample $N_{\rm match,new}$. The ratio $N_{\rm match,fin}/N_{\rm match,ini}$ can be seen as an estimate of the filter survival fraction of the variable sources in that catalogue[7].

The number of *Gaia*-CRF3 sources in the initial sample of 34 million candidates is 1 141 892, of which 722 211 are included in the GLEAN sample. This gives a filter survival percentage of about 63%.

The largest catalogues, containing more than 500 000 sources, give in general a filter survival percentage of roughly between 60% and 70%, with an average value of 65%. This estimate is also in agreement with those inferred by considering the 'QSO' objects in the APOP catalogue and the e-ROSITA AGN catalogue. The filter survival percentage derived from the other smaller catalogues is higher, ranging from about 70% to almost 80%.

The purity of the GLEAN sample is the number of genuine variable AGN included in it over the total number of GLEAN objects. A lower limit to the purity of the GLEAN sample can be obtained from the ratio between the number of *Gaia*-CRF3 sources in the sample and the total number of sources in the sample, which is around 83%. However, as derived from the cross-match with the catalogues in Table 1, 128 282 of the ∼150 000 CANOE objects are present in other AGN catalogues. This in principle raises the purity lower limit of the GLEAN sample to about 97%. However, as we cannot exclude that the common sources still include contaminants, we conservatively estimate the sample purity to be around 95%.

From the cross-match with the AGN catalogues in Table 1, we find that 21 735 sources are new AGN candidates. The distribution of astrometric parameters of these new AGN candidates is

[7] Here we did not take into account that less than 3000 sources in GLEAN did not pass the filter selection due to their peculiarities (see Sect. 3.3). However, these represent less than 0.3% of the sample, and so they cannot significantly change the filter survival fraction estimates.

Table 1. Results of the cross-match of the GLEAN and CANOE samples with large AGN catalogues.

Catalogue	N_{cat}	$N_{match,ini}$	$N_{match,fin}$	Ratio	$N_{match,new}$	Reference
WISE C75	20 907 127	925 988	602 782	0.65	28 574	1
WISE R90	4 543 530	788 714	526 388	0.67	4 569	1
Gaia–unWISE	2 734 464	1 363 764	819 677	0.60	108 345	2
Gaia-DR2	2 690 021	1 025 599	652 900	0.64	85 662	3
***Gaia*-CRF3**	**1 614 173**	**1 141 892**	**722 211**	**0.63**	**0**	**4**
SDSS DR16Q Superset v3	1 440 615	295 426	196 907	0.67	1 360	5
AllWISE AGN	1 354 775	494 067	323 225	0.65	2 316	6
MILLIQUAS	1 115 619	411 742	273 422	0.66	5 545	7
SDSS DR16Q v4	750 414	291 484	195 946	0.67	1 201	5
LQAC5	592 809	259 127	174 076	0.67	89	8
LQRF	100 165	81 560	61 159	0.75	40	9
BROS	88 211	6 304	4 510	0.72	304	10
APOP (QSO)	86 821	72 107	54 407	0.63	33	11
LAMOST5	52 453	38 341	29 188	0.76	92	12
2QZ	49 425	19 254	13 794	0.72	184	13
e-ROSITA	21 952	5 122	3 319	0.65	289	14
OCARS	13 589	6 541	5 099	0.78	147	15
Seyfert	11 101	7 802	5 578	0.71	26	16

References. (1) Assef et al. (2018); (2) Shu et al. (2019); (3) Bailer-Jones et al. (2019); (4) Gaia Collaboration (2022); (5) Lyke et al. (2020); (6) Secrest et al. (2015); (7) Flesch (2021); (8) Souchay et al. (2019); (9) Andrei et al. (2009); (10) Itoh et al. (2020); (11) Qi et al. (2015); (12)Yao et al. (2019); (13) Croom et al. (2004); (14) Liu et al. (2022); (15) Malkin (2018); (16) Rakshit et al. (2017).

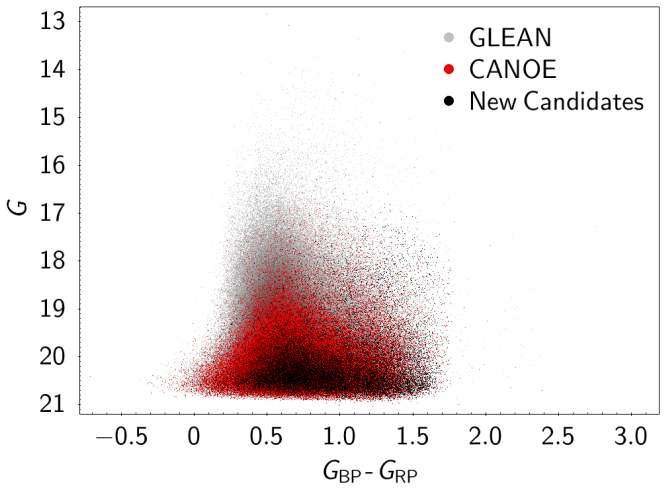


Fig. 25. *Gaia G* versus $G_{BP} - G_{RP}$ colour index for the sources in the GLEAN and CANOE samples, and for the new variable AGN candidates.

shown in Fig. 21, while Fig. 25 displays their colour–magnitude diagram, G versus $G_{BP} - G_{RP}$. The new sources approximately cover the same range of $G_{BP} - G_{RP}$ colour indices as the GLEAN and CANOE objects, but they lie among the faintest sources and tend to avoid the region of the bluest colours. The excess in large negative proper motions discussed above appears to be largely due to these new sources, and the effect could be related to their faintness, though we cannot rule out a certain percentage of stellar contamination.

7. Cross-match with radio catalogues

As mentioned in Sect. 1, a fraction of AGN are radio-loud. This fraction is generally assumed to be around 10%, but actually diminishes with increasing redshift and decreasing luminosity

Table 2. Results of the cross-match between the GLEAN sources and radio catalogues.

