A&A 674, A31 (2023)
https://doi.org/10.1051/0004-6361/202243423

Astronomy & Astrophysics
Special issue

Gaia Data Release 3

Gaia Data Release 3

Apsis. III. Non-stellar content and source classification

L. Delchambre[1,⋆], C. A. L. Bailer-Jones[2], I. Bellas-Velidis[3], R. Drimmel[4], D. Garabato[5], R. Carballo[6], D. Hatzidimitriou[7,3], D. J. Marshall[8], R. Andrae[2], C. Dafonte[5], E. Livanou[7], M. Fouesneau[2], E. L. Licata[4], H. E. P. Lindstrøm[4,9,10], M. Manteiga[11], C. Robin[12], A. Silvelo[5], A. Abreu Aramburu[13], M. A. Álvarez[5], J. Bakker[24], A. Bijaoui[14], N. Brouillet[15], E. Brugaletta[16], A. Burlacu[17], L. Casamiquela[15,18], L. Chaoul[19], A. Chiavassa[14], G. Contursi[14], W. J. Cooper[20,4], O. L. Creevey[14], A. Dapergolas[3], P. de Laverny[14], C. Demouchy[21], T. E. Dharmawardena[2], B. Edvardsson[22], Y. Frémat[23], P. García-Lario[24], M. García-Torres[25], A. Gavel[26], A. Gomez[5], I. González-Santamaría[5], U. Heiter[26], A. Jean-Antoine Piccolo[19], M. Kontizas[7], G. Kordopatis[14], A. J. Korn[26], A. C. Lanzafame[16,27], Y. Lebreton[28,29], A. Lobel[23], A. Lorca[30], A. Magdaleno Romeo[17], F. Marocco[31], N. Mary[12], C. Nicolas[19], C. Ordenovic[14], F. Pailler[19], P. A. Palicio[14], L. Pallas-Quintela[5], C. Panem[19], B. Pichon[14], E. Poggio[14,4], A. Recio-Blanco[14], F. Riclet[19], J. Rybizki[2], R. Santoveña[5], L. M. Sarro[32], M. S. Schultheis[14], M. Segol[21], I. Slezak[14], R. L. Smart[4], R. Sordo[33], C. Soubiran[15], M. Süveges[34], F. Thévenin[14], G. Torralba Elipe[5], A. Ulla[35], E. Utrilla[30], A. Vallenari[33], E. van Dillen[21], H. Zhao[14], and J. Zorec[36]

(Affiliations can be found after the references)

Received 25 February 2022 / Accepted 28 May 2022

ABSTRACT

Context. As part of the third *Gaia* Data Release, we present the contributions of the non-stellar and classification modules from the eighth coordination unit (CU8) of the Data Processing and Analysis Consortium, which is responsible for the determination of source astrophysical parameters using *Gaia* data. This is the third in a series of three papers describing the work done within CU8 for this release.
Aims. For each of the five relevant modules from CU8, we summarise their objectives, the methods they employ, their performance, and the results they produce for *Gaia* DR3. We further advise how to use these data products and highlight some limitations.
Methods. The Discrete Source Classifier (DSC) module provides classification probabilities associated with five types of sources: quasars, galaxies, stars, white dwarfs, and physical binary stars. A subset of these sources are processed by the Outlier Analysis (OA) module, which performs an unsupervised clustering analysis, and then associates labels with the clusters to complement the DSC classification. The Quasi Stellar Object Classifier (QSOC) and the Unresolved Galaxy Classifier (UGC) determine the redshifts of the sources classified as quasar and galaxy by the DSC module. Finally, the Total Galactic Extinction (TGE) module uses the extinctions of individual stars determined by another CU8 module to determine the asymptotic extinction along all lines of sight for Galactic latitudes $|b| > 5°$.
Results. *Gaia* DR3 includes 1591 million sources with DSC classifications; 56 million sources to which the OA clustering is applied; 1.4 million sources with redshift estimates from UGC; 6.4 million sources with QSOC redshift; and 3.1 million level 9 HEALPixes of size 0.013 deg^2 where the extinction is evaluated by TGE.
Conclusions. Validation shows that results are in good agreement with values from external catalogues; for example 90% of the QSOC redshifts have absolute error lower than 0.1 for sources with empty warning flags, while UGC redshifts have a mean error of 0.008 ± 0.037 if evaluated on a clean set of spectra. An internal validation of the OA results further shows that 30 million sources are located in high confidence regions of the clustering map.

