Astronomy & Astrophysics manuscript no. gdr3_extragalactic
Tuesday 14th June, 2022

Gaia Data Release 3: The extragalactic content⋆

Gaia Collaboration, C.A.L. Bailer-Jones[1], D. Teyssier[2], L. Delchambre[3], C. Ducourant[4], D. Garabato[5], D. Hatzidimitriou[6,7], S.A. Klioner[8], L. Rimoldini[9], I. Bellas-Velidis[7], R. Carballo[10], M.I. Carnerero[11], C. Diener[12], M. Fouesneau[1], L. Galluccio[13], P. Gavras[14], A. Krone-Martins[15,16], C.M. Raiteri[11], R. Teixeira[17], A.G.A. Brown[18], A. Vallenari[19], T. Prusti[20], J.H.J. de Bruijne[20], F. Arenou[21], C. Babusiaux[22,21], M. Biermann[23], O.L. Creevey[13], D.W. Evans[12], L. Eyer[24], R. Guerra[25], A. Hutton[26], C. Jordi[27], U.L. Lammers[25], L. Lindegren[28], X. Luri[27], F. Mignard[13], C. Panem[29], D. Pourbaix†[30,31], S. Randich[32], P. Sartoretti[21], C. Soubiran[4], P. Tanga[13], N.A. Walton[12], U. Bastian[23], R. Drimmel[11], F. Jansen[33], D. Katz[21], M.G. Lattanzi[11,34], F. van Leeuwen[12], J. Bakker[25], C. Cacciari[35], J. Castañeda[36], F. De Angeli[12], C. Fabricius[27], Y. Frémat[37], A. Guerrier[29], U. Heiter[38], E. Masana[27], R. Messineo[39], N. Mowlavi[24], C. Nicolas[29], K. Nienartowicz[40,9], F. Pailler[29], P. Panuzzo[21], F. Riclet[29], W. Roux[29], G.M. Seabroke[41], R. Sordo[19], F. Thévenin[13], G. Gracia-Abril[42,23], J. Portell[27], M. Altmann[23,43], R. Andrae[1], M. Audard[24,9], K. Benson[41], J. Berthier[44], R. Blomme[37], P.W. Burgess[12], D. Busonero[11], G. Busso[12], H. Cánovas[2], B. Carry[13], A. Cellino[11], N. Cheek[45], G. Clementini[35], Y. Damerdji[3,46], M. Davidson[47], P. de Teodoro[25], M. Nuñez Campos[26], A. Dell'Oro[32], P. Esquej[14], J. Fernández-Hernández[48], E. Fraile[14], P. García-Lario[25], E. Gosset[3,31], R. Haigron[21], J.-L. Halbwachs[49], N.C. Hambly[47], D.L. Harrison[12,50], J. Hernández[25], D. Hestroffer[44], S.T. Hodgkin[12], B. Holl[24,9], K. Janßen[51], G. Jevardat de Fombelle[24], S. Jordan[23], A.C. Lanzafame[52,53], W. Löffler[23], O. Marchal[49], P.M. Marrese[54,55], A. Moitinho[15], K. Muinonen[56,57], P. Osborne[12], E. Pancino[32,55], T. Pauwels[37], A. Recio-Blanco[13], C. Reylé[58], M. Riello[12], T. Roegiers[59], J. Rybizki[1], L.M. Sarro[60], C. Siopis[30], M. Smith[41], A. Sozzetti[11], E. Utrilla[26], M. van Leeuwen[12], U. Abbas[11], P. Ábrahám[61,62], A. Abreu Aramburu[48], C. Aerts[63,64,1], J.J. Aguado[60], M. Ajaj[21], F. Aldea-Montero[25], G. Altavilla[54,55], M.A. Álvarez[5], J. Alves[65], R.I. Anderson[66], E. Anglada Varela[48], T. Antoja[27], D. Baines[2], S.G. Baker[41], L. Balaguer-Núñez[27], E. Balbinot[67], Z. Balog[23,1], C. Barache[43], D. Barbato[24,11], M. Barros[15], M.A. Barstow[68], S. Bartolomé[27], J.-L. Bassilana[69], N. Bauchet[21], U. Becciani[52], M. Bellazzini[35], A. Berihuete[70], M. Bernet[27], S. Bertone[71,72,11], L. Bianchi[73], A. Binnenfeld[74], S. Blanco-Cuaresma[75], T. Boch[49], A. Bombrun[76], D. Bossini[77], S. Bouquillon[43,78], A. Bragaglia[35], L. Bramante[39], E. Breedt[12], A. Bressan[79], N. Brouillet[4], E. Brugaletta[52], B. Bucciarelli[11,34], A. Burlacu[80], A.G. Butkevich[11], R. Buzzi[11], E. Caffau[21], R. Cancelliere[81], T. Cantat-Gaudin[27,1], T. Carlucci[43], J.M. Carrasco[27], L. Casamiquela[4,21], M. Castellani[54], A. Castro-Ginard[18], L. Chaoul[29], P. Charlot[4], L. Chemin[82], V. Chiaramida[39], A. Chiavassa[13], N. Chornay[12], G. Comoretto[2,83], G. Contursi[13], W.J. Cooper[84,11], T. Cornez[69], S. Cowell[12], F. Crifo[21], M. Cropper[41], M. Crosta[11,85], C. Crowley[76], C. Dafonte[5], A. Dapergolas[7], P. David[44], P. de Laverny[13], F. De Luise[86], R. De March[39], J. De Ridder[63], R. de Souza[17], A. de Torres[76], E.F. del Peloso[23], E. del Pozo[26], M. Delbo[13], A. Delgado[14], J.-B. Delisle[24], C. Demouchy[87], T.E. Dharmawardena[1], S. Diakite[88], E. Distefano[52], C. Dolding[41], H. Enke[51], C. Fabre[89], M. Fabrizio[54,55], S. Faigler[90], G. Fedorets[56,91], P. Fernique[49,92], F. Figueras[27], Y. Fournier[51], C. Fouron[80], F. Fragkoudi[93,94,95], M. Gai[11], A. Garcia-Gutierrez[27], M. Garcia-Reinaldos[25], M. García-Torres[96], A. Garofalo[35], A. Gavel[38], E. Gerlach[8], R. Geyer[8], P. Giacobbe[11], G. Gilmore[12], S. Girona[97], G. Giuffrida[54], R. Gomel[90], A. Gomez[5], J. González-Núñez[45,98], I. González-Santamaría[5], J.J. González-Vidal[27], M. Granvik[56,99], P. Guillout[49], J. Guiraud[29], R. Gutiérrez-Sánchez[2], L.P. Guy[9,100], M. Hauser[1,101], M. Haywood[21], A. Helmer[69], A. Helmi[67], M.H. Sarmiento[26], S.L. Hidalgo[102,103], T. Hilger[7], N. Hładczuk[25,104], D. Hobbs[28], G. Holland[12], H.E. Huckle[41], K. Jardine[105], G. Jasniewicz[106], A. Jean-Antoine Piccolo[29], Ó. Jiménez-Arranz[27], J. Juaristi Campillo[23], F. Julbe[27], L. Karbevska[9,107], P. Kervella[108], S. Khanna[67,11], M. Kontizas[6], G. Kordopatis[13], A.J. Korn[38], Á Kóspál[61,1,62], Z. Kostrzewa-Rutkowska[18,109], K. Kruszyńska[110], M. Kun[61], P. Laizeau[111], S. Lambert[43], A.F. Lanza[52], Y. Lasne[69], J.-F. Le Campion[4], Y. Lebreton[108,112], T. Lebzelter[65], S. Leccia[113], N. Leclerc[21], I. Lecoeur-Taibi[9], S. Liao[114,11,115], E.L. Licata[11], H.E.P. Lindstrøm[11,116,117], T.A. Lister[118], E. Livanou[6], A. Lobel[37], A. Lorca[26], C. Loup[49], P. Madrero Pardo[27], A. Magdaleno Romeo[80], S. Managau[69], R.G. Mann[47], M. Manteiga[119], J.M. Marchant[120], M. Marconi[113], J. Marcos[2], M.M.S. Marcos Santos[45], D. Marín Pina[27], S. Marinoni[54,55], F. Marocco[121],

D.J. Marshall[122], L. Martin Polo[45], J.M. Martín-Fleitas[26], G. Marton[61], N. Mary[69], A. Masip[27], D. Massari[35], A. Mastrobuono-Battisti[21], T. Mazeh[90], P.J. McMillan[28], S. Messina[52], D. Michalik[20], N.R. Millar[12], A. Mints[51], D. Molina[27], R. Molinaro[113], L. Molnár[61,123,62], G. Monari[49], M. Monguió[27], P. Montegriffo[35], A. Montero[26], R. Mor[27], A. Mora[26], R. Morbidelli[11], T. Morel[3], D. Morris[47], T. Muraveva[35], C.P. Murphy[25], I. Musella[113], Z. Nagy[61], L. Noval[69], F. Ocaña[2,124], A. Ogden[12], C. Ordenovic[13], J.O. Osinde[14], C. Pagani[68], I. Pagano[52], L. Palaversa[125,12], P.A. Palicio[13], L. Pallas-Quintela[5], A. Panahi[90], S. Payne-Wardenaar[23], X. Peñalosa Esteller[27], A. Penttilä[56], B. Pichon[13], A.M. Piersimoni[86], F.-X. Pineau[49], E. Plachy[61,123,62], G. Plum[21], E. Poggio[13,11], A. Prša[126], L. Pulone[54], E. Racero[45,124], S. Ragaini[35], M. Rainer[32,127], P. Ramos[27,49], M. Ramos-Lerate[2], P. Re Fiorentin[11], S. Regibo[63], P.J. Richards[128], C. Rios Diaz[14], V. Ripepi[113], A. Riva[11], H.-W. Rix[1], G. Rixon[12], N. Robichon[21], A.C. Robin[58], C. Robin[69], M. Roelens[24], H.R.O. Rogues[87], L. Rohrbasser[9], M. Romero-Gómez[27], N. Rowell[47], F. Royer[21], D. Ruz Mieres[12], K.A. Rybicki[110], G. Sadowski[30], A. Sáez Núñez[27], A. Sagristà Sellés[23], J. Sahlmann[14], E. Salguero[48], N. Samaras[37,129], V. Sanchez Gimenez[27], N. Sanna[32], R. Santoveña[5], M. Sarasso[11], M. Schultheis[13], E. Sciacca[52], M. Segol[87], J.C. Segovia[45], D. Ségransan[24], D. Semeux[89], S. Shahaf[130], H.I. Siddiqui[131], A. Siebert[49,92], L. Siltala[56], A. Silvelo[5], E. Slezak[13], I. Slezak[13], R.L. Smart[11], O.N. Snaith[21], E. Solano[132], F. Solitro[39], D. Souami[108,133], J. Souchay[43], A. Spagna[11], L. Spina[19], F. Spoto[75], I.A. Steele[120], H. Steidelmüller[8], C.A. Stephenson[2,134], M. Süveges[135], J. Surdej[3,136], L. Szabados[61], E. Szegedi-Elek[61], F. Taris[43], M.B. Taylor[137], L. Tolomei[39], N. Tonello[97], F. Torra[36], J. Torra†[27], G. Torralba Elipe[5], M. Trabucchi[138,24], A.T. Tsounis[139], C. Turon[21], A. Ulla[140], N. Unger[24], M.V. Vaillant[69], E. van Dillen[87], W. van Reeven[141], O. Vanel[21], A. Vecchiato[11], Y. Viala[21], D. Vicente[97], S. Voutsinas[47], M. Weiler[27], T. Wevers[12,142], Ł. Wyrzykowski[110], A. Yoldas[12], P. Yvard[87], H. Zhao[13], J. Zorec[143], S. Zucker[74], and T. Zwitter[144]

(Affiliations can be found after the references)

First submitted 31 January 2022. Resubmitted 27 April 2022. Accepted 27 April 2022

ABSTRACT

The *Gaia* Galactic survey mission is designed and optimized to obtain astrometry, photometry, and spectroscopy of nearly two billion stars in our Galaxy. Yet as an all-sky multi-epoch survey, *Gaia* also observes several million extragalactic objects down to a magnitude of $G \sim 21$ mag. Due to the nature of the *Gaia* onboard-selection algorithms, these are mostly point-source-like objects. Using data provided by the satellite, we have identified quasar and galaxy candidates via supervised machine learning methods, and estimate their redshifts using the low resolution BP/RP spectra. We further characterise the surface brightness profiles of host galaxies of quasars and of galaxies from pre-defined input lists. Here we give an overview of the processing of extragalactic objects, describe the data products in *Gaia* DR3, and analyse their properties. Two integrated tables contain the main results for a high completeness, but low purity (50–70%), set of 6.6 million candidate quasars and 4.8 million candidate galaxies. We provide queries that select purer sub-samples of these containing 1.9 million probable quasars and 2.9 million probable galaxies (both ~95% purity). We also use high quality BP/RP spectra of 43 thousand high probability quasars over the redshift range 0.05–4.36 to construct a composite quasar spectrum spanning restframe wavelengths from 72–1000 nm.

1. Introduction

The primary objective of the *Gaia* mission is to study the structure and origin of our Galaxy by measuring the distribution, kinematics, and physical properties of its constituent stars (Gaia Collaboration et al. 2016). The satellite and its observing strategy were therefore designed to optimize the measurement of astrometry, photometry, and spectroscopy of point sources. Nonetheless, by observing the entire sky multiple times down to a limiting magnitude of $G \simeq 21$ mag, *Gaia* has observed millions of extragalactic objects since it started observing in mid 2014. Various data on many of these objects are provided as part of the third *Gaia* data release (DR3), covering both previously-identified objects and new candidate objects identified using the *Gaia* data. The purpose of this paper is to summarize how extragalactic objects were identified, what their properties are, and what data on them are provided in *Gaia* DR3.

Extragalactic objects are classified or analysed by several modules in the *Gaia* data processing system. These modules were provided by different coordination units (CUs) within the Data Processing and Analysis Consortium (DPAC) and operate largely independently. They are as follows: CU3 Astrometry, which assembled a list of extragalactic point sources from external catalogues to use in defining the astrometric reference frame (Gaia Collaboration & Klioner et al. 2022); CU4 Extended Objects (EO), which analyses the surface brightness profiles of an input list of objects to look for physical extension; CU7 Variability, which uses photometric light curves to characterise variability; CU8 Astrophysical Parameters, which uses astrometry, photometry, and the BP/RP spectra to classify objects and to estimate redshifts. Whereas the modules from CU3 and CU4 work on a predefined lists of extragalactic objects identified in other surveys, the Vari module in CU7 and the Discrete Source Classifier (DSC) module in CU8 use supervised machine learning to discover new objects. These classifiers use only *Gaia* data. The inclusion of additional data, such as infrared photometry, should improve the classification performance (sample completeness and purity). However, a key principle of the DPAC is to provide homogeneous classifications based only on the *Gaia* data, unaffected by issues with other catalogues, such as incompleteness.

It is important to realise that there is no common definition of quasar or galaxy across the various *Gaia* modules. A common

⋆ Table 8 is only available in electronic form at the CDS at http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/

definition is also not possible, because each module uses different data to classify or select objects, including different training sets. But broadly speaking, the term 'extragalactic' in the context of this paper refers to unresolved or barely resolved individual objects more than 50 Mpc from the Sun.

If *Gaia* were to obtain noise-free, unbiased parallaxes, then identifying extragalactic objects would be simple: They would be all the objects with parallaxes below some threshold. Yet we do not have this luxury: Despite the high precision of *Gaia* DR3 parallaxes – around 0.5 mas at $G = 20$ mag and 0.25 mas at $G = 19$ mag (Lindegren et al. 2021b) – this is not nearly enough to reliably identify extragalactic objects through a simple cut on parallaxes (or proper motions). Indeed, 657 million objects in *Gaia* DR3 have raw parallaxes below 0.25 mas, the vast majority of which are of course stars in our Galaxy. This is not to say that parallaxes, and moreover proper motions, are not useful, however, and we do indeed make use of them in our classifications and analyses.

Most of the extragalactic candidates we have identified are bundled into two integrated tables in *Gaia* DR3, called `qso_candidates` and `galaxy_candidates`. As their names make clear, the construction of these tables has been driven primarily by the desire to be complete, rather than pure. Together these tables contain around 11.3 million unique objects and have global purities of 50–70%, although they are significantly higher when we exclude the Galactic plane, high density regions around clusters and galaxies, and the faintest sources. These tables are nonetheless a significant improvement over the `gaia_source` table, which has 1.8 billion objects and an extragalactic purity of around 0.2%. Our rationale for producing completeness-driven integrated tables is that it is easier for users to then select a sub-sample of purer objects (according to their own criteria) from our integrated tables, than it would be to find objects (in `gaia_source`) that had been removed from purity-driven tables. In Sect. 8 we recommend how to extract a purer sub-sample (∼96%) from the two integrated tables.

This paper is not the first to deal with classifying extragalactic objects using *Gaia* data. Initial studies cross-matched *Gaia* positions to other catalogues to analyse the properties of quasars and galaxies (e.g. Paine et al. 2018; Souchay et al. 2019). Several studies have made cuts on the astrometry (e.g. Heintz et al. 2018; Gaia Collaboration 2018), sometimes combined with classification using non-*Gaia* data (e.g. Fu et al. 2021), and others have applied machine learning methods to a number of Gaia metrics (e.g. Bailer-Jones et al. 2019) to identify extragalactic objects. Purer samples should be attainable when combining *Gaia* data with more discriminatory data, albeit at the loss of completeness if *Gaia* is the larger survey, and some studies report good results here (e.g. Wu et al. 2021). Other studies have used the Gaia data to characterise specific types of extragalactic object, such as gravitational lenses (e.g. Krone-Martins et al. 2018; Delchambre et al. 2019).

This paper is structured as follows. Section 2 summarizes the extragalactic processing modules that deliver results in *Gaia* DR3 and Sect. 3 describes the various tables that provide these results. Section 4 presents the properties of the extragalactic objects, such as sky distributions, spectra, surface brightnesses, and light curves. In Sect. 5 we provide some basic comparisons between the results of the different modules, and in Sect. 6 we compare the results to external surveys. In Sect. 7 we compute composite quasar spectra from individual quasar spectra at a range of redshifts. Section 8 describes a purer, and necessarily less complete, sub-sample of the integrated extragalactic tables. We conclude in Sect. 9 with some suggested use cases.

Many more details on the topics discussed here can be found in the extensive online documentation that accompanies this data release.[1] We point in particular to the table and field descriptions there. Several other release papers provide details that are not in the documentation. These are Delchambre et al. (2022) for the CU8 classification and redshift estimation modules, Rimoldini et al. (2022) for the CU7 variability classifier and Carnerero et al. (2022) for the resulting selection of Active Galactic Nuclei (AGN), Ducourant et al. (2022) for the CU4 surface brightness profile analysis, and Gaia Collaboration & Klioner et al. (2022) for the *Gaia*-CRF3 (Celestial Reference Frame 3). Readers may also want to consult De Angeli et al. (2022) for a description of the BP/RP spectrophotometry and Lindegren et al. (2021b) for the astrometric (parallax and proper motion) processing (the latter unchanged from *Gaia* EDR3).

2. Extragalactic processing modules

The modules in the *Gaia* data processing system that deal explicitly with extragalactic objects are as follows. DSC and Vari classify *Gaia* objects using supervised machine learning methods. Vari additionally provides characterisations of the light curves. UGC (Unresolved Galaxy Classifier), QSOC (Quasar Classifier), and OA (Outlier Analyser) analyse the results from DSC, the first two computing redshifts. EO analyses the surface brightness profiles of an input source list. We summarize these modules here, leaving more detailed descriptions to the individual processing papers cited below. We also include in our analysis the list of quasars identified for *Gaia*-CRF3.

Some sources, in particular galaxies, are partially resolved by *Gaia*. Their two-dimensional structure – combined with the fact that *Gaia* observes sources over a range of position angles – can induce a spurious (non-intrinsic) photometric variability or an apparent astrometric variability, the latter potentially being interpreted by the astrometric processing (Lindegren et al. 2021b) as spuriously large parallaxes and proper motions. The DSC and Vari modules take advantage of these spurious measurements to help them classify extragalactic sources.

2.1. Discrete Source Classifier (CU8-DSC)

The Discrete Source Classifier uses the BP/RP spectrum together with the mean G-band magnitude, the variability in this band, the parallax, and the proper motion to classify each *Gaia* source probabilistically into five classes: quasar; galaxy; anonymous (essentially single star); white dwarf; binary star. DSC is trained empirically on *Gaia* data with labels for the quasar and galaxy classes coming from Sloan Digital Sky Survey (SDSS) spectroscopic classifications. The distributions of the training data in colour and magnitude are shown in Fig. 1. The training data define the classes (see Sect. 6.2), so these are not the same class definition adopted by other modules that contribute extragalactic source identifications to *Gaia* DR3. DSC comprises three classifiers. Specmod uses the BP/RP spectrum only and gives results for all five classes in DSC. Allosmod uses various photometric and astrometric features and only gives results for quasars, galaxies, and single stars. Specmod and Allosmod are nonetheless trained on a common set of data that has complete data for both classifiers. One consequence of this is that Specmod is also applied to some types of sources it was not trained on, for example galaxies that lack measured parallaxes and proper motions. Combmod combines the Specmod and Allosmod classification

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

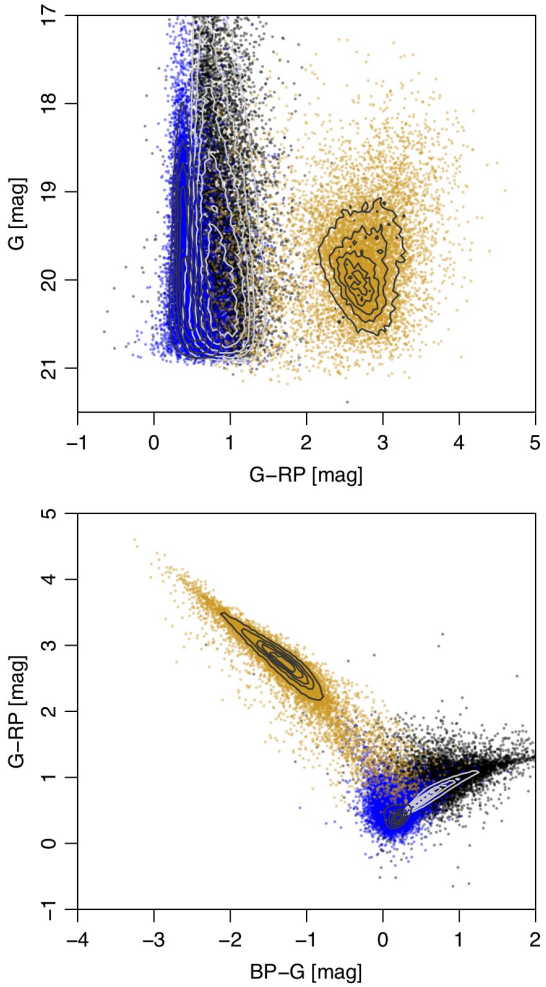

Fig. 1. Colour–magnitude diagram (top) and colour–colour diagram (bottom) of the DSC training data for the quasars (blue) and galaxies (orange) as well as stars (black). The contours in each panel show the variation in source density on a linear scale. The points are equal-sized random subsets of sources from each class. There is significant overlap, in particular between stars and quasars: in reality the former dominate by a factor of about a thousand, and so overlap much more than is shown here. Plots for each class separately are provided in the online documentation.

probabilities in a Bayesian manner to give probabilities for all five classes (using the algorithm described in the appendix of Delchambre et al. 2022). Probabilities from all three classifiers are provided in the `astrophysical_parameters` table. DSC is described in more detail in Delchambre et al. (2022) and in the online documentation.

DSC incorporates a global class prior that reflects the rareness of quasars and galaxies. This makes it hard to achieve a high purity even for a good classifier. For example, if only one in every thousand sources were extragalactic, then even if a classifier had a 99.9% accuracy, the resulting sample would only be around 50% pure. For this reason one must report results not on a balanced validation set, but on one that reflects this prior.[2]

In addition to posterior probabilities, DSC also provides two class labels. The first, `classlabel_dsc`, is assigned the name of the class that achieves the highest posterior probability in Combmod that is greater than 0.5. If none of the output probabilities are above 0.5 then this class label is `unclassified`. This tends to produce a complete but impure sample of objects when we properly account for extragalactic rareness. The analyses in Delchambre et al. (2022) and Bailer-Jones (2021) using SDSS spectroscopically-confirmed objects shows a completeness for quasars and galaxies objects of over 90%, but a global purity of only about 20–25%. For Galactic latitudes above 11.5° the purities increase to 41%. Additional filtering increases this further (see Sect. 8). The second class label, `classlabel_dsc_joint` defines a purer set of quasars and galaxies, and is assigned by requiring both Specmod and Allosmod probabilities to be above 0.5 for the corresponding class. This gives completenesses of 38% on quasars and 83% on galaxies, and purities on both classes of 63%. For Galactic latitudes above 11.5° the purities increase to about 80%.