Name	Band (GHz)	Sky (%)	N_{cat}	N_{cross}
FIRST	1.4	25.6	946 432	13 133
VLASS	3.0	82	3 381 277	31 378
NVSS	1.4	82	1 773 484	11 041

(Jiang et al. 2007; Kratzer & Richards 2015). We cross-matched the GLEAN sample with the catalogues of the radio sky surveys FIRST (Gordon et al. 2021), NVSS (Condon et al. 1998), and VLASS (Lacy et al. 2020) using a 1.5 arcsec radius. Table 2 shows, for each catalogue, the observing radio frequency, the percentage of the sky covered, the number of objects N_{cat} in the catalogue, and the number of GLEAN sources N_{cross} with a radio counterpart. The distribution on the sky of the GLEAN-radio pairs is plotted in Fig. 26. Figure 27 shows the distribution of radio fluxes, highlighting the greater depth of the FIRST and VLASS catalogues with respect to NVSS and the much larger number of objects in the VLASS. The number of non-duplicated variable AGN candidates with radio counterparts is 33 706, which represents about 4% of the GLEAN sample.

Under the assumption that the 1.4–3.0 GHz spectrum can be approximated by a power law $F_\nu \propto \nu^{-\alpha}$, we calculated the 1.4–3.0 GHz spectral index for the 17 399 counterparts of the GLEAN sources with radio data in both bands. When 1.4 GHz information from both FIRST and NVSS was available, we chose the latter, meaning that we used NVSS for about 58% of the sources. The spectral index is plotted in Fig. 28; its median value is 0.39 and the standard deviation 0.88. The median value does not change significantly if we set a lower limit of 3 mJy or even 10 mJy to the VLASS flux. By comparing these results to those by Gordon et al. (2021), we find good agreement, taking into account that we are mostly dealing with compact sources.

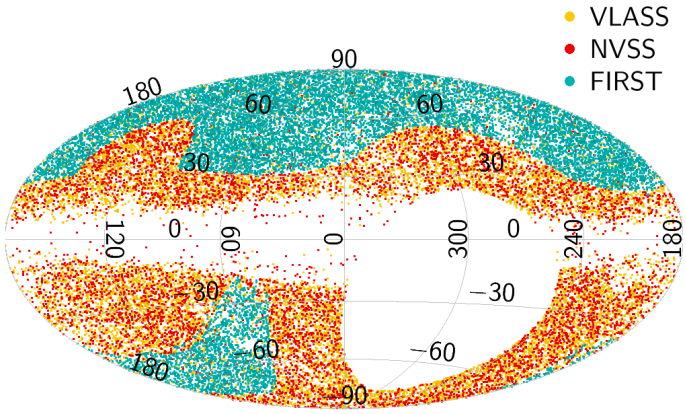


Fig. 26. Distribution on the sky of the radio counterparts of the GLEAN sources in the FIRST (cyan), NVSS (red), and VLASS (orange) catalogues, in Galactic coordinates.

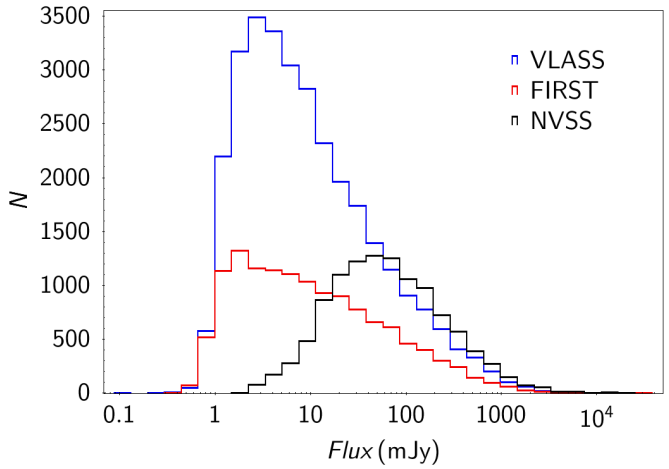


Fig. 27. Radio flux densities (mJy) of the counterparts of the GLEAN sources in the FIRST (red), NVSS (black), and VLASS (blue) catalogues.

If we apply the classical condition $\alpha < 0.5$ to define a flat spectrum (e.g., Urry & Padovani 1995), we find 9949 sources (57%) with a flat spectrum, which is a distinctive feature of a blazar source. A reliable spectral index for a variable source should preferentially be calculated with contemporaneous data in the two bands. Here this is not possible, and this must be kept in mind when evaluating the results. For instance, blazars have flat radio spectra, and indeed ~75% of the 2058 confirmed blazars in the BZCAT5 catalogue for which we could estimate the radio spectral index show values smaller than 0.5 (see Fig. 28), but still ~25% of blazars display a steep spectrum.

Spectral indices with non-contemporaneous data have been used to identify blazar candidates, as in the cases of the CRATES (Healey et al. 2007) and BROS (Itoh et al. 2020) catalogues. In particular, the selection criterion for the BROS blazars was to have $\alpha_{\rm radio} < 0.6$, as derived from the Fermi 4LAT sources (Abdollahi et al. 2020), and the spectral index was obtained using radio data from 0.15 GHz TGSS (Intema et al. 2017) and 1.4 GHz NVSS catalogues. However, among the 4209 BROS expected 'flat-spectrum' sources in Fig. 28, only 2327 (55%) have a flat spectrum according to our criterion. In the above discussion, we have assumed that the broad-band radio spectrum