Key words. methods: data analysis – methods: statistical – Galaxy: fundamental parameters – dust, extinction – quasars: general – catalogs

1. Introduction

The ESA *Gaia* mission was designed to create the most precise three dimensional map of the Milky way, along with its kinematics, through the repeated observation of about two billion stars. *Gaia* observes all objects in the sky down to an apparent G magnitude of about 21 mag, which includes millions of galaxies and quasars (Gaia Collaboration 2016). The data collected between 25 July 2014 and 28 May 2017 (34 months) have been processed by the *Gaia* Data Processing and Analysis Consortium (DPAC) to provide the third data release of the *Gaia* catalogue, *Gaia* DR3.

For sources with $G \leq 17$ mag, typical positional uncertainties are on the order of 80 μas; parallax uncertainties on the order of 100 μas; proper motion uncertainties on the order of 100 μas yr^{-1}; and G magnitude uncertainties on the order of 1 mmag. In addition to this exquisite astrometric and photometric performance, *Gaia* provides high-resolution spectroscopy ($R = \lambda/\Delta\lambda \approx 11\,700$) centred around the calcium triplet (845–872 nm), hence its name radial velocity spectrometer (RVS), as well as low-resolution spectrophotometry from two instruments: the blue photometer (BP) covering the wavelength range 330–680 nm with $30 \leq R \leq 100$, and the red photometer (RP) covering the wavelength range 640–1050 nm with $70 \leq R \leq 100$ (Carrasco et al. 2021).

⋆ Corresponding author: L. Delchambre,
e-mail: ldelchambre@uliege.be

Eight coordination units (CUs) were set up within the DPAC, each focusing on a particular aspect of the *Gaia* processing: CU1 for managing the computer architecture; CU2 for the data simulations; CU3 for the core astrometric processing; CU4 for the analysis of non-single stars, Solar System objects, and extended objects; CU5 for the photometric BP/RP processing; CU6 for the spectroscopic RVS processing; CU7 for the variability analysis; and CU8 for the determination of the astrophysical parameters (APs) of the observed sources. Finally, a ninth CU is responsible for the catalogue validation, access, and publication.

This paper is the third in a series of three papers describing the processing done within CU8. The first of these, Creevey et al. (2023), summarises the work done in CU8 and the various APs it produces. The second, Fouesneau et al. (2023), describes stellar APs. The present paper discusses the object classification and the non-stellar APs produced by CU8, namely the redshifts of extragalactic sources and total Galactic extinction map. We describe the results and methods of the relevant modules, as they have evolved since their description given prior to launch (Bailer-Jones et al. 2013), while focusing on technical details. A thorough scientific analysis of these results, seen from a cross-CU perspective, can be found in performance verification papers like in Gaia Collaboration (2023), where the classification and characterisation of the extragalactic sources are discussed in more details.

We provide an overview of the data products from the classification and non-stellar modules in Sect. 2. The Discrete Source Classifier (DSC), which classifies sources probabilistically into five classes that are known a priori from its training set (quasar, galaxy, star, white dwarf, and physical binary star), is described in Sect. 3. The Outlier Analysis (OA), which complements the DSC classification through a clustering algorithm applied to BP/RP spectra of sources with low DSC probability, is described in Sect. 4. The quasar classifier (QSOC) and Unresolved Galaxy Classifier (UGC), both based on BP/RP spectra, make use of the DSC probabilities in order to identify quasars and galaxies and subsequently determine their redshifts; these are described in Sects. 5 and 6, respectively. Finally, the global stellar parameters of giant stars, as inferred from BP/RP spectra, allow the Total Galactic Extinction (TGE) module to derive the Galactic extinction seen along a given line-of-sight as described in Sect. 7. Finally, we summarise the improvements that are currently foreseen for *Gaia* DR4 in Sect. 8. Additional information on the design and performance of the modules can be found in the *Gaia* online documentation.