2.2. Quasar Classifier (CU8-QSOC)

The Quasar Classifier module (QSOC) estimates the redshift of sources classified as quasars by DSC-Combmod using their BP/RP spectra. For this selection, QSOC uses a very loose cut on the DSC quasar probability, `classprob_dsc_combmod_quasar` $\geq$ 0.01. This prioritizes completeness at the expense of purity to ensure that most of the objects that are suspected to be quasars are given a redshift estimate. The QSOC redshifts are inferred with a chi-square approach, whereby the BP and RP spectra are compared to a composite quasar spectrum taken at various trial redshifts in the range $0 \lesssim z \lesssim 6$. The composite spectrum is built upon a semi-empirical library of quasars from the SDSS DR12Q sample (Pâris et al. 2017). Each SDSS spectrum is first extrapolated to the wavelength range covered by BP/RP before being converted into a BP/RP spectrum using the available instrument model. More details of the algorithm can be found in Delchambre et al. (2022). In addition to the best point estimate of the redshift, QSOC also estimates lower and upper confidence intervals, `redshift_qsoc_lower` and `redshift_qsoc_upper`, which are the 15.9% and 84.1% quantiles of a log-normal distribution. The module also sets various processing flags in `flags_qsoc`, reflecting potential issues and/or degeneracies that may occur during the prediction phase.

2.3. Unresolved Galaxy Classifier (CU8-UGC)

The Unresolved Galaxy Classifier (UGC) estimates the redshift of sources classified as galaxies by DSC-Combmod with probability `classprob_dsc_combmod_galaxy` $\geq$ 0.25. UGC uses the BP/RP spectrum together with a supervised machine learning algorithm, the Support Vector Machine (SVM) (Cortes & Vapnik 1995; Chang & Lin 2011). A regression model (t-SVM) is trained on a set of 6000 sources selected from galaxies in the SDSS DR16 archive (Ahumada et al. 2020; Blanton et al. 2017) that are cross-matched to sources observed by *Gaia*. The BP/RP spectra and the SDSS redshifts of the sources in this set are used as training input and output, respectively. The SDSS galaxies were selected to have redshifts in the range $0 \leq z \leq 0.6$ and magnitudes $17 \leq G \leq 21$. Additional conditions were applied to specific parameters that influence the quality of the observed spectra. A test set of 250 000 galaxies, selected in a similar manner as the training set, was used to estimate the performance of the model, as reported in Delchambre et al. (2022).

[2] In practice we can use a validation set with more convenient class fractions, and then adjust the confusion matrix to reflect the prior, as explained in Sect. 3.4 of Bailer-Jones et al. (2019).

This set was also used to estimate statistical uncertainties of the redshift predictions in redshift bins of width 0.02. The bias – the mean difference between predicted and observed redshifts – was found to be −0.006 with a root mean squared error of 0.039 for the entire redshift range $0.0 \leq z \leq 0.6$. However, the uneven distribution of redshift and magnitude causes the performance to be better for lower redshifts than for higher ones. For each estimated `redshift_ugc` we further determine the lower and upper prediction level, `redshift_ugc_lower` and `redshift_ugc_upper`, corresponding to the bias and the 1σ error of the SVM model in the closest bin. See the online documentation for more details.

2.4. Outlier Analysis (CU8-OA)

The Outlier Analysis (OA) module was originally intended to analyse those sources that receive low classification probabilities for all DSC classes. As DSC-Combmod tends to give rather extreme probabilities – near to 0.0 or 1.0 – we used OA to analyse all sources that have all DSC-Combmod probabilities less than 0.999. This corresponds to 56 million sources. OA uses a Self-Organizing Map (Kohonen 1982), an unsupervised neural network that groups together similar data on a two-dimensional grid of neurons, in our case 30×30. The data here are the BP/RP spectra. From this we compute a prototype spectrum of each neuron as the mean of all spectra assigned to that neuron. We further compute various statistics for each neuron, such as the mean G, $G_{\rm BP}$, $G_{\rm RP}$, parallax, and Galactic latitude. We also compute a quality index that is based on the intra-neuron distance distribution; it takes seven discrete values from 0 to 6, where 0 represents the best quality neurons and 6 the poorest ones. The method of allocation to these is described in the online documentation. Finally, we compute a class label for each neuron by finding the best match between its prototype and a series of labelled templates, although neurons with quality index 6 are not assigned a label. This information is given in the `oa_neuron_information` and `oa_neuron_xp_spectra` tables, and an interactive visualization tool that can explore these tables is available (Álvarez et al. 2021).

2.5. Variability (CU7)

Extragalactic objects can also be identified via their photometric variability. Galaxies with active nuclei show variability in their accretion, such as in Seyfert galaxies and quasars, and in the case of blazars variability can be intrinsic or geometrical, related to a relativistic plasma jet directed towards us.

Using a supervised classification method Vari-Classification described in Rimoldini et al. (2022), we identified 1.0 million Active Galactic Nuclei (AGN) and 2.5 million galaxy candidates from the variability of the *Gaia* light curves. Epoch photometry in the G, $G_{\rm BP}$, and $G_{\rm RP}$ bands are published for AGN candidates, and for those galaxies that are part of the *Gaia* Andromeda Photometric Survey (GAPS; Evans et al. 2022) or that might be misclassified as real variables in *Gaia* DR3 (and so published in one of the variability tables). Indeed, the apparent variability of galaxies in the *Gaia* data is mostly an artefact of their extension combined with the Gaia on-board detection algorithm and scanning law (see Sect. 4.4.2) and so does not justify the release of their time series (which are meant only for genuine variable objects). Nevertheless, the characteristics of these artificial brightness variations made it possible to identify galaxies as if they were variable objects. Light curve statistics for all sources with light curves are published in the `vari_summary` table.

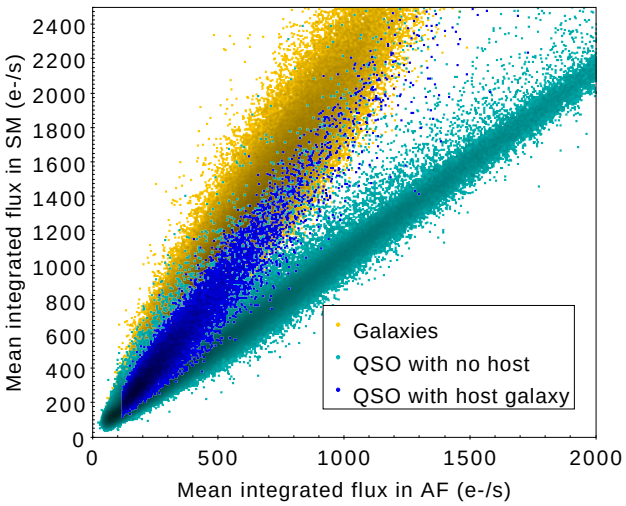

Fig. 2. Comparison of the flux collected in the AF and SM windows in the *Gaia* focal plane for quasars (with and without a detected host galaxy in *Gaia*) and for galaxies. Objects with an extension detectable by *Gaia* lie above the turquoise diagonal of quasars with no host galaxy.

Further analysis and characterisation of the variable AGN classifications (by the module Vari-AGN) led to a higher purity selection of about 872 000 objects (Carnerero et al. 2022), whose AGN-specific metrics are published in the `vari_agn` table, and repeated in the `qso_candidates` table (see Sect. 3). The purity of this AGN sample was estimated to be about 95%. The galaxy sample in `galaxy_candidates` is perhaps even purer, estimated at 99%, although with a lower completeness at around 40% (Rimoldini et al. 2022).

2.6. Surface brightness profile (CU4)

A source is recorded by *Gaia* only if the on-board video processing unit determines its light profile to be sufficiently steep at its centre (Gaia Collaboration et al. 2016). While this is intended to accept only point sources, it does pick up some extended objects (see section 2). The resulting selection function has been assessed theoretically by de Bruijne et al. (2015) and de Souza et al. (2014).

The CU4 surface brightness profile module attempts to reconstruct the two-dimensional light profile of extragalactic sources in the following way (see Ducourant et al. 2022 for more details). Gaia scans each source at a range of transit angles during the course of its mission. These observations are mainly one-dimensional (nine one-dimensional Astro Field (AF) windows plus the two-dimensional Sky Mapper (SM) window), but after a sufficient number of transits, most of the surface of the source has been covered by these transits. The CU4 module attempts to reproduce these observed windows from a large number of simulations of images of galaxies, each with different shape parameters from which *Gaia*-like windows are extracted. The parameters that produce the best fit to the observations are taken as the profile of the source.

The module is only applied to a pre-selected list of extragalactic sources (summarized below). Fits are made for the flux profiles for two types of objects: quasars and their decomposition into quasar and host galaxy; and galaxies.

For the quasars, the module first compares the mean integrated flux of the source in the small AF window (707 mas x 2121 mas) to the mean integrated flux in the large SM window (4715 mas x 2121 mas) (Gaia Collaboration et al. 2016). A larger flux in the SM window is interpreted as a detectable host galaxy, and the surface brightness profile is fit as a combination of an exponential circular profile for the central active nucleus and a Sérsic profile (including ellipticity and position angle) for the host galaxy (see Fig. 2). The surface brightness profile parameters of the host galaxy are produced only when there is no other source present within 2.5″, and only for those sources with a half light radius smaller than 2.5″, in order to avoid too large an extrapolation of the profile and so to increase the reliability of the parameters.

For the galaxies, all the objects processed exhibit flux excess in the SM window when compared to the mean flux in AF window (see Fig. 2), indicating that these sources are clearly extended. Two independent surface brightness profiles are fit for all objects: a Sérsic and a de Vaucouleurs profile.

The pre-defined list of extragalactic sources for these two types of processing was determined as follows. For quasars, several major catalogues of quasars and candidates were compiled: AllWISE (Assef et al. 2018; Secrest et al. 2015), HMQ (Flesch 2015), LQAC3 (Souchay et al. 2015), SDSS-DR12Q (Pâris et al. 2017), ICRF2 (Ma et al. 2009), and a selection of unpublished classifications of *Gaia* DR2 quasars based on photometric variability (Rimoldini et al. 2019). Together this gave a list of 1.4 million sources. Of these, we retained for analysis in *Gaia* DR3 a subset of 1 103 691 sources, each of which has at least 25 *Gaia* observations that together cover at least 86% of the surface area of the source. For the galaxies, a machine learning analysis of *Gaia* DR2 combined with the WISE survey (Cutri & et al. 2012) was used to identify 1.9 million galaxy candidates (Krone-Martins et al. 2022). The same filtering of sources as for the quasars reduced this to 914 837 galaxies to be analysed.

2.7. Gaia-CRF3 (CU3)

One of the outputs of the astrometric solution in *Gaia* DR3 is the selection of a set of sources whose positions and proper motions define the celestial reference frame of *Gaia* DR3, called *Gaia*-CRF3. These correspond to sources cross-matched between *Gaia* and several external quasar catalogues, and selected according to specific quality metrics. The procedure to define this source list is described in Gaia Collaboration & Klioner et al. (2022) and Gaia Collaboration et al. (2021). This source sample also represents an official realisation of the International Celestial Reference System (ICRS) at optical wavelengths, as acknowledged by Resolution B3 of the IAU (2021).

3. *Gaia* DR3 tables with extragalactic content

The extragalactic content of *Gaia* DR3 is provided through a number of tables and fields. These list, among other measures, the outputs of the modules described in the previous section.

The `gaia_source` table provides two dedicated flags (`in_qso_candidates` and `in_galaxy_candidates`) that indicate the presence of a given source in the respective tables of the same name (described below). It also lists the DSC-Combmod probabilities for the quasar and galaxy classes (Sect. 2.1). The table `astrophysical_parameters` lists all the parameters produced by the modules in CU8, namely DSC, QSOC, UGC, and OA. Further results from OA (Sect. 2.4) are provided in the `oa_neuron_information` table. These tables contain all sorts of objects, not just (candidate) extragalactic ones. The tables `vari_classification_result` and `vari_agn` provide information on AGN identified through the photometric light-curves (Sect. 2.5). As a complement to the *Gaia*-CRF3 table carried over from *Gaia*-EDR3 (table `agn_cross_id`), there is a new table `gaia_crf3_xm` in *Gaia* DR3 that provides the complete cross-match information between the *Gaia*-CRF3 sources and the external catalogues in which they were identified (Gaia Collaboration & Klioner et al. 2022).

3.1. Integrated tables: `qso_candidates` and `galaxy_candidates`

In addition to the above tables, two integrated tables – `qso_candidates` and `galaxy_candidates` – are a compilation of the results from all processing modules that have classified or analysed extragalactic objects. While some of their columns are copies of information available in the above-mentioned tables, the rest are provided exclusively through these integrated tables. This is the case for the DSC class labels and the redshifts stemming from QSOC and UGC, as well as the results from the surface brightness profile analysis. These two integrated tables are limited to sources that are more likely to be extragalactic, and have been selected using a number of different selection rules that are defined in the online documentation. Below we provide just a summary of these rules.

The `qso_candidates` table is constructed as follows.

- Sources for which the quasar class probability was larger than 0.5 for any of the three DSC classifiers (Specmod, Allosmod, Combmod – see Sect. 2.1) are included. In addition to this, QSOC sources with reliable redshifts were also added (Sect. 2.2). This reliability is determined from a combination of rules involving quality flags and *Gaia* photometry thresholds (for details see Delchambre et al. 2022).
- Sources based on the analysis of photometric light curves (Vari-Classification, Sect. 2.5) were selected when their class label was set to AGN. This class label is defined in Rimoldini et al. (2022). Almost all of these sources are also part of the Vari-AGN sample, but a handful are not and they have also been added to the integrated quasar table.
- Quasars for which the surface brightness profile was analysed as described in Sect. 2.6 were included provided the presence or not of a host galaxy could be assessed with sufficient confidence. An ancillary table `qso_catalogue_name` provides the name of the external catalogues that were used to select the sources that entered this pipeline.
- All sources used to define the *Gaia*-CRF3 (provided in table `agn_cross_id`, see Sect. 2.7 and Gaia Collaboration & Klioner et al. 2022) are in the quasar table, and a dedicated flag, `gaia_crf_source`, identifies them.
- OA does not contribute any additional sources to the table. We simply add class labels from OA to sources that are included by the above selections. These labels are not necessarily limited to be extragalactic source labels.

The `galaxy_candidates` table is constructed as follows.

- Sources for which the galaxy class probability was larger than 0.5 for any of the three DSC classifiers (Specmod, Allosmod, and Combmod – see Sect. 2.1) are included. In addition to this, UGC sources with reliable redshifts were also

added (Sect. 2.3). The reliability is determined by a combination of two sets of rules, one concerning the quality of the BP/RP spectrum of the source, the other involving the comparison of outputs from three models estimating the redshift (for details see Delchambre et al. 2022).

- Sources identified by Vari-Classification (Sect. 2.5) were selected if their class label was set to GALAXY. For a description of how this class label was defined, see Rimoldini et al. (2022).
- Galaxies for which the surface brightness profile was analysed as described in Sect. 2.6 were included if the light profile parameters could be derived with sufficient quality. In complement to this, an ancillary table `galaxy_catalogue_name` provides the name of the external catalogues that were used to select the sources that entered this pipeline. In *Gaia* DR3 the only applicable catalogue is that described in Krone-Martins et al. (2022).
- As for the `qso_candidates` table, OA does not contribute additional sources. It only provides additional columns, which are filled just for those sources that were processed by OA.

The source lists according to the above selection criteria were concatenated into two lists, one for the `qso_candidates` table and one for the `galaxy_candidates` table. A complete list of the parameters (table columns) available in each table is given in the online documentation. Columns are filled for all sources regardless of how they are selected; thus a source may have a DSC probability that does not meet the above DSC selection criteria, for example (see Table 1). Not all parameters are available for all sources, as not all sources were treated by all modules. There are 6 649 162 sources in the `qso_candidates` table and 4 842 342 in the `galaxy_candidates` table. This large number of sources is mostly due to the selection rules of the DSC module, which favour completeness over purity (see section 2.1 and Delchambre et al. 2022). Users should therefore be aware that there is significant stellar contamination in these tables. For DSC this can be addressed using the `classlabel_dsc_joint` field. We address more generally how to build purer sub-samples in Sect. 8. There are 174 146 sources in common between the two tables, and their union contains 11.3 million sources.

Table 1 gives an overview of how many sources from each module contribute to the integrated tables. Source overlaps between the modules within each table are shown in Tables 2 and 3, and graphically represented in the Venn diagrams in Figs. 3 and 4. Information about the distribution of the parameters featured in the tables is provided in the next section.

To estimate the overall purity of the integrated tables, we must be aware that modules with different purities can contribute the same source to a table. The estimation can be simplified, however, when we consider that all modules except DSC have similar high purities. Specifically, for the `qso_candidates` we assume that the modules other than DSC have an average purity of 96%, compared to a global DSC purity of 24%. From Fig. 3 we see that 4.1 million sources are contributed only by DSC, with the remaining 2.6 million contributed by the other modules. This gives an overall purity of the `qso_candidates` table of 52%. In a similar way, we estimate the overall purity of the `galaxy_candidates` to be 69%. We show how to obtain a purer sub-sample in Sect. 8.

Table 1. Number of sources from each of the extragalactic processing modules contributing to the `qso_candidates` and `galaxy_candidates` tables (second column), or to the set of parameters featured for the respective modules (third column). The difference between the two columns indicates the number of sources where parameters are provided despite the sources not being eligible according to the selection rules of that module. A given source can be contributed by more than one module.

Module	Selected sources	Featuring parameters
`qso_candidates`	6 649 162	
DSC	5 543 896	6 647 511
QSOC	1 834 118	6 375 063
Vari-Classification	1 035 207	1 122 361
Vari-AGN	872 228	872 228
Surface brightness	925 939	1 084 248
Gaia-CRF3	1 614 173	1 614 173
OA	N/A	2 803 225
`galaxy_candidates`	4 842 342	
DSC	3 726 548	4 841 799
UGC	1 367 153	1 367 153
Vari-Classification	2 451 364	2 477 273
Surface brightness	914 837	914 837
OA	N/A	1 901 026

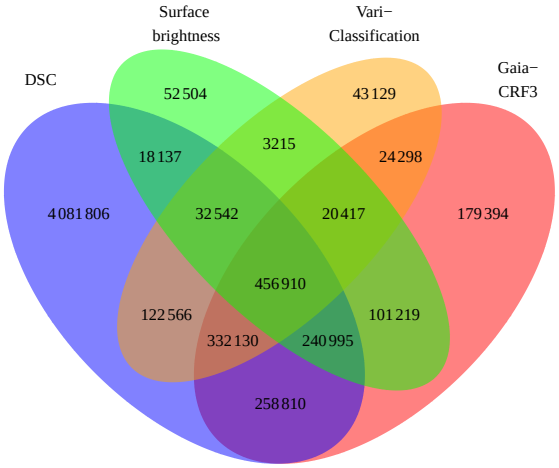

Fig. 3. Quadruple Venn diagram for contributions to the `qso_candidates` table from DSC, the Surface brightness sample, Vari-Classification, and *Gaia*-CRF3.

4. Basic properties

4.1. Parameter distributions

4.1.1. Integrated tables

Figure 5 shows the sky distribution on a logarithmic density scale of all sources in the `qso_candidates` and `galaxy_candidates` tables. As already noted, there is considerable contamination in these due to misclassifications and the completeness-driven nature of the tables (i.e. the absence of filtering in some modules). This is apparent from the overdensities around the Large and Small Magellanic Clouds (LMC and

Table 2. Source overlaps between the modules contributing to the `qso_candidates` table. See text for details about the module names.

Module	Surface brightness	Vari-classification	Vari-AGN	OA	QSOC	DSC
Gaia-CRF3	819 541	833 755	722 211	550 807	672 454	1 288 845
Surface brightness		513 084	483 786	278 078	458 241	748 584
Vari-classification			872 184	245 318	477 971	944 148
Vari-AGN				186 836	442 436	814 315
OA					896 173	2 085 554
QSOC						1 097 229

Table 3. As 2 for modules contributing to the `galaxy_candidates` table.

Module	Vari-classification	UGC	OA	DSC
Surface brightness	634 550	388 552	434 880	530 411
Vari-Classification		972 929	1 070 865	1 529 594
UGC			190 583	1 351 222
OA				840 409

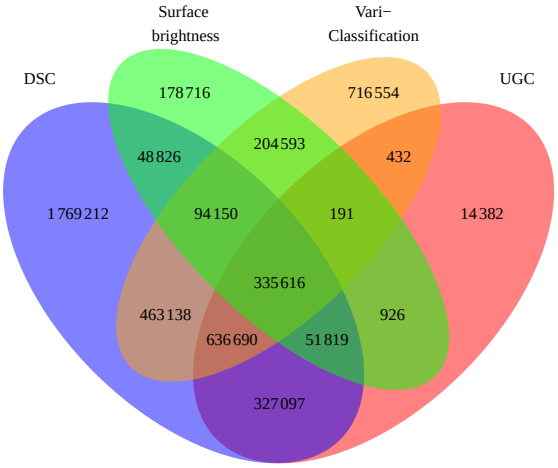

Fig. 4. Quadruple Venn diagram for contributions to the `galaxy_candidates` table from DSC, the Surface brightness sample, Vari-Classification, and UGC.

SMC). If we exclude generous regions around the LMC and SMC (defined in appendix B), then the number of sources in the `qso_candidates` table drops to 3.95 million (59% of the full table) and the number of sources in the `galaxy_candidates` table drops to 4.67 million (96% of the full table). Some patterns are also an artefact of the use of input lists for some of the modules. Many of these sources are also faint, with poorer data in *Gaia* DR3, as can be seen in Fig. 6. There is also a small fraction of sources that are too bright to be genuine quasars or galaxies, which is an inevitable consequence of even a small misclassification probability and limited filtering.

The *Gaia* colour–colour diagram (CCD) and colour–magnitude diagram (CMD) are shown in Fig. 7. Quasars and galaxies separate quite well, but recall that *Gaia* observes primarily those galaxies with point-source like cores. What is not seen in these diagrams is the distribution of the stars, which outnumber true quasars and galaxies by a factor of 500–1000 in *Gaia* DR3, and which make it hard to identify extragalactic objects based only on their *Gaia* colours.

4.1.2. DSC subset of the integrated tables

DSC is the dominant contributor to the `qso_candidates` and `galaxy_candidates` tables, so we look here at two subsets for each table defined by the DSC class labels (Sect. 2.1). The first is selected by `classlabel_dsc`, which gives 5 243 012 quasars in the `qso_candidates` table (class `quasar`) and 3 566 085 galaxies in the `galaxy_candidates` table (class `galaxy`). Through comparison to SDSS spectroscopic classifications, and accommodating for the significant contamination by stars, we estimate these samples to have rather low purities of 24% and 22% respectively (see Bailer-Jones 2021, summarized in Delchambre et al. 2022, and Sect. 6.2 below). The second subset is the purer one identified using `classlabel_dsc_joint`, which gives 547 201 quasars in the `qso_candidates` table and 251 063 galaxies in the `galaxy_candidates` table. These two sets are estimated to have higher purities of 62% and 64% respectively, and of 79% and 82% respectively if we look only at higher latitudes ($|b| > 11.54°$).