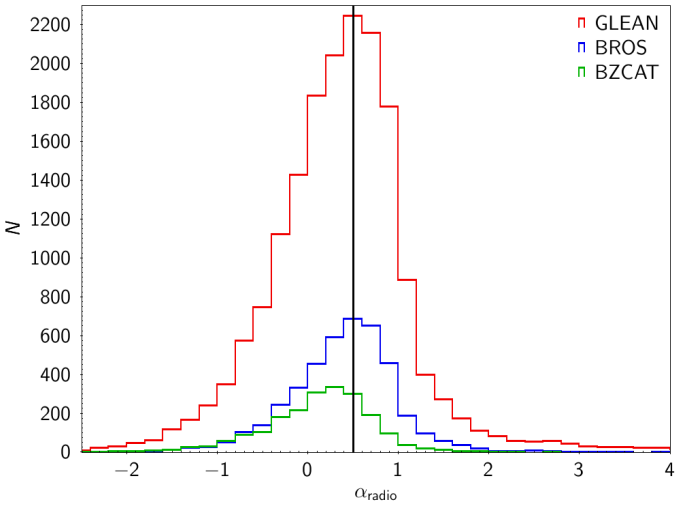


Fig. 28. Distribution of the 1.4–3.0 GHz spectral index for the 17 399 radio counterparts of the GLEAN variable AGN (red) for the 2058 of them that are included in the BZCAT5 catalogue of confirmed blazars (green), and for the 4209 blazar candidates in the BROS catalogue (blue). The vertical line indicates the value 0.5 below which a radio spectrum is defined as 'flat'.

can be approximated by a power law. Deviations from a power-law SED would modify the above numbers.

We investigated the percentage of radio-loud sources in our sample. The classical definition of a radio-loud source is that $R = F_{5\,\rm GHz}/F_B > 10$ (e.g., Urry & Padovani 1995), where $F_{5\,\rm GHz}$ and F_B are the flux densities at 5 GHz and in the optical B band, respectively. For the 17 399 sources for which $\alpha_{\rm radio}$ could be estimated, the $F_{5\,\rm GHz}$ flux density was derived from that at 3.0 GHz in the hypothesis that the estimated $\alpha_{\rm radio}$ also fairly describes the 3–5 GHz spectrum.

In order to calculate F_B, we made the assumption that we can also approximate the spectrum with a power law in the optical. Therefore, we first obtained Johnson-Cousins V and R magnitudes from *Gaia* magnitudes according to the relationships provided by Riello et al. (2021). We then calculated the corresponding flux densities using the zeropoints by Bessell et al. (1998), and then corrected them for Galactic reddening according to Schlegel et al. (1998) and Fitzpatrick (1999). The resulting optical spectral index $\alpha_{\rm opt}$ is shown in Fig. 29. The average value is 0.64 ± 0.77. Finally, for each source we derived F_B from F_V using its own $\alpha_{\rm opt}$.

The distribution of the radio-loudness parameter R is shown in Fig. 30. The number of radio-loud sources is 16 459, which represents 95% of the 17 399 sources for which we could calculate a radio spectral index. If we simply generalised this result to all sources with a radio counterpart, taking into account the different sky coverage of *Gaia* with respect to the radio surveys, we would infer that the number of radio-loud sources in our GLEAN sample is of the order of 4%.

8. Lensed quasars

The GLEAN sample includes more than 100 known gravitationally lensed quasars[8]. We investigated the possibility of deriving robust measurements of the time lag between the observed flux variations corresponding to the various images of a lensed

[8] For a complete list, see the Gravitationally Lensed Quasar Database at https://research.ast.cam.ac.uk/lensedquasars/

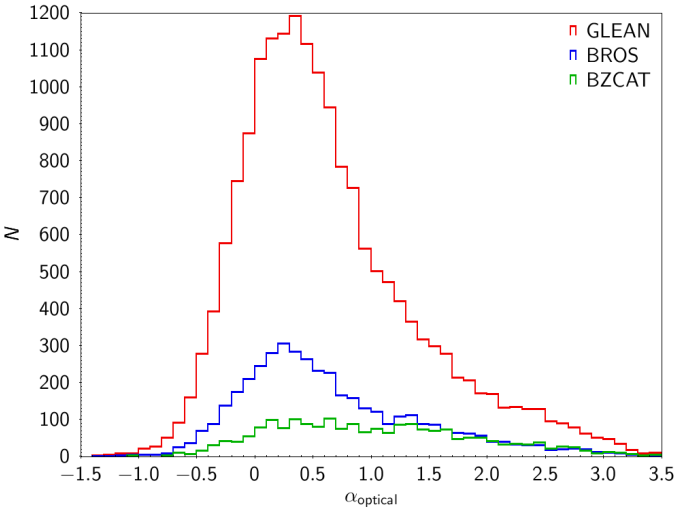


Fig. 29. Same as Fig. 28 for the optical spectral index.

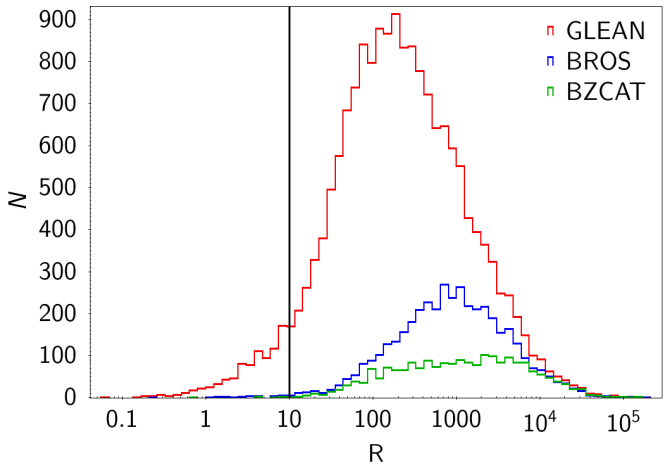


Fig. 30. As Fig. 28 for the distribution of the radio-loudness parameter R. The vertical line indicates the value $R = 10$, above which a source is classically defined as 'radio-loud'.