2. Overview of the non-stellar astrophysical parameters from CU8 in *Gaia* DR3

The five non-stellar modules together contribute to 110 unique fields in the *Gaia* DR3. Table 1 provides an overview of the tables and fields that each of the modules contributes to, including the resulting number of entries in each table. These fields are spread over eight different tables and concern about 1.6 billion unique sources. Figure 1 sketches the inter-dependency between these modules, the selection they apply on the DSC probabilities, their input, output, and the number of sources for which they produce results in *Gaia* DR3. The different selection policies from each module are clearly seen in this plot; each leads to a different associated completeness and purity. The filtering applied by each module on the results they produced is not mentioned here, although we should generally not expect the number of sources satisfying the provided DSC selection criteria to be equal to the number of sources for which there are results in *Gaia* DR3 for each module.

3. Source classification (DSC)

3.1. Objectives

DSC classifies *Gaia* sources probabilistically into five classes: quasar, galaxy, star, white dwarf, and physical binary star. These classes are defined by the training data, which are *Gaia* data, with labels provided by external catalogues. DSC comprises three classifiers: Specmod uses BP/RP spectra to classify into all five classes; Allosmod uses various other features to classify into just the first three classes; Combmod takes the output class probabilities of the other two classifiers and combines them to give combined probabilities in all five classes.

3.2. Method

3.2.1. Algorithms and I/O

Specmod uses an ExtraTrees classifier, which is an ensemble of classification trees. Each tree maps the 100-dimensional input space of the BP/RP spectrum – 60 samples each, minus 5 samples that are rejected at the edges of each spectrum – into regions that are then identified with each of the five classes. By using an ensemble of hundreds of trees, these individual discrete classifications are turned into class probabilities.

Allosmod uses a Gaussian Mixture Model (GMM). For each class, the distribution of the training data in an eight-dimensional feature space is modelled by a mixture of 25 Gaussians. This is done independently for all three classes (quasar, galaxy, star). Once appropriately normalised and a suitable prior applied, each GMM gives the probability that a feature vector (i.e. a new source) is of that class. The eight features are as follows; they are fields in the *Gaia* source table or are computed from these fields:

- sine of the Galactic latitude, $\sin$ `b`,
- parallax, `parallax`,
- total proper motion, `pm`,
- unit weight error (uwe),
  $= \sqrt{\frac{\texttt{astrometric\_chi2\_al}}{\texttt{astrometric\_n\_good\_obs\_al}-5}}$,
- G band magnitude, `phot_g_mean_mag`,
- colour $G_{\rm BP} - G$, `bp_g`,
- colour $G - G_{\rm RP}$, `g_rp`,
- The relative variability in the G band (relvarg),
  $= \sqrt{\texttt{phot\_g\_n\_obs}/\texttt{phot\_g\_mean\_flux\_over\_error}}$.

All eight features must exist for a given source for Allosmod to provide a probability. As explained below, we exploit some of the 'failures' of these features to help identify objects. For example, galaxies should have true proper motions (and parallaxes) very close to zero. Yet they sometimes have larger measured proper motions in *Gaia* DR3 on account of their physical extent combined with the variability in the calculation of the centroid during each scan made by *Gaia* (obtained at different position angles). This can give rise to spuriously large proper motions (although the uncertainties are also larger). In many cases, these solutions are rejected by the astrometric solutions (to give the so-called 2p solutions; see Lindegren et al. 2021 for the definitions), meaning that many galaxies lack parallaxes and proper motions and are therefore not processed by Allosmod.

Allosmod models the distribution of the data, and so it provides likelihoods. When combined with the class prior, this gives posterior class probabilities, which are the output from Allosmod. Specmod, in contrast, is a tree-based model that does not strictly provide posterior probabilities. Moreover, its output is influenced by the distribution in the training data (see below). However, by

Table 1. Individual contributions of the non-stellar CU8 modules to the *Gaia* DR3.