Figure 8 shows the *Gaia* colour–colour diagrams for quasars in the `qso_candidates` table according to these two subsets. The upper panels show the DSC-Combmod probabilities. In the upper left panel we see that there are sources far away from the main clump of quasars, but the lower panel reveals that there are very few of them. These are all removed in the `classlabel_dsc_joint = quasar` set (right column), which shows only high Combmod probabilities. Figure 9 shows the corresponding colour–colour diagrams for the `galaxy_candidates` table. Again we see how the set defined by `classlabel_dsc_joint = galaxy` has a tighter distribution and higher Combmod probabilities than the less pure set defined by `classlabel_dsc = galaxy`. Similar figures showing the quasar and galaxy populations together are shown in Delchambre et al. (2022). These also show that use of the joint label preferentially removes fainter, lower signal-to-noise

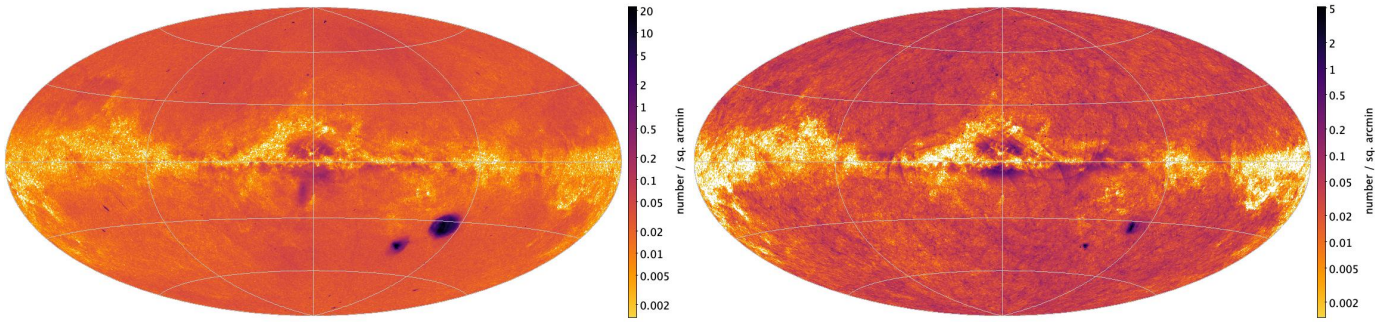

Fig. 5. Galactic sky distribution of all the sources in the `qso_candidates` table (left) and `galaxy_candidates` table (right). The plot is shown at HEALpixel level 7 (0.210 sq. deg.) in Hammer–Aitoff projection. The colour scale, which is logarithmic, covers the full range for each panel, so is different for each panel.

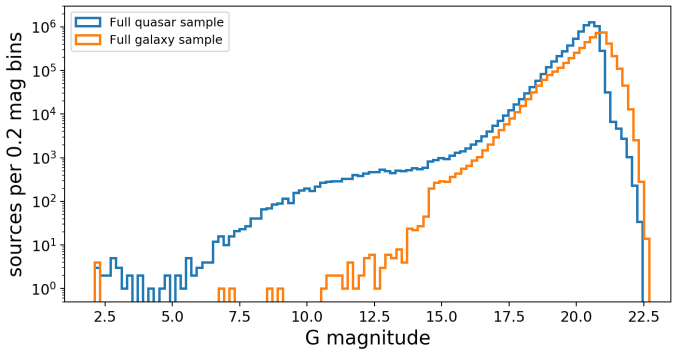

Fig. 6. G-band magnitude distribution of all objects in the `qso_candidates` (blue) and `galaxy_candidates` (orange) table on a logarithmic scale. The brightest known quasar (3C273 – `source_id` 3700386905605055360) has a G magnitude of 12.8.

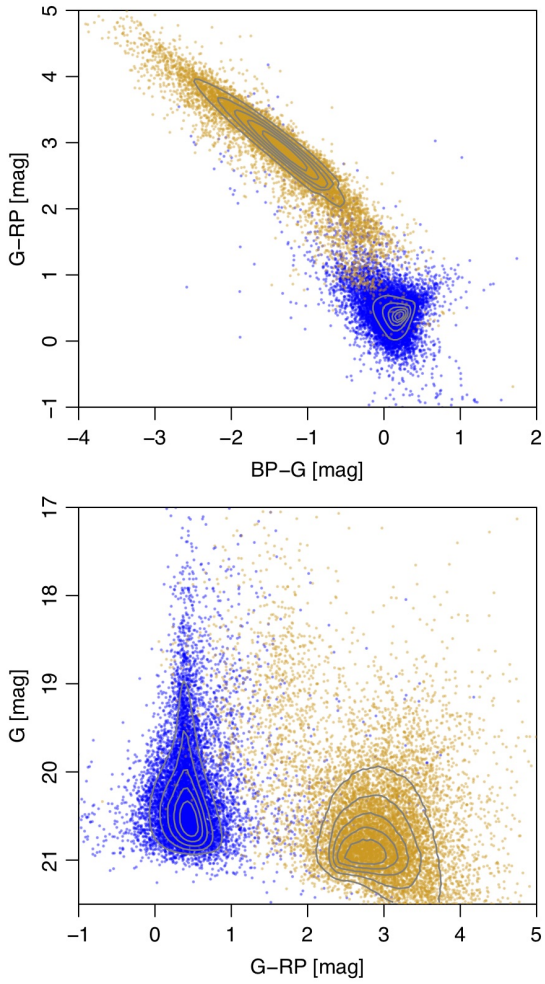

Fig. 7. Colour–colour diagram (top) and colour–magnitude diagram (bottom) for all sources in the `qso_candidates` table (blue) and `galaxy_candidates` table (orange). The contours show density on a linear scale. The points are a random selection of 10 000 sources for each class.

sources, as these are less likely to get a high probability classification in both Specmod and Allosmod.

One thing to bear in mind is that Specmod and Allosmod do not deal with identical sets of sources, because these classifiers require different input data. In particular, Allosmod requires parallaxes and proper motions, that is 5p or 6p astrometric solutions (see Lindegren et al. 2021a for the definition of these solutions). Galaxies often only get 2p solutions (no parallax or proper motion) on account of their physical extent. Of the 3 566 085 million sources in the `galaxy_candidates` table with `classlabel_dsc = galaxy`, 3 367 211 have all three photometric bands, but of these, only 1 015 462 have parallaxes and proper motions and so can be classified by Allosmod (these numbers are for the whole sky, so including the LMC and SMC). As `classlabel_dsc_joint` can only be set to galaxy when Allosmod results are present, the change in the distribution we see in Fig. 9 for the two class labels is partially due to this. Plots in Delchambre et al. (2022) show the change when only considering the subset with 5p or 6p solutions. Most quasars, in contrast, do have 5p or 6p solutions: Of the 5 243 012 sources in the `qso_candidates` with `classlabel_dsc = quasar`, 5 086 531 have all three photometric bands, of which 4 815 212 have parallaxes and proper motions.

Because DSC is not the only contributor to the integrated tables, some of the sources in these tables have DSC class labels that are not the class of the table. In the `qso_candidates` table, 156 970 sources have `classlabel_dsc` set to `galaxy`, and 12 302 have `classlabel_dsc_joint` set to `galaxy`. In the `galaxy_candidates` table, the numbers with these two class-labels set to `quasar` are 12 933 and 234 respectively.

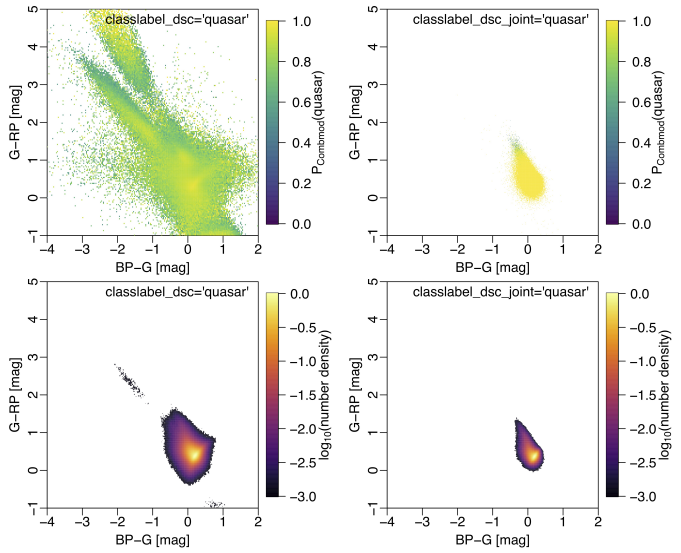

Fig. 8. Colour–colour diagram for sources in the `qso_candidates` table, excluding regions around the LMC and SMC. The left column shows sources with `classlabel_dsc = quasar` (2.77 million sources), the right column shows sources with `classlabel_dsc_joint = quasar` (the purer subset, 0.52 million sources). These numbers refer to the number of sources plotted, which are those that have all *Gaia* bands. The upper panel shows the mean DSC-Combmod probability for the quasar class (the field `classprob_dsc_combmod_quasar`). The lower panel shows the density of sources on a log scale relative to the peak density in that panel (densities 1000 times lower than the peak are not shown).

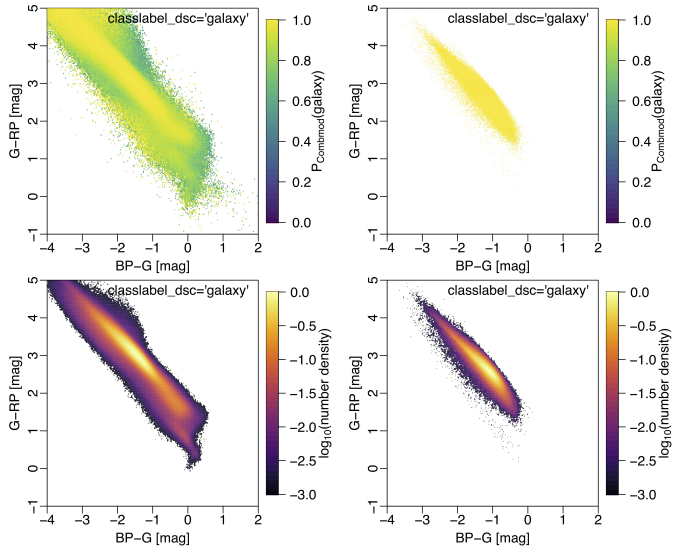

Fig. 9. As Fig. 8, but for sources in the `galaxy_candidates` table. There are 3.24 million sources with `classlabel_dsc = galaxy` and 0.25 million sources with `classlabel_dsc_joint = galaxy` (in both cases excluding the regions around the LMC and SMC, and requiring all three *Gaia* bands).

4.2. BP/RP spectra

Gaia observes all of its targets with the low resolution ($30 \leq \lambda/\Delta\lambda \leq 100$) BP/RP slitless spectrograph (Carrasco et al. 2021). 1.6 billion of these were used by DSC-Specmod for classification (section 2.1; Delchambre et al. 2022), but only a fraction of these are published in *Gaia* DR3. Spectra for all sources brighter than $G = 17.35$ mag with at least 15 retained observations in each of BP and RP are published in *Gaia* DR3, amounting to 220 million sources. This includes few extragalactic sources, so a small set of these were added. In total, BP/RP spectra of 163 000 quasar candidates and 26 500 galaxy candidates in the integrated tables are published in *Gaia* DR3. Of these, 119 000 and 12 600 respectively are in the purer sub-samples defined in Sect. 8.

As described in De Angeli et al. (2022), spectra are published as a set of coefficients of basis functions, from which spectra at arbitrary samplings can be produced using a published software tool. Internal to CU8, the spectra were sampled using the tool SMS-gen (Creevey et al. 2022), which is what we used to produce the spectra shown in this section. In all cases the spectra are the mean (epoch-averaged) spectra over a time span of up to 34 months.

4.2.1. Quasars

Figure 10 shows the BP/RP spectra for 42 944 quasars with published BP/RP coefficients (field `has_xp_continuous = true` in `gaia_source`), `classprob_dsc_combmod_quasar` > 0.01 and spectroscopically confirmed redshift in the Milliquas 7.2 quasar catalogue of Flesch (2021) (`type = Q`). A search radius of 1″ was used to match the *Gaia* sources to their Milliquas counterparts, leading to a redshift coverage of $0.052 \leq z \leq 4.358$. The cut on the DSC Combmod quasar probability ensures that obvious stellar contaminants contained in our cross-match are discarded. The median magnitude of the sources in Figure 10 is $G = 18.53$ mag. *Gaia* observes much fainter quasars, but the BP/RP spectra of many of these will only be released in *Gaia* DR4. While we clearly see common quasar emission lines in this averaged plot, they are not necessarily visible in the low signal-to-noise ratio (S/N) spectra of individual faint quasars. Similarly, wiggles that are an artefact of the Hermite spline representation of the BP/RP spectra (De Angeli et al. 2022) tend to lower the contrast of these emission lines compared to the continuum. These wiggles smooth out faint spectral features, and can be confused with emission lines, as both have comparable strength in low S/N spectra. Typically, though, the strongest spectral features – Lyα, C IV, Hβ, and Hα – are retained in $G < 20$ mag spectra. We also see in Fig. 10 that regions at wavelengths below 430 nm and above 650 nm in BP, and below 630 nm and above 950 nm in RP, contain little flux: spectral features in these regions generally have low S/N, complicating their detection by the DSC and QSOC algorithms.

4.2.2. Galaxies

Figure 11 shows four representative spectra of galaxies as observed by *Gaia* (top row) and their corresponding SDSS spectra (bottom row). The first SDSS spectrum on the left shows only absorption lines, suggesting an early type galaxy with little or no star formation activity (the few spikes are caused by cosmic rays). These lines are barely detectable, if at all, in the low-resolution BP/RP spectrum. The two middle spectra show strong emission lines characteristic of active star formation. The strongest is the Hα emission with [N II] lines on either side. This set of three lines is unresolved in the RP spectrum where it merges into a single and wide emission feature. Similarly, in the BP spectrum the Hβ and [O III] emission lines are merged into another wide peak. The last spectrum on the right is classified as a 'GALAXY AGN' in SDSS. The corresponding BP/RP

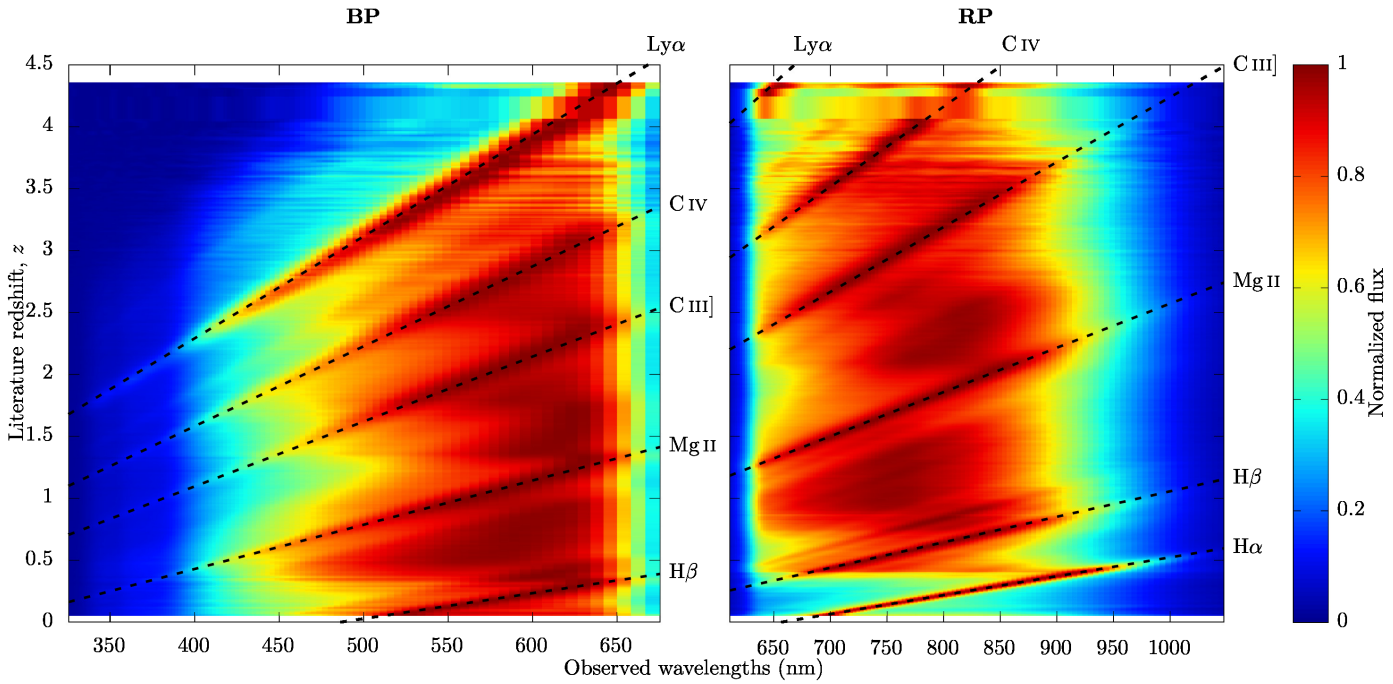

Fig. 10. Distribution of the BP flux (left) and RP flux (right) as sampled by SMS-gen (Creevey et al. 2022) of 42 944 quasars published in *Gaia* DR3 that have spectroscopically confirmed redshifts in the Milliquas 7.2 quasar catalogue of Flesch (2021) (`type = Q`). Dotted lines show the dominant quasar emission lines. Spectra are individually normalized in order to have a maximum flux of 1.0 and are then averaged in redshift bins of 0.01, with the inverse variance of the sampled fluxes used as the weight during the computation of the mean.

spectrum, due to the much lower resolution – and the already mentioned wiggles – shows much less prominent features.

4.3. Surface brightness profiles

4.3.1. Quasars

The majority of the 1 103 691 quasars analysed in terms of surface brightness lie in the diagonal of Fig. 2. These sources are considered point-like with no host galaxy detectable by *Gaia*. A group of 64 498 exhibit a clear extension, indicative of a host galaxy, as evidenced by larger fluxes in the SM window than in the AF window (Sect. 2.6). For these sources the flag `host_galaxy_detected = true` is set. Among these, a robust solution from the fitting process was derived for 15 867 sources and their surface brightness profile is given in the catalogue. The flag `host_galaxy_flag` indicates the outcome of the fitting process for all sources considered. Values of 1 and 2 are good fits, indicating detection of a host galaxy. 3 indicates that no host could be found, whereas 4 is a poor fit. Sources with `host_galaxy_flag` = 5 or 6 show no evidence of a host galaxy in our analysis, due to non-convergence of the algorithm or the presence of a close neighbour, respectively.

Figure 12 shows the spatial distribution on the sky of the 1 103 691 quasars analysed. The coverage is inhomogeneous due to the limited sky coverage of the catalogues that constitute the quasar input list (Sect. 2.6) but it also reflects the scanning law of *Gaia*, as we only analyse sources that have at least 25 focal plane transits. The empty zones correspond either to the Galactic plane or to zones of lower frequency of scanning in *Gaia* DR3.

The distribution of the Sérsic index for all the quasars has a mode at 0.9 and a mean of 1.9. These values are consistent with quasars hosted by galaxies with disk-like light profiles, in agreement with a recent study of the surface brightness of host galaxies from the Hyper Suprime-Cam Subaru Strategic Program (Li et al. 2021).

The distribution of position angles of host galaxies is roughly uniform, as expected, although there is a small excess at around 90°. These are sources with negligible ellipticity for which the position angle is meaningless. In such cases, our fitting algorithm favours a 90°position angle. The same is true for the galaxy sample discussed in the next section.

226 160 of the quasars processed have a spectroscopic redshift listed in Milliquas 7.2 (Flesch 2021) (selection `TYPE = Q`). 2084 of these have a host galaxy detected by *Gaia*. Figure 13 shows the distribution of these redshifts. As expected, the quasars with a host galaxy have small redshifts (mean z=0.54) whereas those without a visible host galaxy have larger redshifts (mean z=1.71). In a few cases the host is detected for larger redshifts. These sources are usually very faint (G>20 mag) and suffer either from uncertainties in the light profile fit or in the redshift measurement. The host galaxies resolved by *Gaia* have an effective radius (encompassing half of the total light) distribution with a peak at around 800 mas.

4.3.2. Galaxies

The surface brightness profile module processed 914 837 galaxies. We see from Fig. 2 that all of these have a clear spatial extension.

The distribution of the effective radius of the de Vaucouleurs profile as measured by *Gaia* is shown in Fig. 14 as a function of the *Gaia* redshifts (given by `redshift_ugc` in table `galaxy_candidates`). The redshifts are all below about 0.5, with a mean value of 0.16. As expected, the closer a source is to us, the larger its effective radius. There is a slight accumulation of effective radii at 8000 mas, which corresponds to the bound

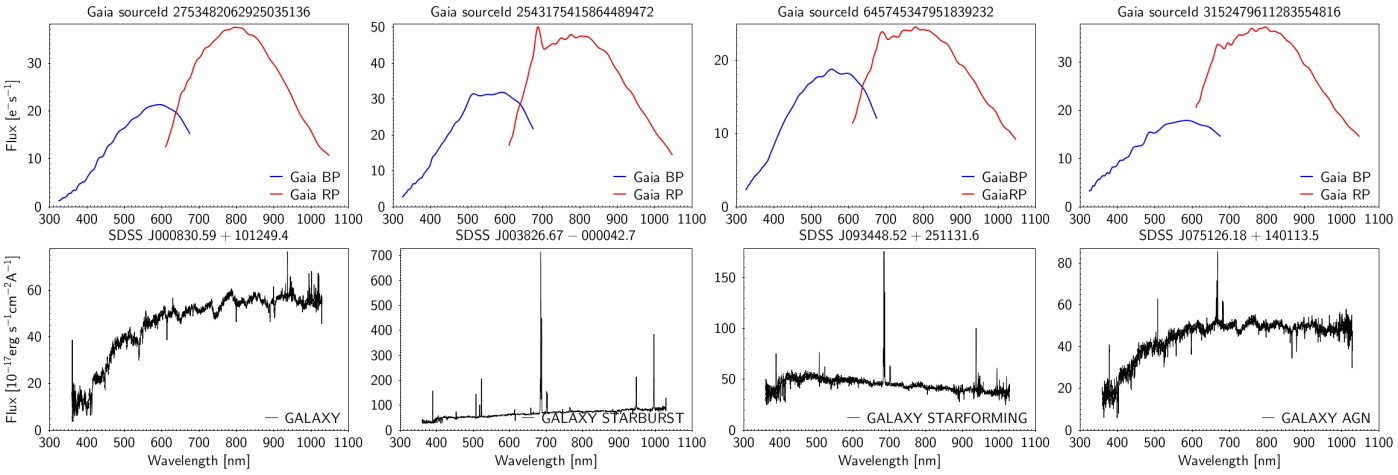

Fig. 11. Galaxy spectra. Top row: Representative mean BP and RP *Gaia* spectra for four galaxies. Bottom row: The spectra for the same galaxies as observed with the SDSS-BOSS spectrograph (the SDSS class and subclass, if defined, are shown).

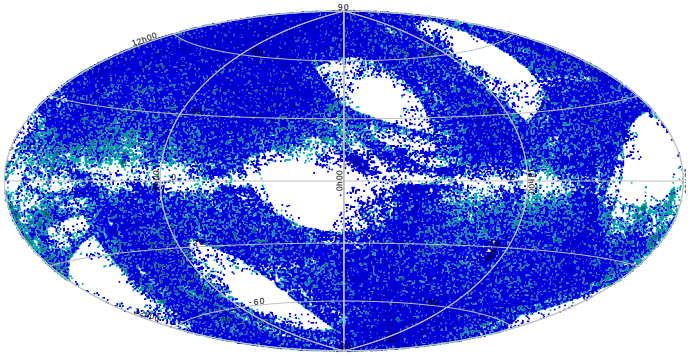

Fig. 12. Distribution in Galactic coordinates (Hammer–Aitoff projection) of the quasars processed by the surface brightness profile module. Blue points are quasars with a host galaxy detected (`host_galaxy_detected = true`) and turquoise points are those without a host galaxy.

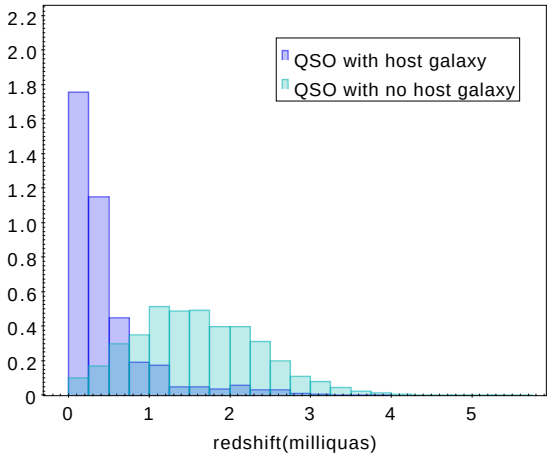

Fig. 13. Normalized distribution of the redshifts from Milliquas v7.2aa (Flesch 2021) of quasars that were analysed for surface brightness profiles. Blue shows the 2000 quasars for which a host galaxy was detected by *Gaia*, and turquoise the remaining 224 000 quasars for which no host galaxy was detected.

of the parameter search domain, with the results that for larger galaxies the radius would remain at 8000 mas.