quasar, which is the first step that can lead to the determination of the value of the Hubble constant (e.g., Tewes et al. 2013; Wong et al. 2020, and references therein). This is a difficult task, because quasars are characterised by smooth variability on timescales of months and because microlensing by the stars of the lensing galaxy can produce additional features, which are different in the light curves of the various images. Long-term monitoring with good sampling is therefore necessary to match the light curve of one image with that of another image through the application of the right shift in time and brightness. The detection of well-defined characteristic patterns of variability substantially improves the time-lag estimate. We find such an example in the double-lensed quasar DESJ0501-4118 (Lemon et al. 2019), which is shown in Fig. 31. The characteristic variability behaviour, with a double bump in the light curve of the brighter image (image 1), which can be recognised in the light curve of the fainter image (image 2) after some delay, means that a robust time-lag determination could indeed be possible. Microlensing effects by stars within the lensing galaxy seem important here and, as mentioned before, can explain differences between the two light curves that cannot be accounted for by shifts in time and magnitude. Because of these effects, the simple application of a discrete correlation function (DCF; Edelson & Krolik 1988; Hufnagel & Bregman 1992), a method which was specifically designed to cross-correlate unevenly sampled data trains, gives somewhat unstable results, which depend on the DCF time-lag bin. A detailed treatment of the microlensing effects is beyond the scope of this paper. However, an estimate of the time lag can be obtained by considering cubic spline interpolations through the binned light curves, which highlight the long-term trend while smoothing the short-term oscillations.

We first calculated the cubic spline interpolation through the 30-day binned light curve of image 1 ($G_{1,\mathrm{spline}}$). We then shifted this spline by a quantity τ in time and a quantity ζ in magnitude to find the values of these two parameters that lead to the best match with the light curve of image 2 – whose data points $G_2(t_i)$ have errors $\sigma_2(t_i)$; these minimise the reduced chi-squared (with ν degrees of freedom):

$$\frac{\chi^2}{\nu} = \frac{1}{N-1}\sum_{ij}\left(\frac{G_{1,\mathrm{spline}}(t_j+\tau)+\zeta-G_2(t_i)}{\sigma_2(t_i)}\right)^2 \quad (8)$$

for all N pairs ij of points that are separated by no more than 5 days, that is, for which $|t_j + \tau - t_i| < 5$. The result was $\tau = 121$ days and $\zeta = 0.21$ mag. Figure 31 shows the match between the data of image 1 and image 2 when the former is shifted by 121 days and 0.21 mag. Decreasing the spline bin to 20 days does not change the results, while increasing it to 40 days leads to $\tau = 120$ days, but in both cases the χ^2/ν increases with respect to the 30-day bin.

The DCF between the two light curves and the one between the two splines (see Fig. 31) show a peak at $\tau = 120$ days, but the centroid indicates a somewhat smaller time lag: about 117 days for the DCF on the light curves, and 119 days for that on the splines.

To determine the uncertainty on the time lag, we ran 3000 'flux randomisation–random subset selection' Monte-Carlo simulations (Peterson et al. 1998; Raiteri et al. 2003). The distribution of the lag centroids is shown in Fig. 31; in 68% of cases (1σ) the delay is between 119 and 120.4 days. Altogether, we conclude that the brightness variations of image 2 follow those of image 1 with a time lag of 119–121 days.

9. Summary and conclusions

We present the *Gaia* SOS-AGN module included in the variability analysis pipeline, and the subsequent procedure to select variable AGN candidates. The result is a high-purity variable AGN sample (GLEAN), including more than 872 000 sources. Starting from initial requirements (more than 20 FoV transits in the G band light curve and having some variability metrics defined), the following filters were tailored on the *Gaia*-CRF3 sample and included cuts on the structure function index, the Butler & Bloom (2011) statistics, colour indices, parallax, proper motion, and environment density. We also introduced filters on the effect of scan-angle variations and on the GVD variability probability to avoid contamination by artificially variable nearby galaxies. We notice that the upstream module of General Supervised Classification includes galaxies detectable from their spurious variable signal, stars (in 23 types), and AGN as target categories. Sources with spurious variability due to scan-angle variations are expected to be assigned to the galaxy class, with the AGN class mostly retrieving the extragalactic sources dominated by the emission from an active nucleus (Eyer et al. 2023;

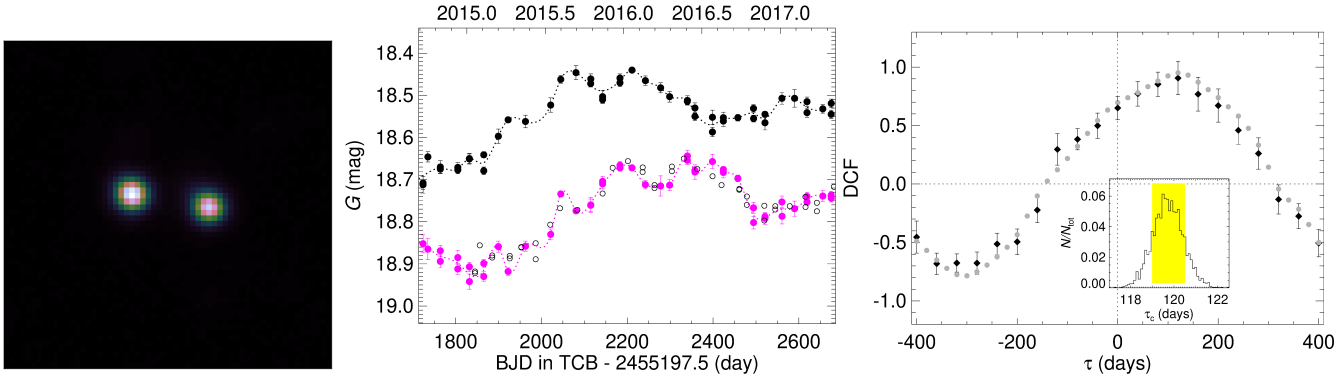


Fig. 31. *Left*: Dark Energy Survey (DES) *g*-band image of the lens system DESJ0501-4118. *Middle*: *Gaia G*-band light curves of image 1 (black dots) and image 2 (magenta dots); the empty circles represent the image 1 light curve shifted in time by 121 days and in brightness by 0.21 mag to match the behaviour of image 2. Dotted lines are cubic spline interpolations through the 30-day binned light curves. *Right*: DCF between the image 1 and image 2 light curves (black diamonds) and between their splines (grey dots); they indicate a time delay of ∼120 days of the flux variations of image 2 with respect to those of image 1. The inset shows the result of 3000 Monte Carlo DCF simulations; the yellow strip highlights the interval of the time-lag centroid values including 68% of cases (1σ).