Figure 15 shows the distribution of these galaxies on the sky. As with the quasars in the previous section, we see an uneven distribution due primarily to the required minimum number of observations.

The distribution of the Sérsic index peaks at around 4.5, which is consistent with the fact that the on-board detection algorithm favours elliptical types (de Souza et al. 2014). A few thousand galaxies have a Sérsic index below 2, indicative of disk galaxies. A visual inspection of a fraction of these reveals that most of them exhibit a compact bright bulge. The effective radius of the Sérsic profile has a peak value around 1800 mas and a de Vaucouleurs radius of around 1000 mas, which is typical of sources with a mean redshift of 0.13.

The ellipticities derived from *Gaia* exhibit a peak value around 0.25. This is more or less what is expected from the projection of oblate ellipsoids (representative of elliptical) onto the plane of the sky and is also observed in other surveys, such as Padilla & Strauss (2008).

4.4. Light curves

4.4.1. AGN

Gaia DR3 includes about a million variable AGN candidates in the `vari_classifier_result` table, which were selected mainly on the basis of their variability properties. For these, the epoch photometry in the G, G_{BP}, and G_{RP} bands is published in the `light_curve` datalink table. A complete description of the selection methods can be found in Rimoldini et al. (2022), and are summarized below. More restrictive criteria were applied to achieve the higher purity sample comprising 872 228 candidates in the `vari_agn` table (Sect. 2.5), the characteristics of which are analysed in Carnerero et al. (2022).

Of the one million G-band light curves of the variable AGN, 90% contain between 20 and 244 focal plane transits covering 795 to 1038 days (after applying time series filters described in Sect. 10.2 of the online documentation). On average they have 39 focal plane transits over 925 days, which is sufficient to follow the long-term variability of most AGN. Figure 16 shows the light curves in the G, G_{BP}, and G_{RP} bands of three sources belonging to different AGN classes: a) the type 1 Seyfert galaxy

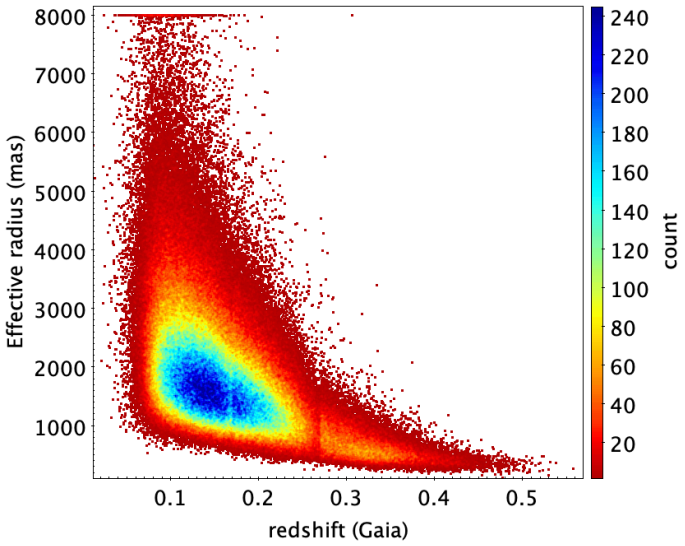

Fig. 14. Distribution of the effective radius (de Vaucouleurs profile) of galaxies processed by the surface brightness profile module as function of the redshifts measured by *Gaia* (`redshift_ugc`).

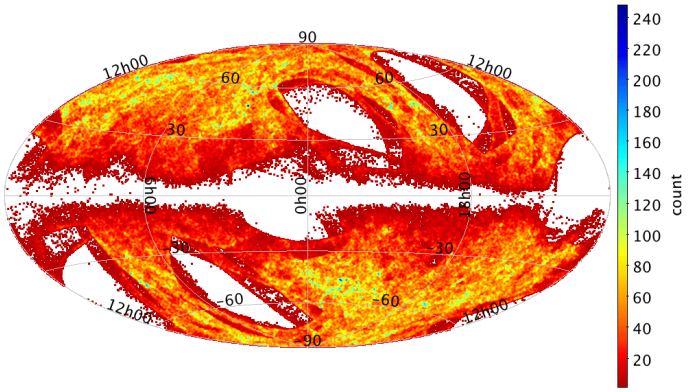

Fig. 15. Distribution in Galactic coordinates (Hammer-Aitoff projection) of the galaxies processed by the surface brightness profile module. The colours show the density on a linear scale.

PG 0921+525; b) the blazar CTA 102, which was observed during its historical 2016–2017 maximum (Raiteri et al. 2017), resulting in the most variable object of the sample; c) the quasar B2 0945+22.

Photometrically-variable AGN candidates from supervised classification were verified and further down-selected by a series of filters that use *Gaia*-CRF3 (Gaia Collaboration & Klioner et al. 2022) as a reference sample. Variability-related constraints were set on: the index of the structure function (Simonetti et al. 1985, `structure_function_index` in the table `vari_agn`); quasar versus non-quasar metrics (Butler & Bloom 2011, `qso_variability` and `non_qso_variability` in table `vari_agn`); the Abbe (also called von Neumann) parameter (`abbe_mag_g_fov` in table `vari_summary`) in the G band versus the renormalized unit weight error of the astrometric solution (`ruwe` in table `gaia_source`). Additional cuts were made in the $G_{\rm BP}-G$ versus $G-G_{\rm RP}$ colour space, on parallaxes and proper motions, on the environment source number density (to avoid crowded sky regions), on the scan angle correlation with photometric variation (to remove artificial effects; see Holl et

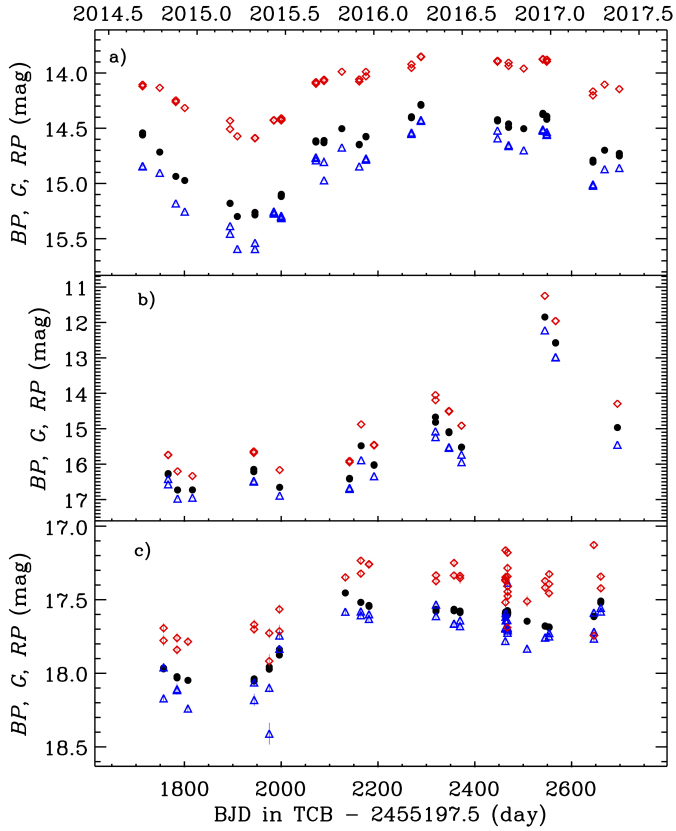

Fig. 16. Light curves in the G (black dots), $G_{\rm BP}$ (blue triangles), and $G_{\rm RP}$ (red diamonds) bands of some variable AGN sources. From top to bottom: a) the type 1 Seyfert galaxy PG 0921+525 (source_id 1019788071166861952); b) the blazar CTA 102 (source_id 2730046556694317312) caught during its historical 2016–2017 outburst; c) the quasar B2 0945+22 (source_id 640411921988216576).

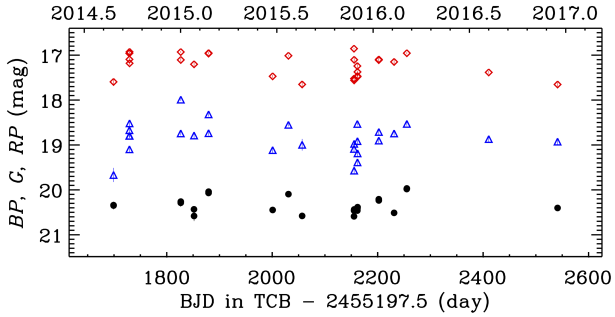

Fig. 17. Light curves in the G (black dots), $G_{\rm BP}$ (blue triangles), and $G_{\rm RP}$ (red diamonds) bands of the known galaxy LEDA 2268723 (source_id 377643902971151872).

al. 2022), and finally on the variability probability (to deal with clearly variable objects).

4.4.2. Galaxies

About 2.5 million galaxies in the `galaxy_candidates` table were selected based on the properties of their light curves. (Only a subset of these light curves are published in *Gaia* DR3; see Sect. 2.5.) *Gaia* scans individual objects multiple times at different position angles. For extended objects this can produce an apparent – but spurious – photometric variability, because on each

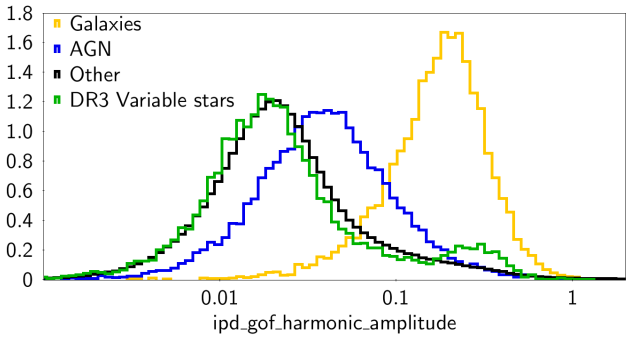

Fig. 18. Distributions (normalized by area) of the field `ipd_gof_harmonic_amplitude` for sources of various classes in the *Gaia* Andromeda Photometric Survey. 'Other' includes constants and those variable objects that were not targeted in *Gaia* DR3.

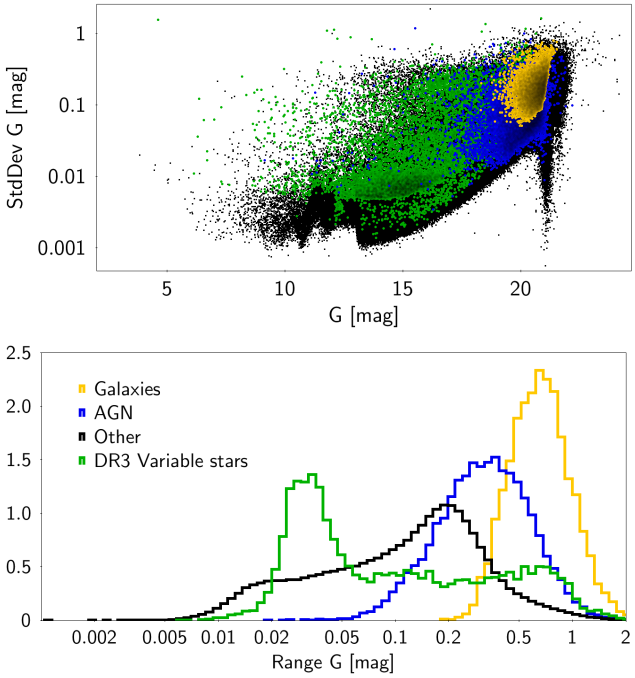

Fig. 19. Statistics of light curves of objects in the *Gaia* Andromeda Photometric Survey. Top: Standard deviation versus median G magnitude. Bottom: Normalized distribution of minimum-to-maximum variability range for G band light curves. Both panels are colour-coded as in Fig. 18. The distributions overlap in the upper panel, with galaxies covering AGN, for example.

scan only part of the total flux is collected by the limited size in the allocated window (Holl et al. 2022). Figure 17 shows the light curve of a known galaxy, in which we see variations in excess of 0.6 mag in G. Figure 18 shows the distribution (normalized by area) of the parameter `ipd_gof_harmonic_amplitude` for galaxies, AGN, *Gaia* DR3 variable stars, and other objects in the *Gaia* Andromeda Photometric Survey. This parameter measures the amplitude of the variation of the Image Parameters Determination goodness-of-fit statistic as function of the scan direction angle. Because galaxies are often extended objects at the *Gaia* resolution, they tend to have a larger value of this parameter than other types of objects. The galaxies that are detected by variability are based on this type of spurious signal. Figure 19 shows how the magnitude variability distribution of galaxies within 5.5° of the Andromeda Galaxy (M31) compares to that of other sources in the same classes as in Fig. 18. We note

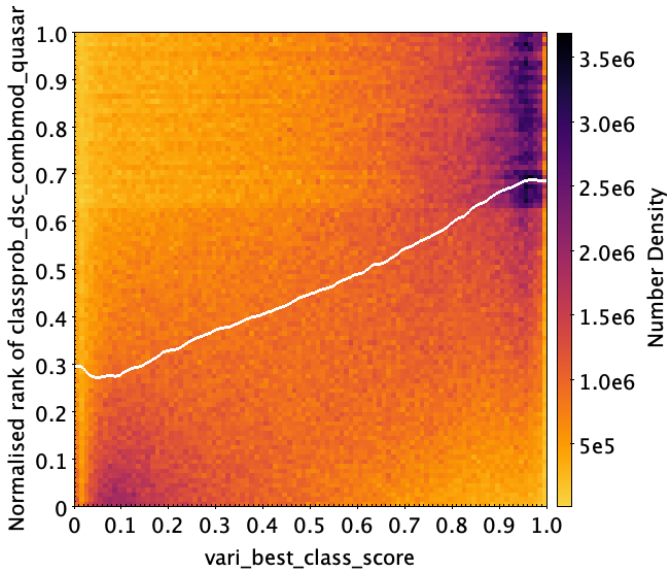

Fig. 20. Comparison of DSC quasar classification probabilities (transformed to normalized ranks) with scores from the variability analysis. Darker colours depict higher densities, and the white line indicates the median rank. We see a broad agreement between the highest and lowest ranked quasars.

that *Gaia* DR3 variable stars still amount to a relatively small fraction of all variables detected in *Gaia*. The brightness variations of galaxies overlap with high-amplitude tails of the distributions of other classes.

5. Internal comparison

5.1. Classification

The various extragalactic modules (Sect. 2) use different methods and data. This leads to a given source being classified differently in different modules, which is apparent in the `qso_candidates` and `galaxy_candidates` tables that collate results from all modules. On top of this comes the fact that different modules use different definitions of quasar and galaxy, in particular in the case of supervised learning algorithms, where the class is defined by the training data set. Tables 4 and 5 show the percentage of different classifications of the overlapping sources between modules based on the class labels where they exist (DSC, DSC-Joint, OA, Vari-Classification) or the existence of parameters (from UGC, QSOC, Vari-AGN, Surface brightness). For example, `classlabel_dsc` and Variability give different classes for 9.5% of their common sources. Such disagreements also come about because some modules focus more on high completeness, whereas others focus more on high purity (partially achieved by filtering). Recall also that the classification from a module appears in the table even if that source would not have been selected for inclusion in the table by that particular module (see Table 1). QSOC for quasars and UGC for galaxies are subsets of DSC selected with the properties described in Sections 2.2 and 2.3. Both use much lower thresholds on the DSC probabilities than do the DSC class labels.

Gaia-CRF3 does not distinguish between galaxies and quasars. Most are expected to be quasars so all are all in the `qso_candidates` table. OA works with a small fraction of sources that are generally faint and noisy so the comparison between OA and other modules should be carefully interpreted.

Table 4. Comparison of the classes of sources in the `qso_candidates` table according by its contributing modules. Each element gives the number of sources with different classifications between any two modules, expressed as a fraction of the number of sources in common between those two tables. Sources labelled unclassified in `classlabel_dsc` and `classlabel_dsc_joint` are excluded. The columns list all the modules that provide classifications. The rows list all modules that add sources to the table: the last four of these are not classifiers, but provide sources based on other labels.

QSO \ Other	DSC classification	DSC Joint classification	Variability classification	OA classification
DSC classification		0.0	5.7	59.5
DSC Joint classification	0.0		0.9	53.9
Variability classification	9.5	0.5		44.6
OA classification	30.2	1.1	7.1	
Vari-AGN	7.3	0.5	0.0	46.2
Surface brightness	20.0	2.3	0.4	46.6
Gaia-CRF3	20.8	2.0	0.2	42.3
QSOC	43.2	0.1	10.8	61.4

Table 5. As Table 4 but for the `galaxy_candidates` table.

Galaxy \ Other	DSC classification	DSC Joint classification	Variability classification	OA classification
DSC classification		0.1	1.5	13.1
DSC Joint classification	0.0		2.8	2.9
Variability classification	37.6	0.0		0.7
OA classification	59.7	6.4	1.3	
UGC	1.2	0.1	1.0	7.3
Surface brightness	42.0	0.0	0.0	2.3

DSC provides posterior class probabilities. `vari_best_class_score` (from Vari) provides the median normalized rank, which also increases from 0 to 1 with increasing reliability, but it is not a probability. To compare these quantities, we map the DSC probabilities into normalized ranks. Figure 20 compares this for `classprob_dsc_combmod_quasar` to `vari_best_class_score`. The deviation from a perfect correlation reflects the difference in input data types, training sets and class definitions, and classification methods in general.

5.2. Redshift

Redshifts are derived by two modules, the results of which are reported in the `qso_candidates` table (from QSOC) and the `galaxy_candidates` table (from UGC). Of the 174 146 sources in common between the two tables, 16 534 have a redshift derived by both modules. These are compared in Fig. 21. 7 469 of these sources have predictions with $|\Delta z| < 0.1$, and the correlation improves when restricting the comparison to QSOC redshifts with higher reliability (black dots in Fig. 21): In this subset, 105 of 166 sources have $|\Delta z| < 0.1$. Specific discrepancies arise from emission line mismatches in the QSOC redshift determination. As QSOC aims to be complete, it processes galaxies, even though UGC – by design – generally gets better predictions on these objects (see Delchambre et al. 2022 for a more detailed explanation of these emission line mismatches). UGC, in contrast, aims to be pure and is accordingly not expected to process a significant number of quasars. Figure 21 shows loci of constant QSOC redshifts. These are probably erroneous matches at the BP/RP spectral borders, where wiggles from the Hermite polynomials are confused with quasar emission lines in the templates.

Figure 22 shows the colour–colour diagram for all sources for which UGC provides a redshift value, colour-coded by redshift. We see that galaxies generally become redder in $G_{\rm BP}-G$, but bluer in $G-G_{\rm RP}$ as redshift increases from 0 to 0.4.

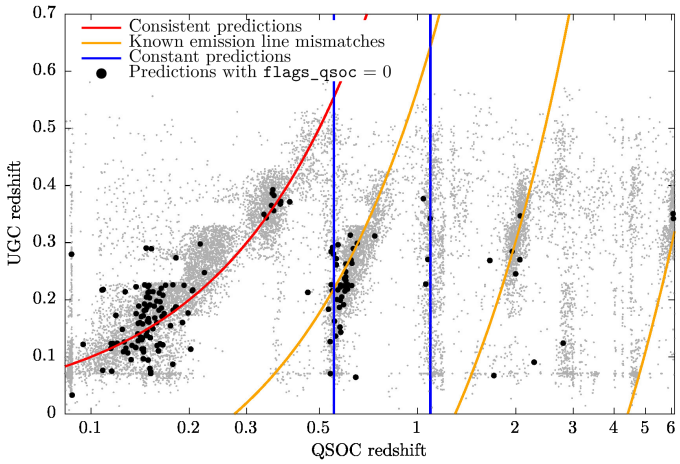

Fig. 21. Comparison between the UGC and QSOC redshifts. Grey dots correspond to all redshifts in common between the two tables, while black dots are restricted to those with `flags_qsoc = 0`, which corresponds to a higher reliability subset. The red curve denotes identical predictions in the two modules. Yellow curves highlight mismatches between common quasar or AGN emission lines, as explained in Delchambre et al. (2022), while the blue vertical lines show constant predictions by QSOC.

5.3. Sources with stellar parameters

The extragalactic tables contain sources for which stellar astrophysical parameters are also reported in *Gaia* DR3. This is expected, because stellar parameters were inferred for sources in-

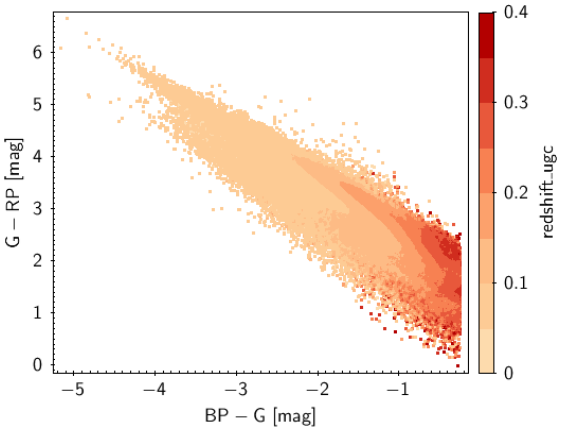

Fig. 22. Colour-colour diagram for the 1 367 153 galaxies for which redshifts are provided by UGC, colour-coded by redshift. A small number of sources have redshifts extending up to 0.6

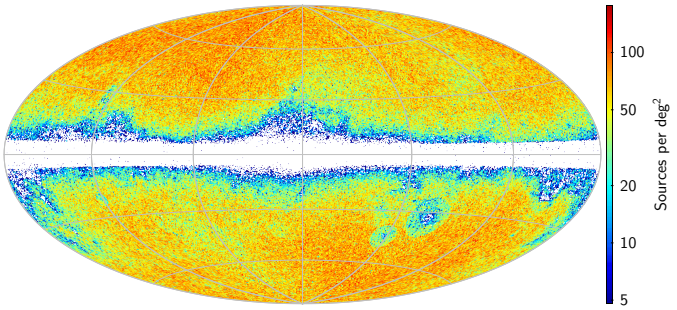

Fig. 23. Distribution Galactic coordinates (Hammer–Aitoff projection) of the 1 897 754 sources from the astrometric selection (i.e. sources in the `qso_candidates` table with `astrometric_selection_flag` set). The plot shows the density of sources per square degree computed from the source counts per pixel at HEALPix level 7 (pixel size $\simeq 0.21\,\mathrm{deg}^2$)

dependently of their classification status (Creevey et al. 2022). There are 255 948 sources in the `qso_candidates` table and 7069 sources in the `galaxy_candidates` table that have effective temperatures derived by the CU8 GSP-Phot module (Fousneau et al. 2022). Checking a variety of metrics such as magnitude, sky distribution, and effective temperature itself, there is nothing apparently peculiar with these sources. Their presence is an inevitable consequence of the known stellar contamination. It is also important to remember that DSC, which is the single largest contributor to these integrated tables, did not filter out sources simply because they were bright (only DSC-Allosmod classifies sources with $G < 14.5$ mag to be stars).

There are also 4027 sources with valid radial velocities in the `qso_candidates` table, and 160 in the `galaxy_candidates` table. Considering that the extragalactic tables are mostly populated with faint sources, these small numbers are essentially due to the intrinsic magnitude limit of sources for which radial velocities could be derived in *Gaia* DR3 (Katz et al. 2022). Those featuring valid radial velocities have magnitudes that are usually incompatible with extragalactic sources, so it is fair to assume that they are stars.

5.4. Astrometric selection

Additional insight into the classified sources can be gained by analysing their astrometric parameters. As has been demonstrated in Gaia Collaboration & Klioner et al. (2022) and Gaia Collaboration et al. (2021), astrometry can be used to improve the purity of a sample of quasar candidates. It is clear, however, that this can only be achieved at the cost of reducing the completeness.

The procedure here is similar to that used in the construction of *Gaia*-CRF3, namely a two-step astrometric filtering of a sample of candidates (Gaia Collaboration & Klioner et al. 2022). In the case of *Gaia*-CRF3, the sample was obtained by cross-matching the *Gaia* EDR3 catalogue with several external quasar catalogues. In the present study, each of the *Gaia* classifiers contributing to the `qso_candidates` table is considered as an additional catalogue, and the same procedure is applied to all the external and *Gaia*-own selections of quasar candidates.

The first step of the astrometric filtering is to select individual sources that have high-quality astrometric solutions in *Gaia* EDR3 and statistically insignificant parallaxes and proper motions (see Gaia Collaboration & Klioner et al. 2022, Sect. 2.1 for the exact mathematical formulations). This step alone is insufficient to find genuine quasars (or extragalactic objects), as about 214 million sources in *Gaia* EDR3, dubbed 'confusion sources', satisfy these astrometric criteria. These are mostly stars of our Galaxy (Gaia Collaboration & Klioner et al. 2022, appendix C). At least at this stage of the *Gaia* project, astrometry cannot be used as an independent quasar classifier, although this may change in the future (see e.g. Heintz et al. 2015, 2018).

A second step of filtering is therefore needed. In this step, only those samples of sources are retained that show near-Gaussian distributions in the uncertainty-normalized parallaxes and proper motions. Since extragalactic sources are faint, the typical uncertainties of their astrometric parameters in *Gaia* DR3 are about two orders of magnitude larger than either the known level of systematic errors in *Gaia* DR3 (Lindegren et al. 2021a) or the known physical systematic effects (Gaia Collaboration et al. 2021). Bearing in mind that the true parallaxes and proper motions of genuine extragalactic sources should be zero, one expects Gaussian distributions of the normalized parameters. This requirement had proven to be very useful to distinguish genuine quasars from the confusion sources.

Both steps of the astrometric filtering obviously reject some genuine quasars that have considerable measured, but spurious, proper motions due to time-varying source structure (see Sect. 2). A prominent example here is 3C273, which is not part of *Gaia*-CRF3 for this reason. The samples considered at the second step of the astrometric filtering can come from a particular external catalogue or from one of the *Gaia* classifiers, but could also be selections according to various criteria (e.g. avoiding the crowded areas on the sky) or intersections of such selections (e.g. sources that were found to be quasars by two classifiers).

An additional characteristic of a sample of genuine extragalactic objects is that its sky distribution should not show overdensities in known stellar structures in our Galaxy and its environments, such as clusters, although it could still be influenced by such structures, for example variable Galactic extinction. This can also be used to help decide whether a particular sample of sources should be retained.

Using this two-step selection procedure we have identified a set of 1 897 754 quasar candidates, which we refer to as the 'astrometric selection'. They are indicated by the `astrometric_selection_flag` in the `qso_candidates` table. The purity of this sample is difficult to estimate, but we believe it to be 98% or perhaps better. The vast majority of these

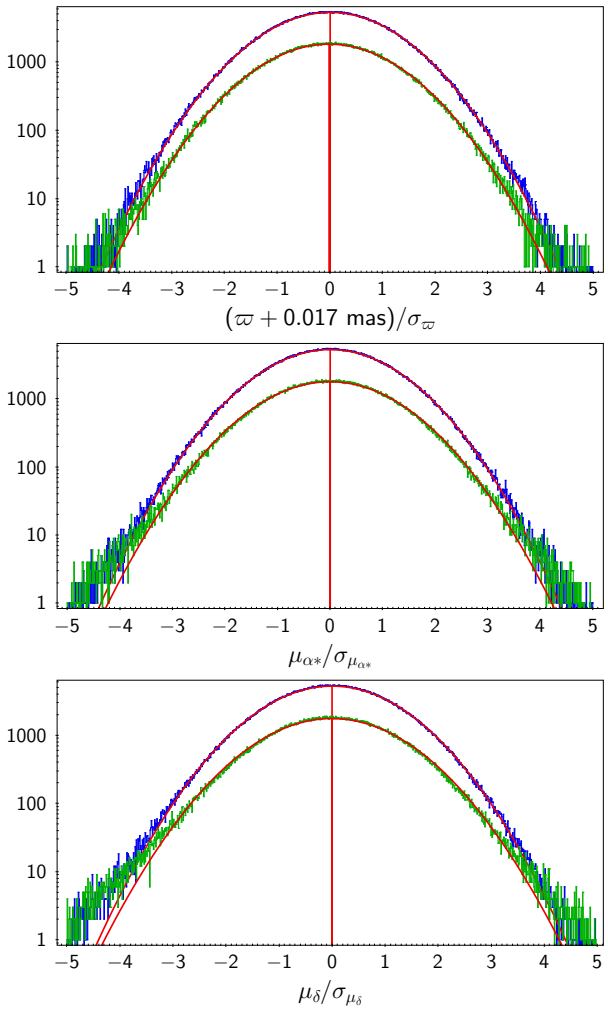

Fig. 24. Distributions of the normalized parallaxes and proper motion components for the sources in the astrometric selection with 5p (blue) and 6p (green) solutions. The red curves show the corresponding best-fit Gaussian distributions. The global parallax zero point of −0.017 mas of *Gaia* DR3 is taken into account (Lindegren et al. 2021b,a). The standard deviations of the best-fit Gaussian distributions for the sources with 5p (6p) solutions are 1.048 (1.068), 1.054 (1.092) and 1.063 (1.109) for the parallaxes, and proper motions in right ascension and declination, respectively. As usual in *Gaia*, the asterisk in $\alpha*$ in the middle panel indicates the implicit factor $\cos\delta$, that is $\mu_{\alpha*} = \dot{\alpha}\cos\delta$.

sources were identified as quasars by at least two independent external catalogues and/or *Gaia* classifiers. The density distribution of these sources on the sky is shown on Fig. 23. This set contains 1 406 729 sources (74%) with 5p astrometric solution and 491 025 sources (26%) with 6p solutions in *Gaia* DR3. The avoidance zone in the Galactic plane as well as the lower density of sources around the LMC and SMC result from the difficulty in reliably identifying quasars in those crowded areas. This concerns both the external catalogues and the *Gaia* classifiers.

Fig. 24 shows the distributions of the normalized parallaxes and proper motions of the astrometric selection. They are close to Gaussian, which suggests a reasonably low level of stellar contamination. The standard deviations of the best-fit Gaussian distributions range from 1.05 to 1.11 and indicate by how much the formal uncertainties of the corresponding astrometric parameters may be underestimated in *Gaia* DR3.

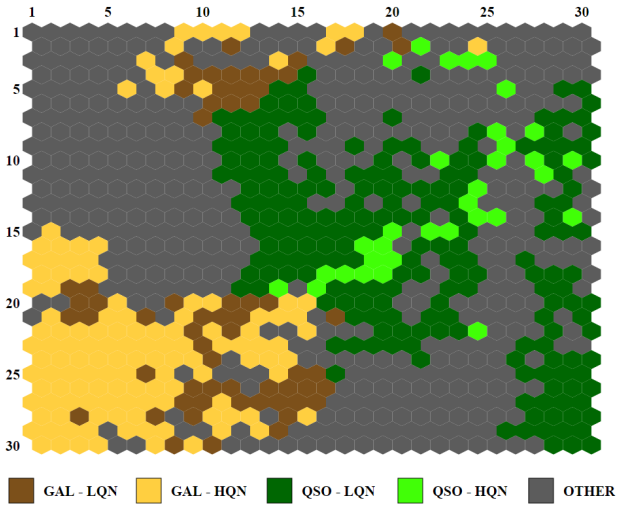

Fig. 25. OA class labels for extragalactic objects. HQN = high quality neuron (quality 0–3), LQN = low quality neuron (quality 4–6).

We attempted a similar astrometric selection for the `galaxy_candidates` table. However, since most of the galaxies have only two-parameter astrometry in *Gaia* DR3, and as more problems in *Gaia* astrometry can be expected for extended sources, the astrometric selection for the galaxy table turned out to be less useful, so we decided not to publish it. Nonetheless, this analysis did reveal the properties of the population of sources in the astrometric selection that were classified as both quasars and galaxies: the astrometric selection from the `qso_candidates` table contains 54 892 sources that are also present in the `galaxy_candidates` table (cf. overall overlap of these tables of 174 146 sources). 99% of those sources have 6p astrometric solutions. The normalized parallaxes and proper motions of this set of sources also have near-Gaussian distributions, but with standard deviations of 1.13–1.25, which is about 10% larger than for the astrometric selection as a whole. This set of sources is probably dominated by AGN for which source structure (i.e. the host galaxy) notably affects the astrometric solution. Indeed, a host galaxy was detected by *Gaia* for 23 805 of these sources (43%). Similar statistics of the normalized astrometric parameters can also be found for the set of sources in the astrometric selection for which a host galaxy was detected by *Gaia* (see Sect. 4.3), which contains 51 586 sources.

Thus we encounter the problem in the optical that is well known in radio astrometry (e.g. Charlot et al. 2020), namely the influence of the source structure on the quality of the astrometry. This topic will need a special attention in the future *Gaia* data releases.

5.5. Analysis of objects with lower probability classifications

The unsupervised algorithm OA was used to analyse the sources with lower DSC class probabilities (Sect. 2.4). Here we focus on those neurons that were assigned to an extragalactic class label (`QSO` or `GAL`). These are shown in Fig. 25 for two different subsets: high quality neurons (HQN), that represent quality categories 0 to 3, and low quality neurons (LQN), that represent categories 4 to 6. We further limit our analyses to those sources that appear in the integrated tables. Figure 26 shows the number of neurons and objects assigned to each quality category. Ap-

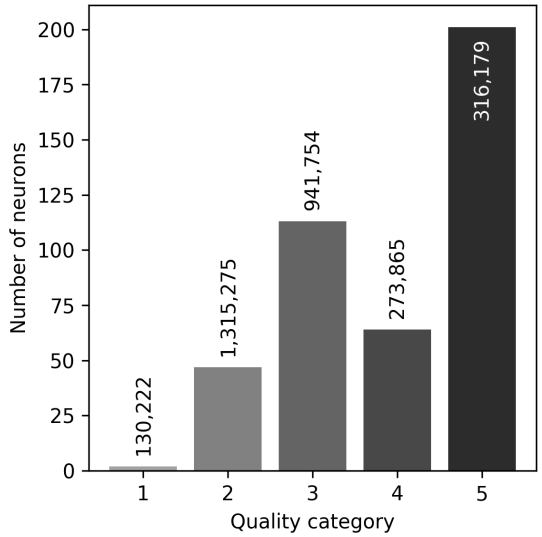

Fig. 26. Distribution of OA neurons labelled as extragalactic for each quality category. The number on each bar gives the number of sources in the `qso_candidates` or `galaxy_candidates` tables. No extragalactic objects appear in any of the best (0) or worst (6) quality neurons.

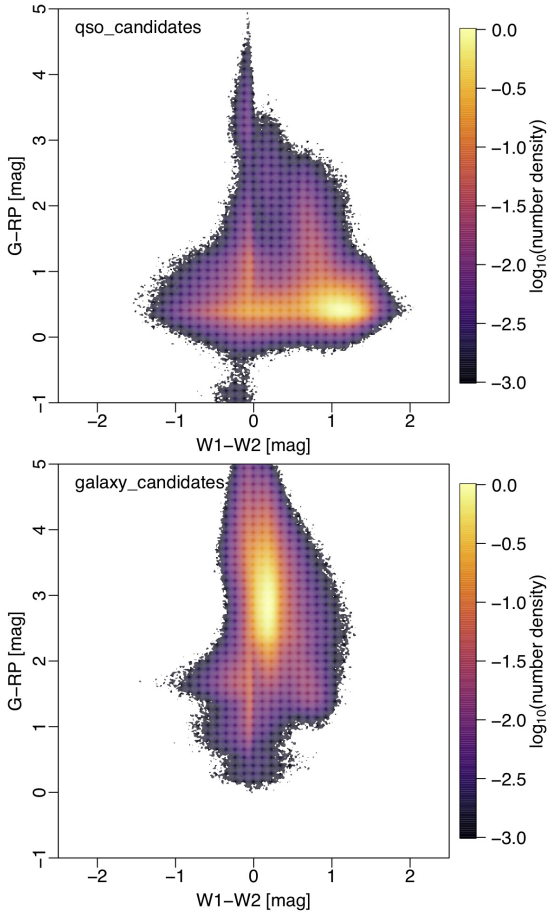

Fig. 27. *Gaia*-catWISE colour–colour diagrams. Top: All sources in the `qso_candidates` table. Bottom: All sources in the `galaxy_candidates` table. In both cases regions around the LMC and SMC have been excluded. The colour scale shows the density of sources on a log scale relative to the peak density (densities 1000 times lower than the peak are not shown).

proximately 80% of the sources are assigned to a high quality neuron.

All OA sources were processed by DSC, as well as QSOC or UGC depending on their DSC probabilities. Table 6 is a contingency table showing the fraction of objects in common between these classifiers and the various OA neurons. For this we use `classlabel_dsc`. QSOC and UGC do not classify sources; for this purpose we just look at the ones they provide redshifts for. Among the galaxies identified by DSC, 83% of them were also found to be galaxies by the OA module, of which 77% landed in a high quality neuron and 6% in a low quality one. The coincidence increases for UGC, with 89% of its galaxies found in a galaxy neuron, of which 83% have high quality. The coincidence for the quasars is substantially lower, around 35% for both DSC and QSOC, with no substantial difference between high and low quality neurons. We also see that a large fraction of those objects that were not classified as a quasar or galaxy by DSC, or that were not analysed by QSOC, are classified as galaxies by OA: 54% and 90%, respectively, where most of them belong to a high quality neuron.

OA processes sources that tend to be faint with noisy BP/RP spectra, some of which OA had to modify (e.g. remove negative fluxes) so that it could process them. Table 6 suggests that the OA classification complements the results from the other modules. OA coincides with DSC and UGC when identifying galaxies in particular, and identifies objects rejected by those modules that may be real galaxies. OA could also potentially help to identify extragalactic candidates that are not in the `qso_candidates` or `galaxy_candidates` tables.

6. External comparison

6.1. WISE and proper motions

To investigate the infrared colours of the sources in the integrated tables, we cross-matched them to the catWISE2020 catalogue (Marocco et al. 2021, including the 2021 catalogue updates) using a 1″matching radius. We found 4.31 million sources (65%) matches in the `qso_candidates` table, and 4.59 million (95%) in the `galaxy_candidates` table. Excluding the regions around the LMC and SMC (defined in appendix B) left 2.99 million matches in the `qso_candidates` table and 4.46 million in the `galaxy_candidates` table. Figure 27 shows the distribution of these sources in a *Gaia*-catWISE colour–colour diagram. We see that most galaxy candidates have W1−W2 colours between 0.0 and 0.5 mag. This agrees with the range identified by Stern et al. (2012) for galaxies without an active nucleus and redshifts below 0.6. The quasar candidates, in contrast, show two overdensities in the catWISE colour. We explore this further by looking at the quasar candidates in the proper motion space, as shown in Fig. 28. The upper panel is for all sources with 5p or 6p solutions (2.87 million sources). We see that the bluer clump at around W1−W2 ≃ 0 mag shows the full range of proper motions. Recall that non-zero proper motions of true quasars are spurious, either due to noise or to time-variable source structure. Nonetheless, the larger proper motions in the bluer clump compared to the redder clump is indicative of contamination by stars (and some galaxies), and the W1−W2 colour would seem to confirm that. Indeed, the lower panel of Fig. 28 is for the purer subset defined by `classlabel_dsc_joint = quasar` (0.50 million sources),

Table 6. Contingency table for OA classifications. Each entry gives the percentage of objects classified by DSC (using `classlabel_dsc`), or processed by QSOC or UGC, that are assigned to OA high-quality neurons (HQN) or low-quality neurons (LQN), for sources in the `qso_candidates` and `galaxy_candidates` tables.

		OA						
		QSO		GAL		other		
		HQN	LQN	HQN	LQN	HQN	LQN	Total
DSC	quasar	20%	15%	1%	1%	45%	18%	2 158 916
	galaxy	7%	2%	77%	6%	3%	5%	851 127
	other	7%	8%	53%	1%	24%	7%	1 993 592
QSOC	quasar	19%	15%	2%	1%	45%	17%	3 069 458
	other	3%	1%	87%	3%	4%	2%	1 934 177
UGC	galaxy	3%	1%	83%	6%	2%	5%	199 093
	other	13%	10%	33%	2%	30%	12%	4 804 542

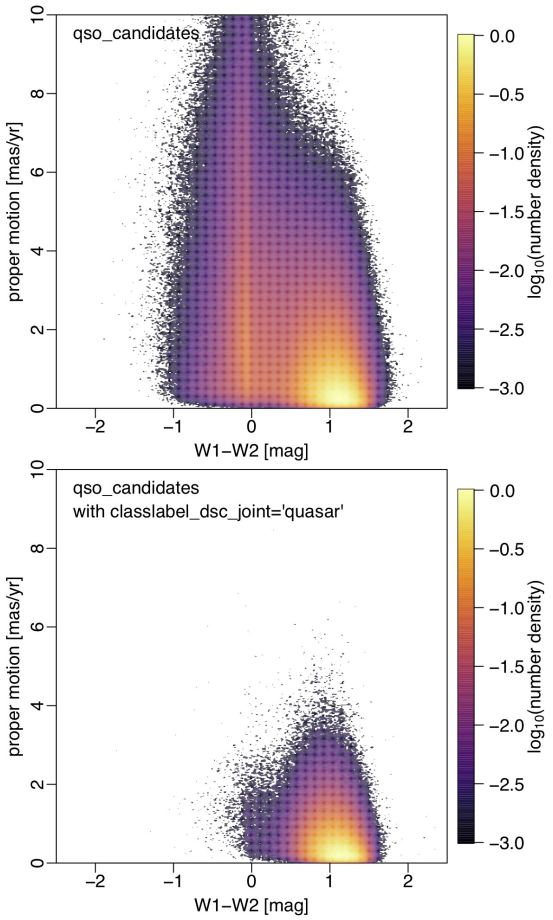

Fig. 28. *Gaia* proper motion vs catWISE W1−W2 colour for all sources in the `qso_candidates` table (top) and the subset with `classlabel_dsc_joint = quasar` (bottom). Regions around the LMC and SMC are excluded. The colour scale shows the density of sources on a log scale relative to the peak density (densities 1000 times lower than the peak are not shown).

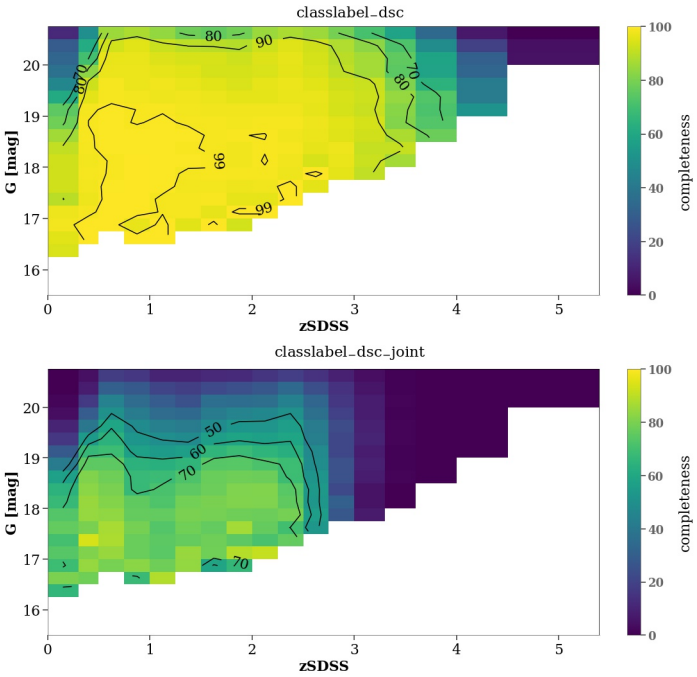

Fig. 29. Completeness of `classlabel_dsc = quasar` (top) and `classlabel_dsc_joint = quasar` (bottom) with respect to the SDSS-DR14Q-*Gaia* DR3 cross-match as a function of G and SDSS redshift. Empty bins (white) have fewer than 25 sources.

and this retains just the redder sources with proper motions that are more consistent with zero (plus noise).

6.2. Quasars

Here we look in more detail at the properties of known quasars in *Gaia*. For this purpose we cross-matched quasars from SDSS-DR14 (Pâris et al. 2018) that have a visually confirmed redshifts (`source_z = VI` or `source_z = DR7Q`) to all *Gaia* sources (those in `gaia_source`) using a 1″ matching radius. Such a match is nominally identical to the quasars selected for training DSC (see Delchambre et al. 2022). However, we further limited this set to those with complete data in all photometric bands, at least five observations in BP and RP, complete astrometry (i.e. 5p or 6p solutions), and with $G < 20.75$ mag. This gave 232 794 sources covering the redshift range 0.038–5.305.

A quasar in SDSS-DR14 is defined according to spectroscopic criteria. Specifically, they are sources with: (a) either at least one broad emission line with a full width at half maximum larger than 500 km s^{-1}, or interesting or complex absorption features; and (b) sufficiently large intrinsic luminosity ($M_i[z = 2] < -20.5$). Since only one broad emission line is required, some objects may otherwise be classified as type 2 AGNs (those with predominantly narrow emission lines). The second part of the first condition aims to include Broad Absorption Line (BAL) quasars. This definition is free of morphological criteria.

The sample defined above is similar to the superset from which the DSC-Allosmod training set was drawn. However,

DSC did not force classifications on them, so we can use it to assess DSC's completeness as a function of magnitude and redshift (further assessments can be found in Bailer-Jones 2021). This is shown in Fig. 29, using the two class labels from DSC. The dependence on redshift is expected because of its weak correlation with $G_{\rm BP}$−$G_{\rm RP}$, which increases the confusion with stars at high redshifts and with galaxies at low redshifts. Nonetheless, the completeness is above 80% for redshifts between 0.3 and 3.6 and $G \leq 20.25$ mag. The lower completeness at fainter magnitudes is also expected, because lower quality data are more likely to be classified by DSC as the majority class of stars, according to the global prior (Sect. 2.1), especially for the more conservative `classlabel_dsc_joint` label. This also explains why the overall completeness is much lower for this label, although it is still above 60% from $z = 0$ to $z = 2.5$ for $G \leq 19.25$ mag. The overall completeness of `classlabel_dsc` is 215 721/232 794=93% and of `classlabel_dsc_joint` is 97 995/232 794=42%. However, given the non-uniform selection function of SDSS for obtaining spectra, we should be careful not to over-interpret this specific assessment of the DSC's completeness.

The $G_{\rm BP} - G_{\rm RP}$ vs redshift relation for the sources classified as quasars by `classlabel_dsc` and `classlabel_dsc_joint` is shown in Fig. 30 (top and bottom panels). The pattern of undulations is expected, and is due to the quasar emission lines moving across the bands with redshift. The tail to the red for $z > 3.5$ corresponds to the Lyα forest entering and then filling the BP band. We see that some low redshift objects classified by SDSS as quasars are classified as galaxies by DSC. These quasars likely have a higher contribution of the host galaxy to the total emission, making them redder. The quasars with `classlabel_dsc_joint = quasar` follow neatly the bluest part of the colour-z relation, avoiding the regions with $G_{\rm BP}$−$G_{\rm RP}$ above the median (compare top and bottom panels). This result and Fig. 29 indicate that this class selects mainly bright quasars, and from these only the bluest ones. `classlabel_dsc = quasar` complements the `classlabel_dsc_joint = quasar` class by covering the quasars that are redder due to their intrinsic emission, Galactic extinction, or local absorption (as in BALs, for example). The middle panel of Fig. 30 shows that the incompleteness of `classlabel_dsc` at high-z is due to the misclassification of many of these quasars as stars (see also Fig. 29). Similarly, this plot shows that the quasars in the envelope of the reddest colours over the range $z = 0.5$–3.5 are also classified as stars.