Rimoldini et al. 2023). Moreover, the selection of variable AGN presented in this paper is tailored to the *Gaia*-CRF3 sample, which includes mostly AGN dominated by the nucleus, rather than by the galaxy. In addition, our sources are characterised by good astrometric solutions, while galaxies dominated by artificial variability have in general astrometric solutions of lower quality. We also note that a small contribution from the artificial variability of the host galaxy would in any case be diluted from the AGN contribution. In conclusion, we expect that only a minor fraction of sources in our sample may be affected in a sensitive way by artificial variability introduced by an extended host galaxy. All filters are based on *Gaia* data only. The GLEAN sample has a 47% completeness when we take the SDSS DR16Q quasar catalogue as a reference, assuming that 88% of the sources are variable. The completeness estimated as the percentage of *Gaia*-CRF3 variable AGN identified by our selection procedure with respect to those in the complete sample is 51%. We find that this value strongly depends on magnitude. We further evaluated the specific impact of the series of filters applied to the sources selected by the SOS-AGN module. When considering the *Gaia*-CRF3 sample, the filter survival percentage is about 63%, which means the cuts are responsible for the removal of 37% of the candidates. Taking into account other large AGN catalogues, the cut survival percentage ranges between about 60% and 80%. The purity of the GLEAN sample is conservatively estimated to be higher than 95%. This result comes from both the comparison with other AGN catalogues and a careful investigation of possible contaminants.

We discuss the properties of the selected AGN, complementing *Gaia* data with data from near-IR, mid-IR, and radio surveys. In particular, we estimate that about 4% of the selected sources are radio-loud according to the classical definition.

Finally, we show the potentiality of *Gaia* light curves to estimate the time lags between the flux variations of the multiple images of lensed quasars. This goal would be more easily achieved by merging *Gaia* data with other datasets.

Acknowledgements. This work presents results from the European Space Agency (ESA) space mission *Gaia*. *Gaia* data are being processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the *Gaia* MultiLateral Agreement (MLA). The *Gaia* mission website is https://www.cosmos.esa.int/gaia. The *Gaia* archive website is https://archives.esac.esa.int/gaia. Full acknowledgements are given in Appendix A.

References

Abdollahi, S., Acero, F., Ackermann, M., et al. 2020, ApJS, 247, 33
Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42
Andrei, A. H., Souchay, J., Zacharias, N., et al. 2009, A&A, 505, 385
Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, ApJ, 772, 26
Assef, R. J., Stern, D., Noirot, G., et al. 2018, ApJS, 234, 23
Bailer-Jones, C. A. L., Fouesneau, M., & Andrae, R. 2019, MNRAS, 490, 5615
Berghea, C. T., Makarov, V. V., Quigley, K., & Goldman, B. 2021, AJ, 162, 21
Bessell, M. S., Castelli, F., & Plez, B. 1998, A&A, 333, 231
Butler, N. R., & Bloom, J. S. 2011, AJ, 141, 93
Carrasco, J. M., Weiler, M., Jordi, C., et al. 2021, A&A, 652, A86
Condon, J. J., Cotton, W. D., Greisen, E. W., et al. 1998, AJ, 115, 1693
Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2004, MNRAS, 349, 1397
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products
De Angeli, F., Weiler, M., Montegriffo, P., et al. 2023, A&A, 674, A2 (*Gaia* DR3 SI)
Edelson, R. A., & Krolik, J. H. 1988, ApJ, 333, 646
Evans, D.W., Eyer, L., Busso, G., et al. 2023, A&A, 674, A4 (*Gaia* DR3 SI)
Eyer, L. 2002, Acta Astron., 52, 241
Eyer, L., & Genton, M. G. 1999, A&AS, 136, 421
Eyer, L., Audard, M., Holl, B., et al. 2023, A&A, 674, A13 (*Gaia* DR3 SI)
Fitzpatrick, E. L. 1999, PASP, 111, 63
Flesch, E. W. 2015, PASA, 32, 10F
Flesch, E. W. 2021, ArXiv e-prints [arXiv:2105.12985]
Fu, Y., Wu, X.-B., Yang, Q., et al. 2021, ApJS, 254, 6
Fusillo, N. P. G., Tremblay, P. E., Cukanovaite, E., et al. 2021, MNRAS, 508, 3877
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Gaia Collaboration (Klioner, S. A., et al.) 2021, A&A, 649, A9
Gaia Collaboration (Klioner, S. A., et al.) 2022, A&A, 667, A148
Gaia Collaboration (Bailer-Jones, C. A. L., et al.) 2023, A&A, 674, A41 (*Gaia* DR3 SI)
Gavras, P., Rimoldini, L., Nienartowicz, K., et al. 2023, A&A, 674, A22 (*Gaia* DR3 SI)
Gentile Fusillo, N. P., Tremblay, P.-E., Gänsicke, B. T., et al. 2019, MNRAS, 482, 4570
Gordon, Y. A., Boyce, M. M., O'Dea, C. P., et al. 2021, ApJS, 255, 30
Healey, S. E., Romani, R. W., Taylor, G. B., et al. 2007, ApJS, 171, 61
Holl, B., Sozzetti, A., Sahlmann, J., et al. 2023, A&A, 674, A10 (*Gaia* DR3 SI)
Hufnagel, B. R., & Bregman, J. N. 1992, ApJ, 386, 473
Hughes, P. A., Aller, H. D., & Aller, M. F. 1992, ApJ, 396, 469
Intema, H. T., Jagannathan, P., Mooley, K. P., & Frail, D. A. 2017, A&A, 598, A78
Itoh, R., Utsumi, Y., Inoue, Y., et al. 2020, ApJ, 901, 3

Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2020, VizieR Online Data Catalog, II/366
Jiang, L., Fan, X., Ivezić, Ž., et al. 2007, ApJ, 656, 680
Kratzer, R. M., & Richards, G. T. 2015, AJ, 149, 61
Lacy, M., Baum, S. A., Chandler, C. J., et al. 2020, PASP, 132, 1009
Lemon, C. A., Auger, M. W., & McMahon, R. G. 2019, MNRAS, 483, 4242
Liao, S., Wu, Q., Qi, Z., et al. 2021, PASP, 133, 094501
Lindegren, L., Bastian, U., Biermann, M., et al. 2021a, A&A, 649, A4
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021b, A&A, 649, A2
Liu, T., Buchner, J., Nandra, K., et al. 2022, A&A, 661, A5
Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8
MacLeod, C. L., Brooks, K., Ivezić, Ž., et al. 2011, ApJ, 728, 26
Malkin, Z. 2018, ApJS, 239, 20
Massaro, F., D'Abrusco, R., Tosti, G., et al. 2012, ApJ, 750, 138
Massaro, E., Maselli, A., Leto, C., et al. 2015, Ap&SS, 357, 75
Mateos, S., Alonso-Herrero, A., Carrera, F. J., et al. 2012, MNRAS, 426, 3271
Montegriffo, P., De Angeli, F., Andrae, R., et al. 2023, A&A, 674, A3 (*Gaia* DR3 SI)
Peterson, B. M., Wanders, I., Horne, K., et al. 1998, PASP, 110, 660
Qi, Z., Yu, Y., Bucciarelli, B., et al. 2015, AJ, 150, 137
Raiteri, C. M., Villata, M., Tosti, G., et al. 2003, A&A, 402, 151
Raiteri, C. M., Villata, M., Carnerero, M. I., et al. 2014, MNRAS, 442, 629
Raiteri, C. M., Villata, M., Acosta-Pulido, J. A., et al. 2017, Nature, 552, 374
Rakshit, S., Stalin, C. S., Chand, H., & Zhang, X.-G. 2017, ApJS, 229, 39
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Richards, G. T., Myers, A. D., Gray, A. G., et al. 2009, ApJS, 180, 67
Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3
Rimoldini, L., Eyer, L., Audard, M., et al. 2022, Gaia DR3 documentation Chapter 10: Variability, Gaia DR3 documentation, European Space Agency; Gaia Data Processing and Analysis Consortium, https://gea.esac.esa.int/archive/documentation/GDR3/Data_analysis/chap_cu7var/
Rimoldini, L., Holl, B., Gavras, P., et al. 2023, A&A, 674, A14 (*Gaia* DR3 SI)
Ross, N. P., Myers, A. D., Sheldon, E. S., et al. 2012, ApJS, 199, 3
Schlafly, E. F., Meisner, A. M., & Green, G. M. 2019, ApJS, 240, 30
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schneider, D. P., Richards, G. T., Fan, X., et al. 2002, AJ, 123, 567
Secrest, N. J., Dudik, R. P., Dorland, B. N., et al. 2015, ApJS, 221, 12
Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, AJ, 134, 2236
Sesar, B., Hernitschek, N., Mitrović, S., et al. 2017, AJ, 153, 204
Shu, Y., Koposov, S. E., Evans, N. W., et al. 2019, MNRAS, 489, 4741
Simonetti, J. H., Cordes, J. M., & Heeschen, D. S. 1985, ApJ, 296, 46
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Souchay, J., Gattano, C., Andrei, A. H., et al. 2019, A&A, 624, A145
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30
Sumi, T., Woźniak, P. R., Eyer, L., et al. 2005, MNRAS, 356, 331
Taylor, M. B. 2005, ASP Conf. Ser., 347, 29
Tewes, M., Courbin, F., & Meylan, G. 2013, A&A, 553, A120
Urry, C. M., & Padovani, P. 1995, PASP, 107, 803
Vaughan, S., Edelson, R., Warwick, R. S., & Uttley, P. 2003, MNRAS, 345, 1271
Wong, K. C., Suyu, S. H., Chen, G. C. F., et al. 2020, MNRAS, 498, 1420
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Yan, L., Donoso, E., Tsai, C.-W., et al. 2013, AJ, 145, 55
Yao, S., Wu, X.-B., Ai, Y. L., et al. 2019, ApJS, 240, 6

Appendix A: Full acknowledgements

This work presents results from the European Space Agency (ESA) space mission *Gaia*. *Gaia* data are being processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the *Gaia* MultiLateral Agreement (MLA). The *Gaia* mission website is https://www.cosmos.esa.int/gaia. The *Gaia* archive website is https://archives.esac.esa.int/gaia.

The *Gaia* mission and data processing have financially been supported by, in alphabetical order by country:

- the Algerian Centre de Recherche en Astronomie, Astrophysique et Géophysique of Bouzareah Observatory;
- the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF) Hertha Firnberg Programme through grants T359, P20046, and P23737;
- the BELgian federal Science Policy Office (BELSPO) through various PROgramme de Développement d'Expériences scientifiques (PRODEX) grants and the Polish Academy of Sciences - Fonds Wetenschappelijk Onderzoek through grant VS.091.16N, and the Fonds de la Recherche Scientifique (FNRS), and the Research Council of Katholieke Universiteit (KU) Leuven through grant C16/18/005 (Pushing AsteRoseismology to the next level with TESS, GaiA, and the Sloan DIgital Sky SurvEy – PARADISE);
- the Brazil-France exchange programmes Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) - Comité Français d'Evaluation de la Coopération Universitaire et Scientifique avec le Brésil (COFECUB);
- the Chilean Agencia Nacional de Investigación y Desarrollo (ANID) through Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) Regular Project 1210992 (L. Chemin);
- the National Natural Science Foundation of China (NSFC) through grants 11573054, 11703065, and 12173069, the China Scholarship Council through grant 201806040200, and the Natural Science Foundation of Shanghai through grant 21ZR1474100;
- the Tenure Track Pilot Programme of the Croatian Science Foundation and the École Polytechnique Fédérale de Lausanne and the project TTP-2018-07-1171 'Mining the Variable Sky', with the funds of the Croatian-Swiss Research Programme;
- the Czech-Republic Ministry of Education, Youth, and Sports through grant LG 15010 and INTER-EXCELLENCE grant LTAUSA18093, and the Czech Space Office through ESA PECS contract 98058;
- the Danish Ministry of Science;
- the Estonian Ministry of Education and Research through grant IUT40-1;
- the European Commission's Sixth Framework Programme through the European Leadership in Space Astrometry (ELSA) Marie Curie Research Training Network (MRTN-CT-2006-033481), through Marie Curie project PIOF-GA-2009-255267 (Space AsteroSeismology & RR Lyrae stars, SAS-RRL), and through a Marie Curie Transfer-of-Knowledge (ToK) fellowship (MTKD-CT-2004-014188); the European Commission's Seventh Framework Programme through grant FP7-606740 (FP7-SPACE-2013-1) for the *Gaia* European Network for Improved data User Services (GENIUS) and through grant 264895 for the *Gaia* Research for European Astronomy Training (GREAT-ITN) network;
- the European Cooperation in Science and Technology (COST) through COST Action CA18104 'Revealing the Milky Way with *Gaia* (MW-Gaia)';
- the European Research Council (ERC) through grants 320360, 647208, and 834148 and through the European Union's Horizon 2020 research and innovation and excellent science programmes through Marie Skłodowska-Curie grant 745617 (Our Galaxy at full HD – Gal-HD) and 895174 (The build-up and fate of self-gravitating systems in the Universe) as well as grants 687378 (Small Bodies: Near and Far), 682115 (Using the Magellanic Clouds to Understand the Interaction of Galaxies), 695099 (A sub-percent distance scale from binaries and Cepheids – CepBin), 716155 (Structured ACCREtion Disks – SACCRED), 951549 (Sub-percent calibration of the extragalactic distance scale in the era of big surveys – UniverScale), and 101004214 (Innovative Scientific Data Exploration and Exploitation Applications for Space Sciences – EXPLORE);
- the European Science Foundation (ESF), in the framework of the *Gaia* Research for European Astronomy Training Research Network Programme (GREAT-ESF);
- the European Space Agency (ESA) in the framework of the *Gaia* project, through the Plan for European Cooperating States (PECS) programme through contracts C98090 and 4000106398/12/NL/KML for Hungary, through contract 4000115263/15/NL/IB for Germany, and through PROgramme de Développement d'Expériences scientifiques (PRODEX) grant 4000127986 for Slovenia;
- the Academy of Finland through grants 299543, 307157, 325805, 328654, 336546, and 345115 and the Magnus Ehrnrooth Foundation;
- the French Centre National d'Études Spatiales (CNES), the Agence Nationale de la Recherche (ANR) through grant ANR-10-IDEX-0001-02 for the 'Investissements d'avenir' programme, through grant ANR-15-CE31-0007 for project 'Modelling the Milky Way in the *Gaia* era' (MOD4Gaia), through grant ANR-14-CE33-0014-01 for project 'The Milky Way disc formation in the *Gaia* era' (ARCHEOGAL), through grant ANR-15-CE31-0012-01 for project 'Unlocking the potential of Cepheids as primary distance calibrators' (UnlockCepheids), through grant ANR-19-CE31-0017 for project 'Secular evolution of galxies' (SEGAL), and through grant ANR-18-CE31-0006 for project 'Galactic Dark Matter' (GaDaMa), the Centre National de la Recherche Scientifique (CNRS) and its SNO *Gaia* of the Institut des Sciences de l'Univers (INSU), its Programmes Nationaux: Cosmologie et Galaxies (PNCG), Gravitation Références Astronomie Métrologie (PNGRAM), Planétologie (PNP), Physique et Chimie du Milieu Interstellaire (PCMI), and Physique Stellaire (PNPS), the 'Action Fédératrice *Gaia*' of the Observatoire de Paris, the Région de Franche-Comté, the Institut National Polytechnique (INP) and the Institut National de Physique nucléaire et de Physique des Particules (IN2P3) co-funded by CNES;
- the German Aerospace Agency (Deutsches Zentrum für Luft- und Raumfahrt e.V., DLR) through grants 50QG0501, 50QG0601, 50QG0602, 50QG0701, 50QG0901, 50QG 1001, 50QG1101, 50QG1401, 50QG1402, 50QG1403, 50QG1404, 50QG1904, 50QG2101, 50QG2102, and 50QG2202, and the Centre for Information Services and High Performance Computing (ZIH) at the Technische