Figure 31 shows the colour–colour diagram for the sample colour-coded by SDSS redshift (top panel) and by *Gaia* `phot_bp_rp_excess_factor` (bottom panel). In the upper panel we see a clear trend of colour with redshift, with low redshifts located in the upper left and redshift increasing as we descend, but with the highest redshifts on the far right. The bottom panel shows that the region of low-z sources corresponds to sources with larger `phot_bp_rp_excess_factor`, which indicates that the combined BP and RP bands contain more flux than the G band. This region overlaps with the location of the galaxies (see plots of the DSC source densities in Delchambre et al. 2022). This suggests that the excess in $G_{\rm BP}$ and $G_{\rm RP}$ with respect to G is due to the wider photometric windows of $G_{\rm BP}$ and $G_{\rm RP}$ compared to G, which allows detection of the quasar's host galaxy emission over a wider region than in the G band. This, together with the red $G_{\rm BP} - G_{\rm RP}$ colours, indicates a shift from quasars dominated by the nucleus to quasars with an important contribution of the host galaxy. This transition can be also appreciated in the colour–colour diagram of the purer sub-

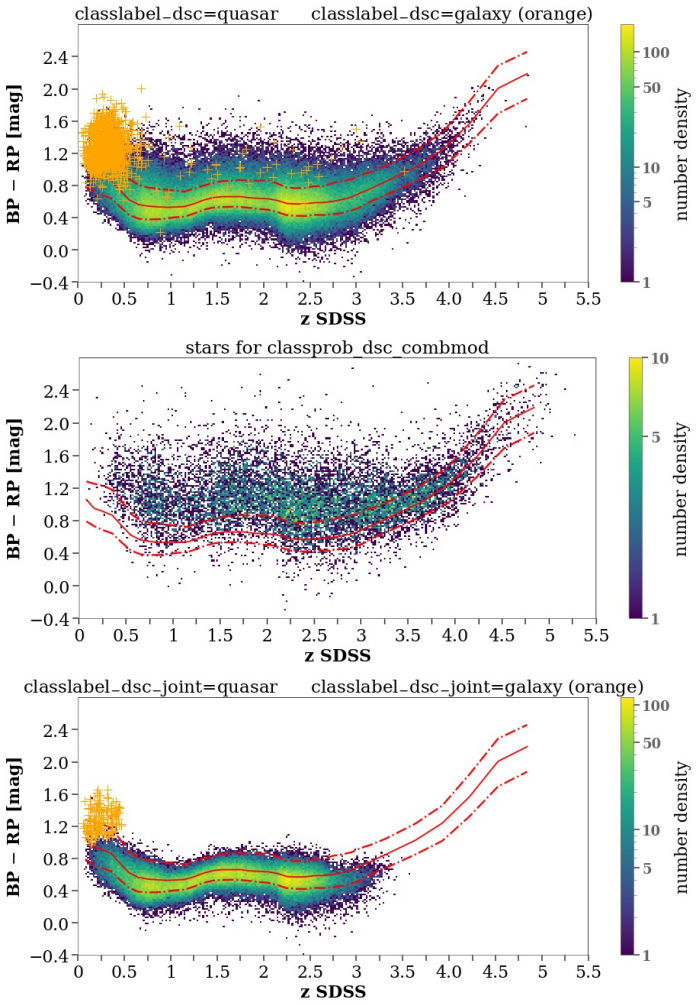

Fig. 30. $G_{\rm BP}$ − $G_{\rm RP}$ vs redshift for the subsets `classlabel_dsc` (top) and `classlabel_dsc_joint` (bottom) of the SDSS-DR14Q-*Gaia* DR3 cross-match. The orange points are `classlabel_dsc = galaxy` (top) and `classlabel_dsc_joint = galaxy` (bottom). The middle panel shows the quasars with `classprob_dsc_combmod`> 0.5 for any of the three stellar classes used in DSC-Combmod. The red curves show the median $G_{\rm BP}$−$G_{\rm RP}$ and the 16% and 84% quantiles for all sources in the cross-match, regardless of the DSC class (so are the same in all panels).

samples from the `qso_candidates` and `galaxy_candidates` tables (Fig. 37), followed by the transition to the general galaxy population.

6.3. Galaxies

We compared the content of the `galaxy_candidates` table with the spectral classes in SDSS DR16. We cross-matched the catalogues with a 1 arcsec radius and removed duplicated sources or ones where `zWarning` was not equal to zero. Of the 4 842 342 sources in the `galaxy_candidates` table, we found 534 154 matches in SDSS. 98.0% of these are classified by SDSS as GALAXY, 1.6% as QSO, and 0.4% as STAR. Table 7 shows these percentages for each module that contributed to the `galaxy_candidates` table. For example, 48 460 sources have `classlabel_dsc_joint = galaxy` in the matched set, and 95.1% of these have SDSS class GALAXY. This percentage is a measure of the purity of `classlabel_dsc_joint` but

Table 7. SDSS spectral classes for objects in the `galaxy_candidates` table. The first row gives the number of sources found in SDSS for the whole table (first column) and for each module (subsequent columns). The following rows give the percentage of sources of each SDSS class among these. The GALAXY row can be thought of as a measure of the purity against sources with SDSS spectral classifications, which by design is dominated by extragalactic sources, so is an overestimate of the purity of a sample selected at random from *Gaia*.

SDSS	`galaxy_candidates`	`classlabel_dsc`	`classlabel_dsc_joint`	Vari-Classification	OA	UGC	Surface brightness	
TOTAL	534 154	477 939	48 460	360 217	105 296	248 196	96 918	#
GALAXY	98.0	97.9	95.1	99.7	95.6	97.9	99.7	%
QSO	1.6	1.7	4.9	0.3	3.9	2.0	0.3	%
STAR	0.4	0.4	0.1	0.0	0.5	0.0	0.0	%

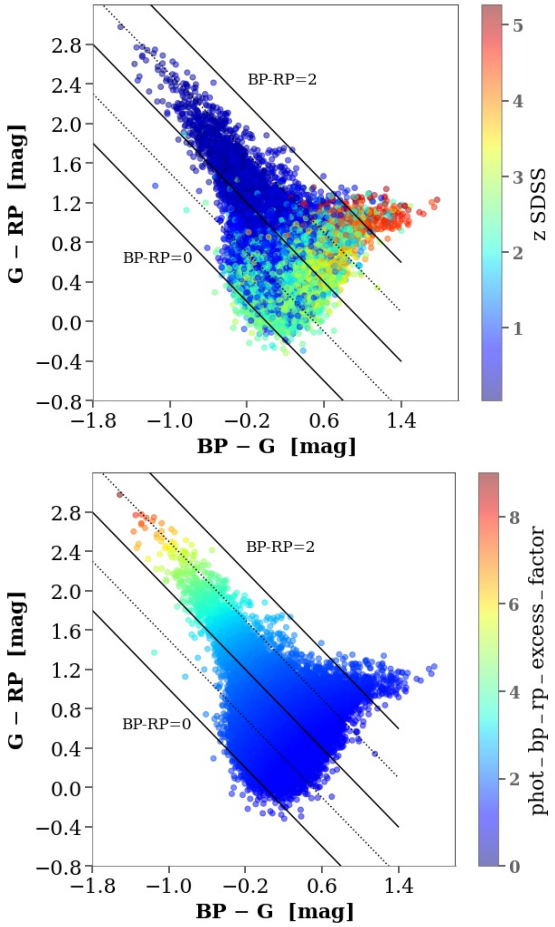

Fig. 31. Colour–colour diagram for the set in Fig. 30 colour-coded by redshift (top) and *Gaia* `phot_bp_rp_excess_factor` (bottom). The straight lines correspond to fixed $G_{\rm BP}-G_{\rm RP}$ colour, with values 2, 1.5, 1, 0.5 and 0.

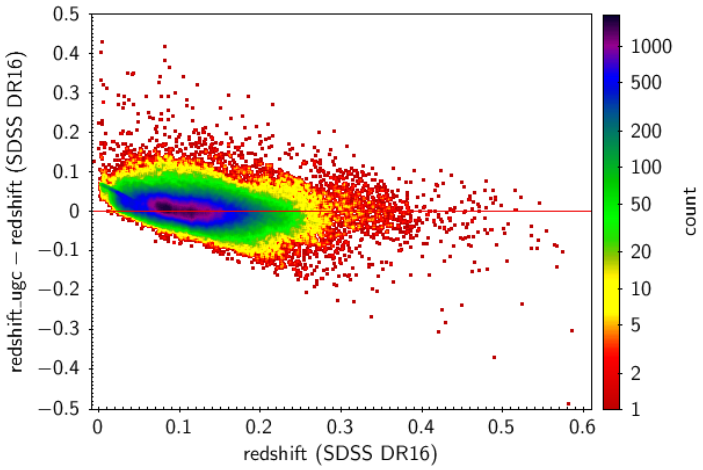

Fig. 32. Difference between `redshift_ugc` and SDSS DR16 redshift, as a function of the latter, for the 248 356 sources in common with SDSS spectral class GALAXY.

only against those sources that have SDSS spectral classifications. It is considerably higher than the purity of this DSC class label reported in Sect. 2.1 (and detailed in Delchambre et al. 2022), even though this purity estimate was also based on SDSS. The reason is that this earlier purity estimate was computed for a set of sources selected at random from *Gaia*: It includes the significant contamination from non-galaxies, which outnumber galaxies by a factor of about one thousand in *Gaia*. The higher figure reported in Table 7, in contrast, is just for those sources that have SDSS spectral classifications. By design of SDSS this includes proportionally very few stars and so very few potential contaminants. The numbers in Table 7 are a therefore a significant overestimate of the true purity and should be treated with caution.

Of the 1 367 153 sources in the `galaxy_candidates` table that have redshifts provided by UGC, 248 356 match to sources in the SDSS-DR16 `specObj` table that are classified as GALAXY. The small discrepancy with the number in Table 7 is due to slightly different cross-match criteria. Figure 32 shows the difference between `redshift_ugc` and the SDSS-DR16 redshift as a function of the latter. The average of this difference is 0.06 with a standard deviation of 0.054 (which reduces to 0.029 when the 67 sources with redshifts above 0.6 are excluded). Generally the agreement is good, although UGC seems to systematically overestimate very low redshifts.

7. Composite quasar spectrum

Composite quasar spectra have many uses. First and foremost they are used as a reference in cross-correlations of individual spectra in order to classify these and determine their redshifts. Composite spectra are also used to identify faint spectral features that would otherwise be undetectable, to calibrate absolute magnitudes through the *k*-correction, as well as to construct colour–colour relations for identifying and characterising quasars based on photometry. Here we construct composite spectra to unveil the capability offered by the *Gaia* BP/RP spectrophotometers to characterise quasars.

In order to build composite BP/RP spectra we use the quasar sample described in Sect. 4.2.1, which is based on the Milliquas 7.2 quasar catalogue of Flesch (2021). Our sample comprises 42 944 sources for which we use the Milliquas redshifts. We rely

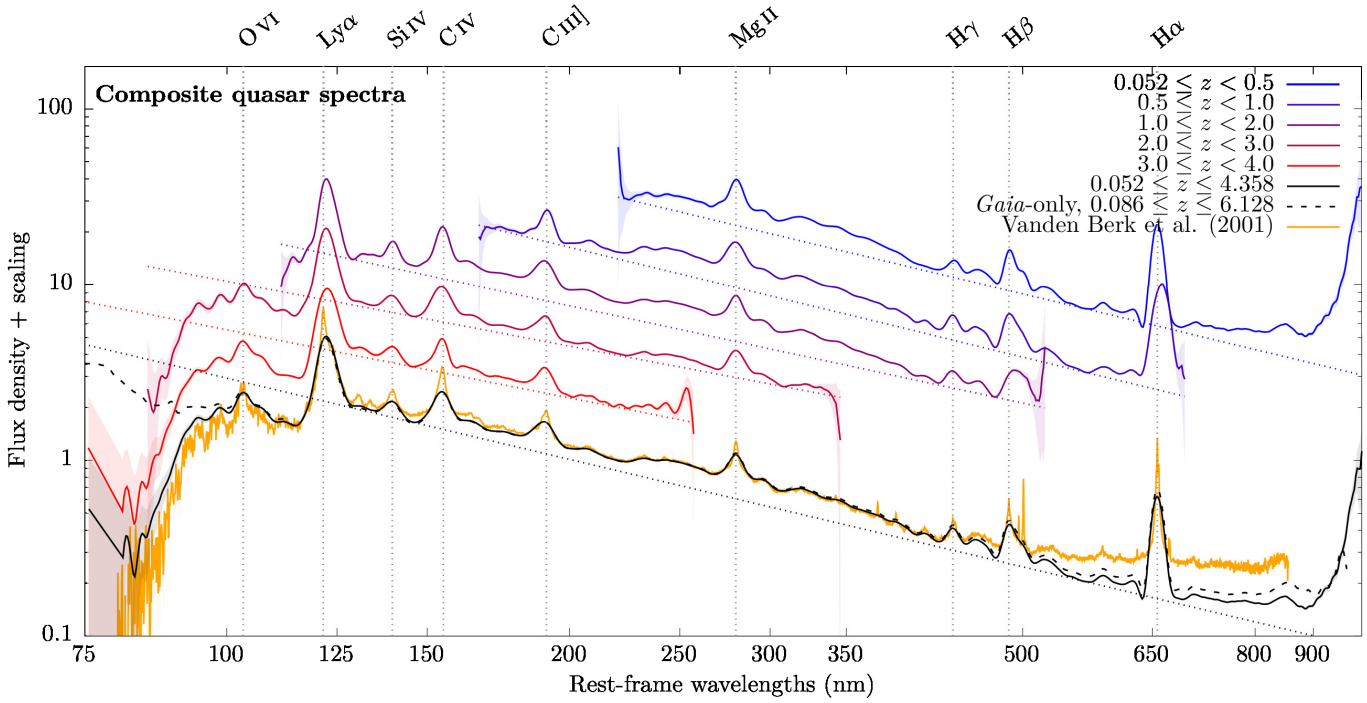

Fig. 33. Composite quasar spectra. The thick solid lines show composites made from the 42 944 BP/RP spectra with spectroscopically-confirmed redshifts from the Milliquas 7.2 quasar catalogue of (Flesch 2021, i.e. `type = Q` in Milliquas). The different colours are for different redshift ranges. The thick dotted black line shows the composite made from 111 563 BP/RP spectra with reliable QSOC redshift estimates (identified using the query given in appendix B.2). The diagonal dotted line under each spectrum shows the quasar continuum, as described in Sect. 7 and defined in Table 9. Vertical dotted lines indicate common quasar emission lines. For comparison purposes we also show the median SDSS composite spectrum of Vanden Berk et al. (2001) (orange line). The flux densities are tabulated in Table 8.

Table 8. Composite quasar spectra shown in Figure 33. The columns are the arbitrarily scaled flux densities, F, and associated uncertainties, $F_{\rm err}$, computed at rest-frame wavelengths λ. The values below are for the Milliquas composite covering the redshift range 0.052 to 4.358. A full electronic version of this table is available with the online version of this article, along with the quasar composite spectra for the narrower redshift ranges shown in Figure 33 as well as the *Gaia*-only composite.

λ (nm)	F	$F_{\rm err}$
75.669	0.00296	0.00292
81.074	0.00157	0.00133
81.318	0.00189	0.00128
81.562	0.00208	0.00124
81.807	0.00207	0.00119
...	...	...
980.799	0.00457	0.00095
983.746	0.00504	0.00142
986.701	0.00499	0.00188
989.666	0.00500	0.00222
992.639	0.00630	0.00297

on an external catalogue instead of the *Gaia* DR3 internal classifications coupled with QSOC redshift determinations, because the various instruments that contributed to the Milliquas catalogue have better spectral resolutions and sensitivities than *Gaia*. We nonetheless also compute a composite *Gaia*-only spectrum based on 111 563 sources coming from the *Gaia* classifications and input lists together with the redshifts from QSOC. The exact sample used for this is defined in appendix B.2. The method by which we compute composite spectra is described in appendix C. The method relies on a single parameter, the logarithmic wavelength sampling of the composite spectra, chosen here to be $\log S = 0.003$ (i.e. $S \approx 1.003$) as a compromise between S/N and execution time. This sampling also applies to transformation matrices – $\mathbf{M_i}$ in Eq. C.1 – that cover the observed wavelength region 309.5–1100.5 nm.

The resulting composite spectrum inferred using the Milliquas sample is shown in Fig. 33 as the solid black line and the values listed in 8. It covers a rest-frame wavelength range from 75.67 nm to 992.64 nm. The *Gaia*-only composite spectrum is shown as the dotted line and covers the wavelength range 47.08 nm to 962.83 nm. The composite spectra are trimmed in order to discard wavelength regions having flux density S/N less than one. After a multiplicative re-scaling of the flux densities so that their continua align, we found absolute differences between these two composites – relative to the Lyα flux density in the Milliquas composite spectrum – of less than 4% over the rest-frame wavelength region 100–900 nm, but up to 60% for regions bluewards of this range. The cause of these larger deviations is either contamination by sources with erroneous redshift estimates, as a consequence of the low purity of QSOC in this very high redshift region (see Delchambre et al. 2022), or border effects in the externally calibrated BP/RP spectra that we describe below. Figure 33 also shows, for comparison, a median composite spectrum from SDSS (Vanden Berk et al. 2001) that covers a rest-frame wavelength range similar to that of our full composite spectrum.

Figure 34 shows the redshift and magnitude distributions of the sources used to build the Milliquas-based composite spectrum. While the redshift distribution is as expected, with imprints of the selection and observational bias of each survey composing the Milliquas catalogue, the sharp drop at $G > 19$ mag is due to the filtering of the set of BP/RP spectra published in *Gaia* DR3. The 17 sources with $G \geq 19$ mag are present because: 11 are

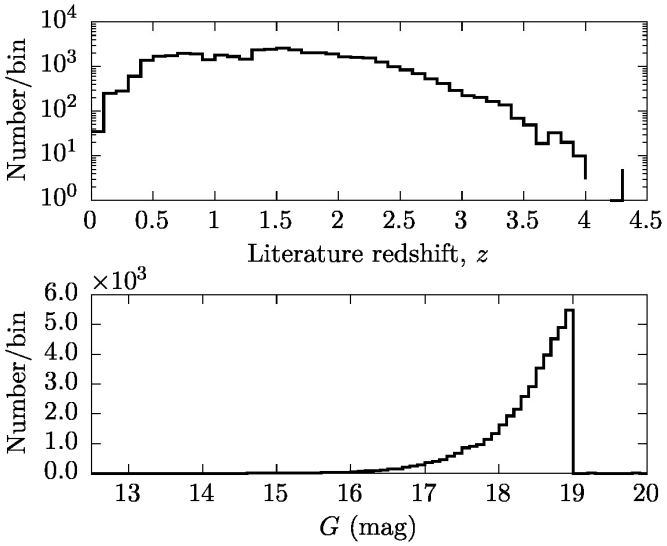

Fig. 34. Distribution of the literature redshift from the Milliquas 7.2 catalogue (Flesch 2021) and *Gaia* G-band magnitudes for the 42 944 sources used in the computation of the composite quasar spectra in Fig. 33.

associated with a best-matching node of the OA module (see Sect. 2.4), four are used as external calibrators by CU5, and two are white dwarf candidates used in Bellazzini at al. (2022).

In addition to the full composite spectrum computed over the whole redshift range from 0.052 to 4.358, we also computed composite spectra over several narrower redshift ranges, chosen such that their logarithmic rest-frame wavelength coverage are approximately of equal size: $0.052 \leq z < 0.5$, $0.5 \leq z < 1$, $1 \leq z < 2$, $2 \leq z < 3$ and $3 \leq z \leq 4.358$. Composite spectra associated with each redshift range are shown in Fig. 33. The number of sources used in each redshift range, as well as the resulting reduced chi-square derived from Eq. C.1, are provided in Table 9 along with the frequency (ν) continuum slope, α_ν and some information on the strongest observed emission lines: the rest-frame wavelengths, the relative flux density at their peak compared to either Lyα or Mg II, and the full width at half maximum (FWHM). Unsurprisingly, all reduced chi-squares are larger than unity: Each composite spectrum evidently does not completely model the intrinsic variance seen in the observations. However, the moderate values that are observed are indicative of reasonable fits to the observations, which is corroborated by visual inspection and by the good agreement with the median composite spectrum of Vanden Berk et al. (2001). The full composite spectrum (with $\chi^2_\nu = 6.6$) also explains 99.2% of the observed variance in the entire set of spectra, although most of this variance comes from the spectral continuum. The maximum S/N ranges from 224 per $\log S$ interval in the composite spectrum with $3 \leq z < 4$, to 645 per $\log S$ interval in the composite spectrum associated with the $0.5 \leq z < 1$ redshift range.

The highest S/N of 610 per $\log S$ interval of the full composite spectrum allows us to identify many more emission lines than are otherwise visible in the BP/RP spectra of individual sources. Whereas only the Lyα, C IV, C III], Mg II, Hβ, and Hα emission lines are commonly seen in the observed BP/RP spectra of quasars, a visual inspection of the full composite spectrum reveals 22 emission lines, which are listed in Table 10. Despite the low resolution of BP/RP spectra (Sect. 4.2), all common quasar emission lines were retrieved during this inspection procedure, in addition to some weak or rarely-seen emission lines. Emission lines from the wavelength region covering the Lyα forest are similarly recovered in an unambiguous way. We achieve good agreement with laboratory wavelength positions, with a maximum absolute difference of $|\Delta\lambda| = |\lambda_{\rm lab} - \lambda_{\rm obs}| = 0.951$ nm for the C III] emission line, where we consider only the nearest emission line in case these are blended. The apparent blueshift of the C III] emission line resides in its asymmetry, which is due to the presence of the Si III] λ 189.203 nm in its neighbourhood. The same rationale applies to the shift of the Lyα and Hβ emission lines ($\Delta\lambda = -0.683$ nm and -0.433 nm, respectively) due to the presence of the N V λ 124.014 nm and [O III] doublets respectively.

Our composite spectra are highly coherent with one another, with little variation depending on the redshift ranges that were used. After a multiplicative re-scaling of the flux densities to align the continua, we found differences relative to the Lyα flux density of less than 6% when compared to the full composite spectrum over several rest-frame wavelength regions (230–900 nm for the $0.052 \leq z < 0.5$ composite spectrum; 180–680 nm for the $0.5 \leq z < 1.0$ composite spectrum; 115–500 nm for the $1 \leq z < 2$ composite spectrum; 85–320 nm for the $2 \leq z < 3$ composite spectrum and 85–240 nm for the $3 \leq z < 4$ composite spectrum). Some border effects that were ignored in the previous comparison can still be seen in all composite spectra. These are due to the low S/N as well as systematic errors at the borders of the externally calibrated BP/RP spectra of quasars that we used to build the composite spectra. Such an effect is particularly noticeable in Fig. 33 redwards of 900 nm, where a sharp rise is visible that is also seen in about 90% of the individual spectra of quasars we used. This artefact is consequently taken as a genuine signal by our method.

The differences noticed when comparing the full composite spectrum to the SDSS one of Vanden Berk et al. (2001) redwards of the Hβ emission line are presumably due to different angular scales of the *Gaia* BP/RP footprint and the SDSS fibres, which are 2.1″(across scan direction; Carrasco et al. 2021) and 3″respectively. This results in the contamination of the composite spectrum by the light from the host galaxy being more suppressed in the *Gaia* observations.

Figure 33 and Table 9 reveal differences in the continuum slopes, α_ν, over the various redshift ranges. These are mostly a result of the different rest-frame wavelength coverage of each composite spectrum. Indeed, the presence of broad Fe multiplets and the Balmer continuum in the wavelength region 200–550 nm – the so-called 300 nm bump – complicates the continuum fitting in this range due to the lack of pure continuum. Consequently the values of α_ν decrease once we consider quasars with redshifts $z > 1$. The value we found on the full composite spectrum, $\alpha_\nu = -0.464 \pm 0.005$, agrees reasonably well with literature values. The median composite spectrum from Vanden Berk et al. (2001) shown in Fig. 33, for example, has $\alpha_\nu = -0.46$.

We inspected the measurements associated with the dominant quasar emission lines given in Table 10. We did not find any meaningful correlation between their values and the redshift range that was used for computing each composite spectrum.

8. How to select purer sub-samples

The `qso_candidates` and `galaxy_candidates` tables collate together results for most extragalactic candidates in *Gaia* DR3 from a number of modules, as described in Sect. 3.1. Overall these tables aim for high completeness, rather than high purity. We have done this intentionally to allow the user to select their

Table 9. Physical quantities derived from the composite spectra of quasars from the Milliquas catalogue shown in Fig. 33. The reduced chi-square, χ^2_ν, is obtained from Eq. C.1. The frequency continuum slope, α_ν, is computed from the wavelength continuum slope, $\alpha_\lambda = -\alpha_\nu - 2$, where the continuum is modelled through a power law of the form $C_\lambda \propto \lambda^{\alpha_\lambda}$. In Fig. 33, the continuum is plotted as the lowest line joining the two most widely separated points in the range 121–600 nm without crossing the spectrum in this range. The emission line location, $\lambda_{\rm obs}$, and maximal line flux density, F, are retrieved from the fit of a quadratic polynomial in the vicinity of the laboratory wavelength, $\lambda_{\rm lab}$, after a local continuum has been subtracted that was computed in the same way as for α_λ. The full-width at half maximum (FWHM) is retrieved from these continuum-subtracted emission lines using a linear interpolation of the spectral flux densities. All uncertainties are calculated, to first order, using the formal uncertainties on the composite spectra obtained from Eq. C.4.

Composite spectrum	Continuum slope, α_ν	Line+$\lambda_{\rm lab}$ (nm)	$\lambda_{\rm obs}$ (nm)	$F/F^a_{\rm ref}$	FWHM (nm)
$0.052 \le z < 0.5$	-0.4506 ± 0.1453	Mg IIλ279.875	280.064 ± 0.234	1.00000 ± 0.01036	8.308 ± 0.214
2 571 sources		H$\gamma\lambda$434.168	435.221 ± 0.588	0.13927 ± 0.00369	10.432 ± 0.586
$\chi^2_\nu = 20.8$		H$\beta\lambda$486.268	486.936 ± 0.081	0.46937 ± 0.00233	12.874 ± 0.131
		H$\alpha\lambda$656.461	656.240 ± 0.010	1.27766 ± 0.00095	17.971 ± 0.035
$0.5 \le z < 1$	-0.4291 ± 0.0570	C III]λ190.873	191.144 ± 0.134	1.80831 ± 0.03931	5.583 ± 0.262
8 810 sources		Mg IIλ279.875	279.644 ± 0.084	1.00000 ± 0.00238	10.613 ± 0.120
$\chi^2_\nu = 7.0$		H$\gamma\lambda$434.168	433.932 ± 0.048	0.25984 ± 0.00099	10.355 ± 0.123
		H$\beta\lambda$486.268	486.825 ± 0.042	0.53950 ± 0.00085	18.263 ± 0.082
		H$\alpha\lambda$656.461	661.947 ± 0.100	1.36875 ± 0.00735	23.063 ± 0.684
$1 \le z < 2$	-0.6115 ± 0.0227	Ly$\alpha\lambda$121.567	122.264 ± 0.029	1.00000 ± 0.00826	4.411 ± 0.097
20 755 sources		Si IVλ139.676	139.925 ± 0.051	0.14881 ± 0.00292	3.637 ± 0.111
$\chi^2_\nu = 4.9$		C IVλ154.906	154.636 ± 0.050	0.29494 ± 0.00154	4.477 ± 0.043
		C III]λ190.873	190.124 ± 0.093	0.13300 ± 0.00039	7.729 ± 0.068
		Mg IIλ279.875	280.056 ± 0.053	0.08408 ± 0.00026	8.026 ± 0.047
		H$\gamma\lambda$434.168	433.450 ± 0.101	0.01752 ± 0.00013	11.683 ± 0.292
		H$\beta\lambda$486.268	490.168 ± 5.268	0.03491 ± 0.00070	23.559 ± 2.499
$2 \le z < 3$	-0.7752 ± 0.0113	O VIλ103.383	103.589 ± 0.152	0.18338 ± 0.00686	3.279 ± 0.342
9 870 sources		Ly$\alpha\lambda$121.567	122.156 ± 0.022	1.00000 ± 0.00160	4.908 ± 0.024
$\chi^2_\nu = 5.8$		Si IVλ139.676	139.738 ± 0.086	0.11467 ± 0.00068	4.927 ± 0.129
		C IVλ154.906	154.465 ± 0.049	0.19586 ± 0.00041	5.854 ± 0.114
		C III]λ190.873	190.589 ± 0.047	0.11133 ± 0.00107	6.742 ± 0.164
		Mg IIλ279.875	279.941 ± 0.091	0.07435 ± 0.00019	9.471 ± 0.078
$3 \le z < 4$	-0.7155 ± 0.0548	O VIλ103.383	103.381 ± 0.112	0.21926 ± 0.00277	5.202 ± 0.935
929 sources		Ly$\alpha\lambda$121.567	122.398 ± 0.086	1.00000 ± 0.00489	5.591 ± 0.083
$\chi^2_\nu = 6.0$		Si IVλ139.676	139.980 ± 0.232	0.12360 ± 0.00895	4.588 ± 0.486
		C IVλ154.906	154.581 ± 0.151	0.24250 ± 0.00306	3.928 ± 0.107
		C III]λ190.873	190.161 ± 0.118	0.13617 ± 0.00097	6.488 ± 0.094
$0.052 \le z \le 4.358$	-0.4639 ± 0.0057	O VIλ103.383	103.500 ± 0.103	0.19285 ± 0.00584	3.784 ± 0.517
42 944 sources		Ly$\alpha\lambda$121.567	122.202 ± 0.037	1.00000 ± 0.00259	5.088 ± 0.036
$\chi^2_\nu = 6.6$		Si IVλ139.676	139.838 ± 0.141	0.11467 ± 0.00146	4.672 ± 0.139
		C IVλ154.906	154.542 ± 0.052	0.21615 ± 0.00132	5.383 ± 0.066
		C III]λ190.873	190.232 ± 0.056	0.11681 ± 0.00072	7.938 ± 0.097
		Mg IIλ279.875	280.006 ± 0.034	0.08202 ± 0.00025	8.891 ± 0.047
		H$\gamma\lambda$434.168	434.164 ± 0.059	0.02165 ± 0.00016	10.706 ± 0.129
		H$\beta\lambda$486.268	486.846 ± 0.109	0.04867 ± 0.00011	16.473 ± 0.179
		H$\alpha\lambda$656.461	656.257 ± 0.014	0.13924 ± 0.00011	18.026 ± 0.032

[a] If Lyα is covered by the composite spectrum, it is used as a reference flux density, $F_{\rm ref}$, otherwise Mg II is used.

own sub-samples using the quality indicators and class probabilities. Here we describe how to select purer sub-samples from these tables.

For the quasars, EO and CRF3 used input lists of quasars identified by other surveys, so their samples are believed to be quite pure, above 90%. From the EO sample we exclude those with close neighbours (`host_galaxy_flag = 6`). For DSC, the joint subset has a purity of 62%, increasing to 79% when the Galactic plane ($|b| < 11.54$ deg) is avoided (Delchambre et al. 2022). The Vari classifier results for AGN, which already exclude the Galactic plane, has been assessed to have a purity of over 90% (Rimoldini et al. 2022). We therefore recommend the query in Table 11 to select a purer sub-sample of quasars. This returns 1 942 825 sources, which is 29% of the original table. Using the same approach at the end of section 3.1, and assuming a 96% purity for the non-DSC modules, we estimate the overall purity of this sub-sample to be 95%. Of these, 1.7 million have published redshifts from QSOC.

We use similar criteria to define a purer galaxy sub-sample, except that here there is no contribution from CRF3. Again we take all of the sources from EO (provided by an input list), the purer subset of DSC (64% pure; 82% outside the Galactic plane), and all of Vari. This query, in Table 12, returns 2 891 132 sources, 60% of the original table. We estimate the purity of this sub-sample to be 94%. Of these, 1.1 million have published redshifts from UGC.

There are 14 471 sources in common between these two purer sub-samples, and their union contains 4.8 million sources.

The sky distributions for these purer sub-samples are shown in Fig. 35 and can be compared with those for the full tables in Fig. 5. We immediately see how the purer sub-samples have lower densities in the Galactic plane. There are still artefacts from the *Gaia* scanning law, which is an indication of the less than perfect purity. We also still see overdensities at the LMC and SMC. This comes mostly from DSC, because unlike the other modules DSC did not do any sky-position-dependent fil-

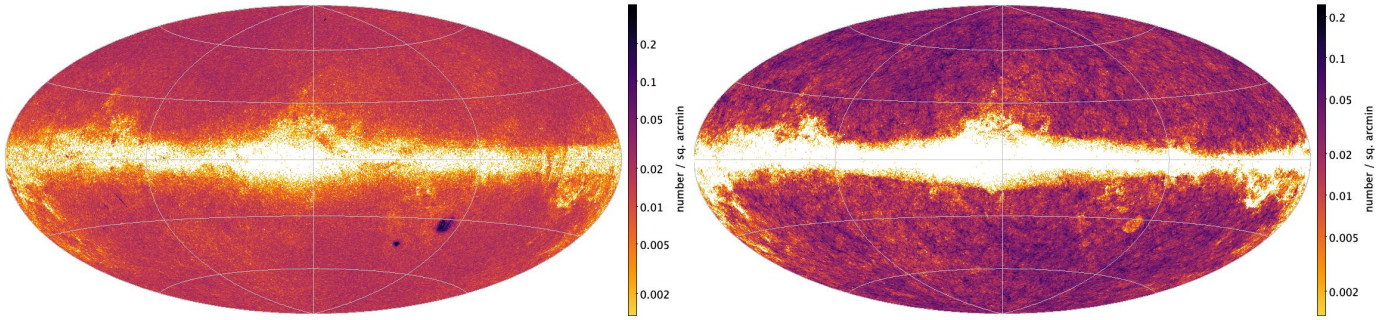

Fig. 35. Galactic sky distribution of all the purer sub-sample of sources in the `qso_candidates` table (left) and `galaxy_candidates` table (right). The plot is shown at HEALpixel level 7 (0.210 sq. deg.) in Hammer–Aitoff projection. The colour scale, which is logarithmic, covers the full range for each panel, so is different for each panel. Compare to Fig. 5 for the full tables.

Table 10. Quasar emission lines found in the Milliquas-based composite spectrum and covering the redshift range $0.052 \leq z \leq 4.35$ (Fig. 33). Each emission line is visually inspected before a quadratic polynomial is fit in the vicinity of its apparent peak using five samples of flux density. The maximum of the quadratic curve provides the observed rest-frame wavelength position, $\lambda_{\rm obs}$, of the line and its maximum flux density compared to Lyα, $F/F_{{\rm Ly}\alpha}$ (where the number of significant digits is provided in parenthesis). Because of the intricacies inherent in the fit of a local continuum to faint, broad, and/or blended emission lines, such a continuum was not subtracted from the flux densities reported here. This explains the differences between this table and the values found in Table 9.

Emission line $+\lambda_{\rm lab}$ (nm)	$\lambda_{\rm obs}$ (nm)	$F/F_{{\rm Ly}\alpha}$
Lyϵ λ 93.780	94.356 ± 1.327	0.344(2)
Lyδ λ 94.974	···	···
C III λ 97.702	98.632 ± 0.237	0.393(2)
N III λ 99.069	···	···
Lyβ λ 102.572	103.423 ± 0.139	0.479(2)
O VI λ 103.383	···	···
Fe II λ 111.208[a]	112.098 ± 0.500	0.328(2)
Lyα λ 121.567	122.250 ± 0.033	1.000(2)
O I λ 130.435	130.495 ± 0.545	0.402(2)
Si II λ 130.682	···	···
Si IV λ 139.676	139.568 ± 0.098	0.427(2)
O IV] λ 140.206	···	···
C IV λ 154.906	154.469 ± 0.061	0.485(3)
C III λ 190.873	189.922 ± 0.102	0.327(3)
Fe III λ 205.271[a]	205.600 ± 0.555	0.230(3)
C II] λ 232.644	232.390 ± 0.350	0.202(3)
[Ne IV] λ 242.383	242.617 ± 0.324	0.198(3)
Mg II λ 279.875	279.888 ± 0.039	0.212(3)
He I λ 318.867	318.612 ± 0.295	0.136(3)
Hγ λ 434.168	434.052 ± 0.067	0.081(3)
Fe II λ 453.780[a]	454.281 ± 0.569	0.070(4)
Hβ λ 486.268	486.701 ± 0.118	0.085(4)
[O III] λ 496.030	···	···
[O III] λ 500.824	···	···
[N I] λ 520.053	522.270 ± 0.615	0.054(4)
He I λ 587.729	588.441 ± 0.411	0.044(4)
Hα λ 656.461	656.283 ± 0.014	0.123(4)
He I λ 706.720	706.666 ± 0.945	0.034(4)
O I λ 844.868	850.091 ± 0.413	0.033(4)
Ca II λ 850.036	···	···

[a] Broad feature composed of Fe multiplets.

tering. Such sources can be easily removed by the user, as shown in appendix B.3.

Table 11. ADQL query to select the purer quasar sub-sample.

```
SELECT * FROM gaiadr3.qso_candidates
WHERE (gaia_crf_source='true' OR
       host_galaxy_flag<6 OR
       classlabel_dsc_joint='quasar' OR
       vari_best_class_name='AGN')
```

Table 12. ADQL query to select the purer galaxy sub-sample.

```
SELECT * FROM gaiadr3.galaxy_candidates
WHERE (radius_sersic IS NOT NULL OR
       classlabel_dsc_joint='galaxy' OR
       vari_best_class_name='GALAXY')
```

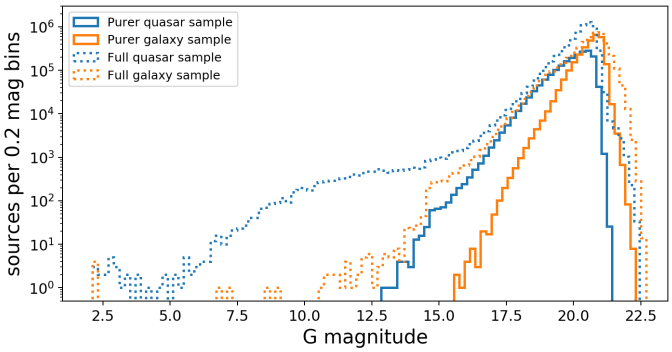

Fig. 36. G-band magnitude distribution of the purer sub-sample of objects in the `qso_candidates` (blue) and `galaxy_candidates` (orange) table on a logarithmic scale. The dotted lines show the distributions for the full tables.

The magnitude distributions of the purer sub-samples are shown in Fig. 36. Compared to the distribution for the full set (dotted lines), we see that the purer sub-sample has excluded the brightest sources (the presence of which appears exaggerated, however, due to the logarithmic number scale). The faintest quasars have also been removed.

Colour–magnitude and colour–colour diagrams of the purer sub-samples are shown in Fig. 37 and can be compared with the same diagrams for the full tables in Fig. 7. This shows that the purer sub-samples have a tighter colour distribution, and remove many of the fainter galaxies.

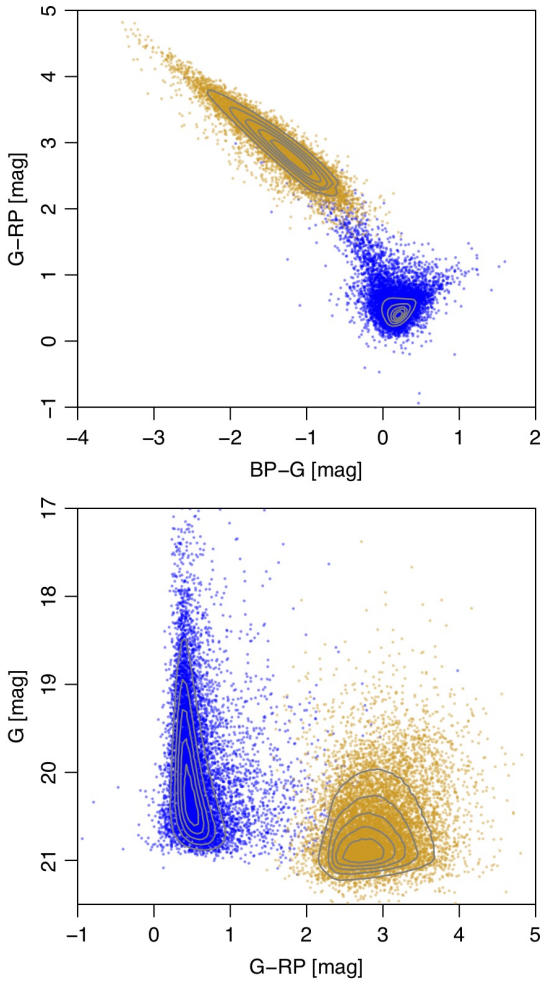

Fig. 37. Colour–colour diagram (top) and colour–magnitude diagram (bottom) for the purer sub-sample of sources in the `qso_candidates` table (blue) and `galaxy_candidates` table (orange). The contours show density on a linear scale. The points are a random selection of 10 000 sources for each class. Compare to Fig. 7 for the full tables.

There are other ways to obtain purer sub-samples of the integrated tables. One could, for example, select on a higher probability threshold of the DSC probabilities (the probabilities for all three DSC classifiers are listed in the `astrophysical_parameters` table). An example is shown in appendix B.3. The variation of purity with threshold is explored in Bailer-Jones (2021). The joint flag used in the purer sub-samples (Tables 11 and Tables 12) corresponds to a threshold of 0.5. One could also use a higher threshold on the `vari_best_class_score` and `vari_agn_membership_score` rankings from the Vari-Classification and Vari-AGN modules listed in the `qso_candidates` table.

In Sect. 5.4, we identified a purer sub-sample of the `qso_candidates` table via an analysis of astrometric distributions that are consistent with a uniform sky distribution of infinitely distant objects. The sources in this astrometric selection are indicated by the boolean flag `astrometric_selection_flag` in the `qso_candidates` table. Although this certainly excludes genuine quasars, it should be a reasonably pure list. Of the 1 897 754 sources in the astrometric sample, 1 801 255 are in common with the purer quasar sub-sample defined in Table 11, and the union of these two sets contains 2 039 324 sources. An equivalent flag is not available for the `galaxy_candidates` table, for reasons discussed in Sect. 5.4.

9. Conclusions

We have described the data products released in *Gaia* DR3 for 11.3 million candidate quasars and galaxies. This set arises from both a classification using the *Gaia* data and from an analysis of sources identified by external surveys cross-matched to *Gaia*. The information on these sources is presented in the `qso_candidates` and `galaxy_candidates` integrated tables. Further information, also on additional lower probability candidates, is provided in several other tables (see Sect. 3). Our integrated tables are completeness driven, so many sources in them will not be true extragalactic objects. We therefore also provide a purer sub-sample of 4.8 million quasars and galaxies (see Sect. 8).

We foresee a number of use cases for our results, including: aiding confirmation of candidates found in other surveys; identifying unusual or rare objects; providing targets for spectroscopic follow-up; providing input data for more focused classifications or characterisations. It is our hope and expectation that the community can build on and improve our results also by combining them with data from other surveys.

As with previous data releases, *Gaia* DR3 is an intermediate data release, this time based on 34 months of mission data. The next data release will be based on 66 months of data, with correspondingly higher S/N and lower systematic errors. We plan to use those data, along with improvements in our algorithms, to update both our classifications and our characterisations of extragalactic objects.

Acknowledgements

This work presents results from the European Space Agency (ESA) space mission *Gaia*. *Gaia* data are processed by the 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` and the archive website is `https://archives.esac.esa.int/gaia`. Further acknowledgments are given in appendix A.

References

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
Álvarez, M. A., Dafonte, C., Manteiga, M., Garabato, D., & Santoveña, R. 2021, Neural Computing and Applications
Assef, R. J., Stern, D., Noirot, G., et al. 2018, ApJS, 234, 23
Astropy Collaboration, Price-Whelan, A., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Bailer-Jones, C. A. L. 2021, GAIA-C8-TN-MPIA-CBJ-094
Bailer-Jones, C. A. L., Fouesneau, M., & Andrae, R. 2019, MNRAS, 490, 5615
Bellazzini at al. 2022, A&A in prep.
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Butler, N. R. & Bloom, J. S. 2011, AJ, 141, 93
Carnerero et al. 2022, A&A
Carrasco, J. M., Weiler, M., Jordi, C., et al. 2021, A&A, 652, A86
Chang, C.-C. & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology, 2, 27:1
Charlot, P., Jacobs, C. S., Gordon, D., et al. 2020, A&A, 644, A159
Cortes, C. & Vapnik, V. 1995, Machine Learning, 20, 273
Creevey, O., Sordo, R., Pailler, F., & et al. 2022, A&A in prep.
Cutri, R. M. & et al. 2012, VizieR Online Data Catalog, II/311

De Angeli et al. 2022, A&A
de Bruijne, J. H. J., Allen, M., Azaz, S., et al. 2015, A&A, 576, A74
de Souza, R. E., Krone-Martins, A., dos Anjos, S., Ducourant, C., & Teixeira, R. 2014, A&A, 568, A124
Delchambre, L., Krone-Martins, A., Wertz, O., et al. 2019, A&A, 622, A165
Delchambre et al. 2022, A&A
Ducourant et al. 2022, A&A in prep.
Evans et al. 2022, A&A in prep.
Flesch, E. W. 2015, PASA, 32, e010
Flesch, E. W. 2021, arXiv e-prints, arXiv:2105.12985
Fousneau et al. 2022, A&A
Fu, Y., Wu, X.-B., Yang, Q., et al. 2021, ApJS, 254, 6
Gaia Collaboration. 2018, A&A, 616, A14
Gaia Collaboration, Klioner, S. A., Mignard, F., et al. 2021, A&A, 649, A9
Gaia Collaboration & Klioner et al. 2022, A&A
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Heintz, K. E., Fynbo, J. P. U., & Høg, E. 2015, A&A, 578, A91
Heintz, K. E., Fynbo, J. P. U., Høg, E., et al. 2018, A&A, 615, L8
Holl et al. 2022, A&A in prep.
Hunter, J. D. 2007, Computing In Science & Engineering, 9, 90
IAU. 2021, Resolution B3, `https://iau.org/administration/resolutions/general_assemblies/`
Katz et al. 2022, A&A
Kohonen, T. 1982, Biological Cybernetics, 43, 59
Krone-Martins, A., Delchambre, L., Wertz, O., et al. 2018, A&A, 616, L11
Krone-Martins, A., Gavras, P., Ducourant, C., et al. 2022, A&A
Li, J., Silverman, J. D., Ding, X., et al. 2021, ApJ, 918, 22
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
Ma, C., Arias, E. F., Bianco, G., et al. 2009, IERS Technical Note, 35, 1
Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. 2021, ApJS, 253, 8
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Padilla, N. D. & Strauss, M. A. 2008, MNRAS, 388, 1321
Paine, J., Darling, J., & Truebenbach, A. 2018, ApJS, 236, 37
Pâris, I., Petitjean, P., Aubourg, É., et al. 2018, A&A, 613, A51
Pâris, I., Petitjean, P., Ross, N. P., et al. 2017, A&A, 597, A79
Pérez, F. & Granger, B. E. 2007, Computing in Science and Engineering, 9, 21
R Core Team. 2020, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria
Raiteri, C. M., Villata, M., Acosta-Pulido, J. A., et al. 2017, Nature, 552, 374
Rimoldini, L., Holl, B., Audard, M., et al. 2019, A&A, 625, A97
Rimoldini et al. 2022, A&A
Secrest, N. J., Dudik, R. P., Dorland, B. N., et al. 2015, ApJS, 221, 12
Simonetti, J. H., Cordes, J. M., & Heeschen, D. S. 1985, ApJ, 296, 46
Souchay, J., Andrei, A. H., Barache, C., et al. 2015, A&A, 583, A75
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
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
Vanden Berk, D. E., Richards, G. T., Bauer, A., et al. 2001, AJ, 122, 549
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Wu, Q., Liao, S., Qi, Z., et al. 2021, arXiv e-prints, arXiv:2111.02131