Universität Dresden for generous allocations of computer time;
– the Hungarian Academy of Sciences through the Lendület Programme grants LP2014-17 and LP2018-7 and the Hungarian National Research, Development, and Innovation Office (NKFIH) through grant KKP-137523 ('SeismoLab');
– the Science Foundation Ireland (SFI) through a Royal Society - SFI University Research Fellowship (M. Fraser);
– the Israel Ministry of Science and Technology through grant 3-18143 and the Tel Aviv University Center for Artificial Intelligence and Data Science (TAD) through a grant;
– the Agenzia Spaziale Italiana (ASI) through contracts I/037/08/0, I/058/10/0, 2014-025-R.0, 2014-025-R.1.2015, and 2018-24-HH.0 to the Italian Istituto Nazionale di Astrofisica (INAF), contract 2014-049-R.0/1/2 to INAF for the Space Science Data Centre (SSDC, formerly known as the ASI Science Data Center, ASDC), contracts I/008/10/0, 2013/030/I.0, 2013-030-I.0.1-2015, and 2016-17-I.0 to the Aerospace Logistics Technology Engineering Company (ALTEC S.p.A.), INAF, and the Italian Ministry of Education, University, and Research (Ministero dell'Istruzione, dell'Università e della Ricerca) through the Premiale project 'MIning The Cosmos Big Data and Innovative Italian Technology for Frontier Astrophysics and Cosmology' (MITiC);
– the Netherlands Organisation for Scientific Research (NWO) through grant NWO-M-614.061.414, through a VICI grant (A. Helmi), and through a Spinoza prize (A. Helmi), and the Netherlands Research School for Astronomy (NOVA);
– the Polish National Science Centre through HARMONIA grant 2018/30/M/ST9/00311 and DAINA grant 2017/27/L/ST9/03221 and the Ministry of Science and Higher Education (MNiSW) through grant DIR/WK/2018/12;
– the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through national funds, grants SFRH/BD/128840/2017 and PTDC/FIS-AST/30389/2017, and work contract DL 57/2016/CP1364/CT0006, the Fundo Europeu de Desenvolvimento Regional (FEDER) through grant POCI-01-0145-FEDER-030389 and its Programa Operacional Competitividade e Internacionalização (COMPETE2020) through grants UIDB/04434/2020 and UIDP/04434/2020, and the Strategic Programme UIDB/00099/2020 for the Centro de Astrofísica e Gravitação (CENTRA);
– the Slovenian Research Agency through grant P1-0188;
– the Spanish Ministry of Economy (MINECO/FEDER, UE), the Spanish Ministry of Science and Innovation (MICIN), the Spanish Ministry of Education, Culture, and Sports, and the Spanish Government through grants BES-2016-078499, BES-2017-083126, BES-C-2017-0085, ESP2016-80079-C2-1-R, ESP2016-80079-C2-2-R, FPU16/03827, PDC2021-121059-C22, RTI2018-095076-B-C22, and TIN2015-65316-P ('Computación de Altas Prestaciones VII'), the Juan de la Cierva Incorporación Programme (FJCI-2015-2671 and IJC2019-04862-I for F. Anders), the Severo Ochoa Centre of Excellence Programme (SEV2015-0493), and MICIN/AEI/10.13039/501100011033 (and the European Union through European Regional Development Fund 'A way of making Europe') through grant RTI2018-095076-B-C21, the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia 'María de Maeztu') through grant CEX2019-000918-M, the University of Barcelona's official doctoral programme for the development of an R+D+i project through an Ajuts de Personal Investigador en Formació (APIF) grant, the Spanish Virtual Observatory through project AyA2017-84089, the Galician Regional Government, Xunta de Galicia, through grants ED431B-2021/36, ED481A-2019/155, and ED481A-2021/296, the Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), funded by the Xunta de Galicia and the European Union (European Regional Development Fund – Galicia 2014-2020 Programme), through grant ED431G-2019/01, the Red Española de Supercomputación (RES) computer resources at MareNostrum, the Barcelona Supercomputing Centre - Centro Nacional de Supercomputación (BSC-CNS) through activities AECT-2017-2-0002, AECT-2017-3-0006, AECT-2018-1-0017, AECT-2018-2-0013, AECT-2018-3-0011, AECT-2019-1-0010, AECT-2019-2-0014, AECT-2019-3-0003, AECT-2020-1-0004, and DATA-2020-1-0010, the Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya through grant 2014-SGR-1051 for project 'Models de Programació i Entorns d'Execució Parallels' (MPEXPAR), and Ramon y Cajal Fellowship RYC2018-025968-I funded by MICIN/AEI/10.13039/501100011033 and the European Science Foundation ('Investing in your future');
– the Swedish National Space Agency (SNSA/Rymdstyrelsen);
– the Swiss State Secretariat for Education, Research, and Innovation through the Swiss Activités Nationales Complémentaires and the Swiss National Science Foundation through an Eccellenza Professorial Fellowship (award PCEFP2_194638 for R. Anderson);
– the United Kingdom Particle Physics and Astronomy Research Council (PPARC), the United Kingdom Science and Technology Facilities Council (STFC), and the United Kingdom Space Agency (UKSA) through the following grants to the University of Bristol, the University of Cambridge, the University of Edinburgh, the University of Leicester, the Mullard Space Sciences Laboratory of University College London, and the United Kingdom Rutherford Appleton Laboratory (RAL): PP/D006511/1, PP/D006546/1, PP/D006570/1, ST/I000852/1, ST/J005045/1, ST/K00056X/1, ST/K000209/1, ST/K000756/1, ST/L006561/1, ST/N000595/1, ST/N000641/1, ST/N000978/1, ST/N001117/1, ST/S000089/1, ST/S000976/1, ST/S000984/1, ST/S001123/1, ST/S001948/1, ST/S001980/1, ST/S002103/1, ST/V000969/1, ST/W002469/1, ST/W002493/1, ST/W002671/1, ST/W002809/1, and EP/V520342/1.

The GBOT programme uses observations collected at (i) the European Organisation for Astronomical Research in the Southern Hemisphere (ESO) with the VLT Survey Telescope (VST), under ESO programmes 092.B-0165, 093.B-0236, 094.B-0181, 095.B-0046, 096.B-0162, 097.B-0304, 098.B-0030, 099.B-0034, 0100.B-0131, 0101.B-0156, 0102.B-0174, and 0103.B-0165; and (ii) the Liverpool Telescope, which is operated on the island of La Palma by Liverpool John Moores University in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias with financial support from the United Kingdom Science and Technology Facilities Council, and (iii) telescopes of the Las Cumbres Observatory Global Telescope Network.

This work made use of software from Postgres-XL (https://www.postgres-xl.org), Java (https://www.oracle.com/java/), and TOPCAT/STILTS (Taylor et al. 2005). This research has made use of NASA's Astrophysics Data System. This research has made use of the NASA/IPAC Extragalactic

Database, which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration, including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, HarvardSmithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.