[1] Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
[2] Telespazio UK S.L. for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[3] Institut d'Astrophysique et de Géophysique, Université de Liège, 19c, Allée du 6 Août, B-4000 Liège, Belgium
[4] Laboratoire d'astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France
[5] CIGUS CITIC - Department of Computer Science and Information Technologies, University of A Coruña, Campus de Elviña s/n, A Coruña, 15071, Spain
[6] Department of Astrophysics, Astronomy and Mechanics, National and Kapodistrian University of Athens, Panepistimiopolis, Zografos, 15783 Athens, Greece
[7] National Observatory of Athens, I. Metaxa and Vas. Pavlou, Palaia Penteli, 15236 Athens, Greece
[8] Lohrmann Observatory, Technische Universität Dresden, Mommsenstraße 13, 01062 Dresden, Germany
[9] Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland
[10] 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
[11] INAF - Osservatorio Astrofisico di Torino, via Osservatorio 20, 10025 Pino Torinese (TO), Italy
[12] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom
[13] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
[14] RHEA for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[15] CENTRA, Faculdade de Ciências, Universidade de Lisboa, Edif. C8, Campo Grande, 1749-016 Lisboa, Portugal
[16] Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, 5226 Donald Bren Hall, 92697-3440 CA Irvine, United States
[17] Instituto de Astronomia, Geofisica e Ciências Atmosféricas, Universidade de São Paulo, Rua do Matão, 1226, Cidade Universitaria, 05508-900 São Paulo, SP, Brazil
[18] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
[19] INAF - Osservatorio astronomico di Padova, Vicolo Osservatorio 5, 35122 Padova, Italy
[20] European Space Agency (ESA), European Space Research and Technology Centre (ESTEC), Keplerlaan 1, 2201AZ, Noordwijk, The Netherlands
[21] GEPI, Observatoire de Paris, Université PSL, CNRS, 5 Place Jules Janssen, 92190 Meudon, France
[22] Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France
[23] Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstr. 12-14, 69120 Heidelberg, Germany
[24] Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland
[25] European Space Agency (ESA), European Space Astronomy Centre (ESAC), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[26] Aurora Technology for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[27] Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
[28] Lund Observatory, Department of Astronomy and Theoretical Physics, Lund University, Box 43, 22100 Lund, Sweden
[29] CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[30] Institut d'Astronomie et d'Astrophysique, Université Libre de Bruxelles CP 226, Boulevard du Triomphe, 1050 Brussels, Belgium
[31] F.R.S.-FNRS, Rue d'Egmont 5, 1000 Brussels, Belgium
[32] INAF - Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, 50125 Firenze, Italy
[33] European Space Agency (ESA, retired)
[34] University of Turin, Department of Physics, Via Pietro Giuria 1, 10125 Torino, Italy
[35] INAF - Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, via Piero Gobetti 93/3, 40129 Bologna, Italy
[36] DAPCOM for Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
[37] Royal Observatory of Belgium, Ringlaan 3, 1180 Brussels, Belgium
[38] Observational Astrophysics, Division of Astronomy and Space Physics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 20 Uppsala, Sweden
[39] ALTEC S.p.a, Corso Marche, 79,10146 Torino, Italy
[40] Sàrl, Geneva, Switzerland

[41] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, United Kingdom
[42] Gaia DPAC Project Office, ESAC, Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[43] SYRTE, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, LNE, 61 avenue de l'Observatoire 75014 Paris, France
[44] IMCCE, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Univ. Lille, 77 av. Denfert-Rochereau, 75014 Paris, France
[45] Serco Gestión de Negocios for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[46] CRAAG - Centre de Recherche en Astronomie, Astrophysique et Géophysique, Route de l'Observatoire Bp 63 Bouzareah 16340 Algiers, Algeria
[47] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, United Kingdom
[48] ATG Europe for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[49] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, 11 rue de l'Université, 67000 Strasbourg, France
[50] Kavli Institute for Cosmology Cambridge, Institute of Astronomy, Madingley Road, Cambridge, CB3 0HA
[51] Leibniz Institute for Astrophysics Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[52] INAF - Osservatorio Astrofisico di Catania, via S. Sofia 78, 95123 Catania, Italy
[53] Dipartimento di Fisica e Astronomia "Ettore Majorana", Università di Catania, Via S. Sofia 64, 95123 Catania, Italy
[54] INAF - Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monte Porzio Catone (Roma), Italy
[55] Space Science Data Center - ASI, Via del Politecnico SNC, 00133 Roma, Italy
[56] Department of Physics, University of Helsinki, P.O. Box 64, 00014 Helsinki, Finland
[57] Finnish Geospatial Research Institute FGI, Geodeetinrinne 2, 02430 Masala, Finland
[58] Institut UTINAM CNRS UMR6213, Université Bourgogne Franche-Comté, OSU THETA Franche-Comté Bourgogne, Observatoire de Besançon, BP1615, 25010 Besançon Cedex, France
[59] HE Space Operations BV for European Space Agency (ESA), Keplerlaan 1, 2201AZ, Noordwijk, The Netherlands
[60] Dpto. de Inteligencia Artificial, UNED, c/ Juan del Rosal 16, 28040 Madrid, Spain
[61] Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Eötvös Loránd Research Network (ELKH), MTA Centre of Excellence, Konkoly Thege Miklós út 15-17, 1121 Budapest, Hungary
[62] ELTE Eötvös Loránd University, Institute of Physics, 1117, Pázmány Péter sétány 1A, Budapest, Hungary
[63] Instituut voor Sterrenkunde, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium
[64] Department of Astrophysics/IMAPP, Radboud University, P.O.Box 9010, 6500 GL Nijmegen, The Netherlands
[65] University of Vienna, Department of Astrophysics, Türkenschanzstraße 17, A1180 Vienna, Austria
[66] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[67] Kapteyn Astronomical Institute, University of Groningen, Landleven 12, 9747 AD Groningen, The Netherlands
[68] School of Physics and Astronomy / Space Park Leicester, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom
[69] Thales Services for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[70] Depto. Estadística e Investigación Operativa. Universidad de Cádiz, Avda. República Saharaui s/n, 11510 Puerto Real, Cádiz, Spain
[71] Center for Research and Exploration in Space Science and Technology, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore MD, USA
[72] GSFC - Goddard Space Flight Center, Code 698, 8800 Greenbelt Rd, 20771 MD Greenbelt, United States
[73] EURIX S.r.l., Corso Vittorio Emanuele II 61, 10128, Torino, Italy
[74] Porter School of the Environment and Earth Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[75] Harvard-Smithsonian Center for Astrophysics, 60 Garden St., MS 15, Cambridge, MA 02138, USA
[76] HE Space Operations BV for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[77] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[78] LFCA/DAS,Universidad de Chile,CNRS,Casilla 36-D, Santiago, Chile
[79] SISSA - Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 265, 34136 Trieste, Italy
[80] Telespazio for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[81] University of Turin, Department of Computer Sciences, Corso Svizzera 185, 10149 Torino, Italy
[82] Centro de Astronomía - CITEVA, Universidad de Antofagasta, Avenida Angamos 601, Antofagasta 1270300, Chile
[83] DLR Gesellschaft für Raumfahrtanwendungen (GfR) mbH Münchener Straße 20 , 82234 Weßling
[84] Centre for Astrophysics Research, University of Hertfordshire, College Lane, AL10 9AB, Hatfield, United Kingdom
[85] University of Turin, Mathematical Department "G.Peano", Via Carlo Alberto 10, 10123 Torino, Italy
[86] INAF - Osservatorio Astronomico d'Abruzzo, Via Mentore Maggini, 64100 Teramo, Italy
[87] APAVE SUDEUROPE SAS for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[88] Mésocentre de calcul de Franche-Comté, Université de Franche-Comté, 16 route de Gray, 25030 Besançon Cedex, France
[89] ATOS for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[90] School of Physics and Astronomy, Tel Aviv University, Tel Aviv 6997801, Israel
[91] Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK
[92] Centre de Données Astronomique de Strasbourg, Strasbourg, France
[93] Institute for Computational Cosmology, Department of Physics, Durham University, Durham DH1 3LE, UK
[94] European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching, Germany
[95] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, 85748 Garching, Germany
[96] Data Science and Big Data Lab, Pablo de Olavide University, 41013, Seville, Spain
[97] Barcelona Supercomputing Center (BSC), Plaça Eusebi Güell 1-3, 08034-Barcelona, Spain
[98] ETSE Telecomunicación, Universidade de Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Galicia, Spain
[99] Asteroid Engineering Laboratory, Space Systems, Luleå University of Technology, Box 848, S-981 28 Kiruna, Sweden
[100] Vera C Rubin Observatory, 950 N. Cherry Avenue, Tucson, AZ 85719, USA
[101] TRUMPF Photonic Components GmbH, Lise-Meitner-Straße 13, 89081 Ulm, Germany
[102] IAC - Instituto de Astrofisica de Canarias, Via Láctea s/n, 38200 La Laguna S.C., Tenerife, Spain
[103] Department of Astrophysics, University of La Laguna, Via Láctea s/n, 38200 La Laguna S.C., Tenerife, Spain

[104] Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands
[105] Radagast Solutions
[106] Laboratoire Univers et Particules de Montpellier, CNRS Université Montpellier, Place Eugène Bataillon, CC72, 34095 Montpellier Cedex 05, France
[107] Université de Caen Normandie, Côte de Nacre Boulevard Maréchal Juin, 14032 Caen, France
[108] LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Université de Paris, 5 Place Jules Janssen, 92190 Meudon, France
[109] SRON Netherlands Institute for Space Research, Niels Bohrweg 4, 2333 CA Leiden, The Netherlands
[110] Astronomical Observatory, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warszawa, Poland
[111] Scalian for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[112] Université Rennes, CNRS, IPR (Institut de Physique de Rennes) - UMR 6251, 35000 Rennes, France
[113] INAF - Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131, Napoli, Italy
[114] Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, People's Republic of China
[115] University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District, Beijing 100049, People's Republic of China
[116] Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen Ø, Denmark
[117] DXC Technology, Retortvej 8, 2500 Valby, Denmark
[118] Las Cumbres Observatory, 6740 Cortona Drive Suite 102, Goleta, CA 93117, USA
[119] CIGUS CITIC, Department of Nautical Sciences and Marine Engineering, University of A Coruña, Paseo de Ronda 51, 15071, A Coruña, Spain
[120] Astrophysics Research Institute, Liverpool John Moores University, 146 Brownlow Hill, Liverpool L3 5RF, United Kingdom
[121] IPAC, Mail Code 100-22, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA
[122] IRAP, Université de Toulouse, CNRS, UPS, CNES, 9 Av. colonel Roche, BP 44346, 31028 Toulouse Cedex 4, France
[123] MTA CSFK Lendület Near-Field Cosmology Research Group, Konkoly Observatory, MTA Research Centre for Astronomy and Earth Sciences, Konkoly Thege Miklós út 15-17, 1121 Budapest, Hungary
[124] Departmento de Física de la Tierra y Astrofísica, Universidad Complutense de Madrid, 28040 Madrid, Spain
[125] Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[126] Villanova University, Department of Astrophysics and Planetary Science, 800 E Lancaster Avenue, Villanova PA 19085, USA
[127] INAF - Osservatorio Astronomico di Brera, via E. Bianchi, 46, 23807 Merate (LC), Italy
[128] STFC, Rutherford Appleton Laboratory, Harwell, Didcot, OX11 0QX, United Kingdom
[129] Charles University, Faculty of Mathematics and Physics, Astronomical Institute of Charles University, V Holesovickach 2, 18000 Prague, Czech Republic
[130] Department of Particle Physics and Astrophysics, Weizmann Institute of Science, Rehovot 7610001, Israel
[131] Department of Astrophysical Sciences, 4 Ivy Lane, Princeton University, Princeton NJ 08544, USA
[132] Departamento de Astrofísica, Centro de Astrobiología (CSIC-INTA), ESA-ESAC. Camino Bajo del Castillo s/n. 28692 Villanueva de la Cañada, Madrid, Spain
[133] naXys, University of Namur, Rempart de la Vierge, 5000 Namur, Belgium
[134] CGI Deutschland B.V. & Co. KG, Mornewegstr. 30, 64293 Darmstadt, Germany
[135] Institute of Global Health, University of Geneva
[136] Astronomical Observatory Institute, Faculty of Physics, Adam Mickiewicz University, Poznań, Poland
[137] H H Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, United Kingdom
[138] Department of Physics and Astronomy G. Galilei, University of Padova, Vicolo dell'Osservatorio 3, 35122, Padova, Italy
[139] CERN, Geneva, Switzerland
[140] Applied Physics Department, Universidade de Vigo, 36310 Vigo, Spain
[141] Association of Universities for Research in Astronomy, 1331 Pennsylvania Ave. NW, Washington, DC 20004, USA
[142] European Southern Observatory, Alonso de Córdova 3107, Casilla 19, Santiago, Chile
[143] Sorbonne Université, CNRS, UMR7095, Institut d'Astrophysique de Paris, 98bis bd. Arago, 75014 Paris, France
[144] Faculty of Mathematics and Physics, University of Ljubljana, Jadranska ulica 19, 1000 Ljubljana, Slovenia

Appendix A: Acknowledgements

The *Gaia* mission and data processing have been financially 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, 50QG1001, 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.

We made use of the following tools in the preparation of this paper: (SIMBAD, Wenger et al. 2000) and VizieR (Ochsenbein et al. 2000) operated at (CDS) Strasbourg; NASA ADS; TOPCAT (Taylor 2005); Matplotlib (Hunter 2007); IPython (Pérez & Granger 2007); Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2018); R (R Core Team 2020).

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University,

University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

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 website 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, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, 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.

Appendix B: ADQL queries

Appendix B.1: Purer sub-samples

The queries that provide the recommended purer sub-samples of quasars and galaxies are shown in Tables 11 and 12. These sub-samples are discussed in Sect. 8.

Appendix B.2: Sources used for the construction of a Gaia-only composite spectrum of quasars

The following query returns the `source_id` of the set of sources used to construct the *Gaia*-only composite spectra of quasars. It selects 111 563 sources from the purer quasar sub-sample (Table 11) that have both a reliable redshift estimate from QSOC (`flags_qsoc` = 0 or `flags_qsoc` = 16, as explained in Delchambre et al. 2022) and BP/RP coefficients that are published in *Gaia* DR3.

```
SELECT source_id
FROM gaiadr3.qso_candidates
JOIN gaiadr3.xp_summary USING (source_id)
WHERE (classlabel_dsc_joint='quasar'
       OR vari_best_class_name='AGN'
       OR host_galaxy_detected='true'
       OR gaia_crf_source='true')
  AND (flags_qsoc = 0 OR flags_qsoc = 16)
```

Appendix B.3: Using DSC probability thresholds

The following query uses thresholds on DSC-Specmod and DSC-Allosmod, which are in `astrophysical_parameters` table, to select sources from the `qso_candidates` table. This example uses thresholds of 0.9 and returns 371 708 sources. This compares to 547 201 returned when using thresholds of 0.5 (which is equivalent to, but slower than, selecting on `classlabel_dsc_joint` = `quasar` supplied in the `qso_candidates` table itself). An analogous query can be used for the `galaxy_candidates` table.

```
SELECT source_id
FROM gaiadr3.qso_candidates
JOIN gaiadr3.astrophysical_parameters USING
     (source_id)
WHERE classprob_dsc_specmod_quasar>0.9 AND
      classprob_dsc_allosmod_quasar>0.9
```

If we want to exclude generous regions around the LMC and SMC from the DSC purer subset, we can use the following query (change the initial line to select the fields you want). It removes 22 705 sources in the LMC and 7456 in the SMC, to leave 517 040 sources.

```
SELECT source_id
FROM (SELECT *
  FROM gaiadr3.qso_candidates
      WHERE classlabel_dsc_joint='quasar') AS temp
JOIN gaiadr3.gaia_source USING (source_id)
WHERE 1!=CONTAINS(
  POINT('ICRS', 81.3, -68.7),
  CIRCLE('ICRS', ra, dec, 9)) AND
  1!=CONTAINS(
  POINT('ICRS', 16.0, -72.8),
  CIRCLE('ICRS', ra, dec, 6))
```

Combining the two queries above to use higher thresholds on the DSC probabilities and to exclude the LMC and SMC, we get the following query, which returns 366 574 sources.

```
SELECT source_id
FROM (SELECT *
  FROM gaiadr3.qso_candidates
  JOIN gaiadr3.astrophysical_parameters USING
  (source_id)
      WHERE classprob_dsc_specmod_quasar>0.9 AND
      classprob_dsc_allosmod_quasar>0.9) AS temp
JOIN gaiadr3.gaia_source USING (source_id)
WHERE 1!=CONTAINS(
  POINT('ICRS', 81.3, -68.7),
  CIRCLE('ICRS', ra, dec, 9)) AND
  1!=CONTAINS(
  POINT('ICRS', 16.0, -72.8),
  CIRCLE('ICRS', ra, dec, 6))
```

Appendix C: Computation of the composite spectra of quasars

The computation of a composite spectrum of quasars from individual spectra may appear to be a straightforward task, but it turns out not to be for a number of reasons.

1. Quasars cover a large range of redshifts, equivalently a large range of rest-frame wavelengths, whereas each spectrum covers only a limited fraction of these rest-frame wavelengths.
2. Spectra have very different apparent luminosities, either because some are intrinsically brighter or fainter, or because of their difference in redshift, or because they are gravitationally lensed. Accordingly, each spectrum contributing to the composite spectrum must be scaled in order to reduce the dispersion of flux in rest-frame wavelength.

3. Spectra often have correlated noise in their fluxes that should be taken into account.
4. Quasars can have different continuum slopes (or any other background signal) that we may want to subtract in order to produce a pure emission line composite spectrum.
5. Spectra used to build the composite spectrum may have different resolutions, sampling and line spread function, as in BP/RP spectra, that should be first homogenized so as to model the sole signal of interest.

With all of these arguments in mind, we developed a new method for computing a composite BP/RP spectrum based on maximum likelihood estimation through the minimization of[3]

$$\chi^2 = \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \mathbf{M_i} \left[\mathbf{P} \boldsymbol{f}_i + \boldsymbol{m} s_i \right] \right\|^2_{\mathbf{W_i}}, \tag{C.1}$$

where

- N is the number of observations (spectra).
- $\boldsymbol{x}_i$ is the i-th observation vector, here taken as a concatenation of the BP and RP spectral coefficients, which are coefficients associated with a linear spline basis functions that represent the BP/RP spectra Carrasco et al. (2021).
- $\mathbf{W_i}$ is the weight matrix associated with $\boldsymbol{x}_i$. If $\mathbf{L_i}$ is the Cholesky decomposition of the covariance matrix associated with $\boldsymbol{x}_i$, $\mathbf{C_i} = \mathbf{L_i}\mathbf{L_i}^T$, then $\mathbf{W_i} = \mathbf{L_i}^{-1}$ such that $\mathbf{C_i}^{-1} = \mathbf{W_i}^T\mathbf{W_i}$.
- $\boldsymbol{m}$ is the composite spectrum we are inferring. We additionally infer the scaling factors, s_i, one associated with each observation i.
- $\mathbf{P}$ is a matrix composed of a set of basis functions used to model a background signal to be subtracted from the spectra. The linear coefficients associated with $\mathbf{P}$ for the i-th observation are given by the column vector $\boldsymbol{f}_i$. An example use of $\mathbf{P}$ would be to model the quasar continua as a low order polynomial (whose coefficients are computed in $\boldsymbol{f}_i$) and to subtract these continua from the spectra in Eq. C.1. This produces a pure emission line composite spectrum in $\boldsymbol{m}$. The matrix $\mathbf{P}$ could be a set of vectors resulting from a previous minimization of Eq. C.1, namely $\mathbf{P} = \mathbf{P_t} = \begin{pmatrix} \mathbf{P_{t-1}} & \boldsymbol{m}_t \end{pmatrix}$ where $\boldsymbol{m}_t$ minimizes $\chi^2_t = \sum_i \left\| \boldsymbol{x}_i - \mathbf{M_i} \left[\mathbf{P_{t-1}} \boldsymbol{f}_i^{(t)} + \boldsymbol{m}_t s_i^{(t)} \right] \right\|^2_{\mathbf{W_i}}$. This method can be seen as a weighted principal component analysis, in the sense that $\mathbf{P_t}$ are the minimal set of t components that minimize χ^2_t. Although mentioned here for completeness, we decided not to subtract the quasar continua in the present study, so we set $\mathbf{P} = 0$ and $\boldsymbol{f}_i = 0$.
- $\mathbf{M_i}$ is a transformation matrix, associated with $\boldsymbol{x}_i$, that projects $\mathbf{P}$ and $\boldsymbol{m}$ into the space of $\boldsymbol{x}_i$. In the present application, the goal of $\mathbf{M_i}$ is twofold. First it isolates the rest-frame wavelength regions from $\mathbf{P}$ and $\boldsymbol{m}$ that correspond to the observed wavelength region in $\boldsymbol{x}_i$ (the source redshift must therefore be taken into account). Second, the shifted and resampled spectrum is converted into BP/RP spectral coefficients through the use of the GaiaXPy simulator. See the documentation on `simulate_continuous` for more information on the calibration procedure.

In the minimization of Eq. C.1, $\boldsymbol{m}$ and $\boldsymbol{a}_i = \begin{pmatrix} \boldsymbol{f}_i \\ s_i \end{pmatrix}$ are free to vary. However, for a given value of $\boldsymbol{m}$, we can differentiate Eq. C.1 with respect to $\boldsymbol{a}_i$ and set the resulting gradient to zero to get[4]

$$\boldsymbol{a}_i = \left(\mathbf{T_i}^T \mathbf{W_i}^T \mathbf{W_i} \mathbf{T_i} \right)^{-1} \mathbf{T_i}^T \mathbf{W_i}^T \mathbf{W_i} \boldsymbol{x}_i \quad \text{where} \quad \mathbf{T_i} = \mathbf{M_i} \begin{pmatrix} \mathbf{P} & \boldsymbol{m} \end{pmatrix}. \tag{C.2}$$

Substituting this last equation into Eq. C.1 makes it depend only on $\boldsymbol{m}$, such that any (global) optimization algorithm can be used with a number of unknowns given by the number of fluxes in $\boldsymbol{m}$. In the present study, we use an expectation-maximization algorithm with momentum in batch mode. The steps of the expectation-maximization algorithm are: (i) compute $\boldsymbol{a}_i$ using Eq. C.2; (ii) fit $\boldsymbol{m}$ to all $\boldsymbol{x}_i$ given the previously computed $\boldsymbol{a}_i$,

$$\boldsymbol{m} = \left(\sum_{i=1}^{N} s_i^2 \, \mathbf{M_i}^T \mathbf{W_i}^T \mathbf{W_i} \mathbf{M_i} \right)^{-1} \left(\sum_{i=1}^{N} s_i \, \mathbf{M_i}^T \mathbf{W_i}^T \mathbf{W_i} \left[\boldsymbol{x}_i - \mathbf{M_i} \mathbf{P} \boldsymbol{f}_i \right] \right). \tag{C.3}$$

Steps (i) and (ii) are then iterated until the reduced chi-square improves by no more that 0.001 for 16 consecutive iterations.

To first order, the covariance matrix associated with the formal uncertainties of the computed composite spectrum can be approximated through the asymptotic normality property of the maximum likelihood estimator as

$$\mathbf{\Sigma_m} \approx \left(\sum_{i=1}^{N} \mathbf{J_i}^T \mathbf{W_i}^T \mathbf{W_i} \mathbf{J_i} \right)^{-1} \tag{C.4}$$

where

$$\mathbf{J_i} = \frac{\partial \mathbf{T_i} \boldsymbol{a}_i}{\partial \boldsymbol{m}} = \mathbf{T_i} \frac{\partial \boldsymbol{a}_i}{\partial \boldsymbol{m}} + \mathbf{M_i} \, s_i$$

and

$$\frac{\partial \boldsymbol{a}_i}{\partial \boldsymbol{m}} = \left(\mathbf{T_i}^T \mathbf{W_i}^T \mathbf{W_i} \mathbf{T_i} \right)^{-1} \begin{pmatrix} -s_i \left[\mathbf{M_i} \mathbf{P} \right]^T \\ \left[\boldsymbol{x}_i - \mathbf{M_i} \mathbf{P} \boldsymbol{f}_i \right]^T - 2 s_i \left[\mathbf{M_i} \boldsymbol{m} \right]^T \end{pmatrix} \mathbf{W_i}^T \mathbf{W_i} \mathbf{M_i}.$$

Equation C.4 is only valid for large values of N. How large N should be is problem dependent. We performed simulations using 65 536 noisy realisations of a problem with $N = 64$ and eight variables in $\boldsymbol{m}$, where all matrices, except $\mathbf{W_i}$, are uniformly distributed in $[-1, 1]$ and $\mathbf{W_i}$ is an orthogonal transformation of matrices whose eigenvalues are uniformly drawn in $[0.001, 1]$. This led to a maximum absolute error in the median correlation coefficients of 0.007.

The Octave/Matlab source code for minimizing Equation C.1 and for computing the approximate covariance matrix from Equation C.4 is available at https://github.com/ldelchambre/gls_mean/.

[3] In our notation, $\|\boldsymbol{x}\|^2_{\mathbf{W}} = \|\mathbf{W}\boldsymbol{x}\|^2 = \boldsymbol{x}^T\mathbf{W}^T\mathbf{W}\boldsymbol{x}$.

[4] Direct inversion of the normal equations is known to be numerically unstable and should be avoided.