Module	Table and field names	Number of non-empty rows
DSC (source classification)	– astrophysical_parameters	
	classprob_dsc_allosmod [(a)]	1 370 759 105
	classprob_dsc_specmod [(b)], classprob_dsc_combmod [(c)]	1 590 760 469
	– gaia_source	
	classprob_dsc_combmod [(c)]	1 590 760 469
	– galaxy_candidates	
	classprob_dsc_combmod [(c)], classlabel_dsc, classlabel_dsc_joint	4 841 799
	– qso_candidates	
	classprob_dsc_combmod [(c)], classlabel_dsc, classlabel_dsc_joint	6 647 511
OA (source classification based on self-organising map)	– oa_neuron_information (78 fields)	900 (1 per neuron)
	– oa_neuron_xp_spectra (7 fields)	78 300 (900 neurons × 87 samples per spectrum)
	– astrophysical_parameters	
	neuron_oa_id, neuron_oa_dist neuron_oa_dist_percentile_rank, flags_oa	56 416 360
	– galaxy_candidates	
	classlabel_oa	1 901 026
	– qso_candidates	
	classlabel_oa	2 803 225
QSOC (quasar redshift determination)	– qso_candidates	
	redshift_qsoc, redshift_qsoc_lower redshift_qsoc_upper, ccfratio_qsoc, zscore_qsoc, flags_qsoc	6 375 063
UGC (galaxy redshift determination)	– galaxy_candidates	
	redshift_ugc, redshift_ugc_lower, redshift_ugc_upper	1 367 153
TGE (Galactic extinction)	– total_galactic_extinction_map (10 fields)	4 177 920 (49 152 in HEALPix level 6, 196 608 in level 7, 786 432 in level 8, 3 145 728 in level 9)
	– total_galactic_extinction_map_opt (7 fields)	3 145 728 (HEALPix level 9)

Notes. [(a)]Corresponding to classprob_dsc_allosmod_quasar, classprob_dsc_allosmod_galaxy and classprob_dsc_allosmod_star. [(b)]Corresponding to classprob_dsc_specmod_quasar, classprob_dsc_specmod_galaxy, classprob_dsc_specmod_star, classprob_dsc_specmod_whitedwarf and classprob_dsc_specmod_binarystar. [(c)]Corresponding to classprob_dsc_combmod_quasar, classprob_dsc_combmod_galaxy, classprob_dsc_combmod_star, classprob_dsc_combmod_whitedwarf and classprob_dsc_combmod_binarystar. See the sections dedicated to each module for a complete description of the fields and tables listed herein. Fields from module-specific tables (i.e. OA and TGE) are not listed here.

using the simple method described in the online documentation we can adjust the outputs from Specmod so that they are analogous to posterior probabilities that incorporate our desired class prior. Allosmod is described in more detail in Bailer-Jones et al. (2019), where it is applied to *Gaia* DR2 data.

The third DSC classifier, Combmod, takes the probabilities from Specmod and Allosmod for a source and combines them into a new posterior probability over all five classes. This is not entirely trivial, because it has to ensure that the global prior is not counted twice, and it has to allow for the fact that Specmod has more classes than Allosmod. The combination algorithm is described in Appendix B.

3.2.2. Class prior

Single stars hugely outnumber extragalactic sources in *Gaia*, and failing to take this into account would give erroneous probabilities and classifications. Specifically, if we were to assume equal priors for all classes, then when the attributes of a given source do not provide a strong discrimination between the classes, the source would be classified as any class with near equal probabilities. However, in reality, the source is far more likely to be a star, because extragalactic sources are so rare. We must therefore set appropriate priors for the classes. Failing to do so corresponds to the well-known base rate fallacy. We choose here to adopt a global prior that reflects the expected fraction of each class (as we define them) in the entire *Gaia* DR3 data set. This prior is given in Table 2. As the relative fraction of extragalactic to Galactic objects that *Gaia* observes varies with quantities such as magnitude and Galactic latitude, we could make the prior a function of these (and potentially other) quantities; but we have not introduced this in *Gaia* DR3.

Using the correct prior is important. A classifier with equal priors would perform worse on the rare objects than a classifier with appropriate priors, because the former would tend to misclassify many stars as being extragalactic. However, we would not notice this if we erroneously validated the classifier on a balanced set (equal numbers in each class), because such a validation set has an artificially low fraction of stars, and hence far too few potential contaminants. The classifier would perform worse but would appear to be performing better. This is demonstrated

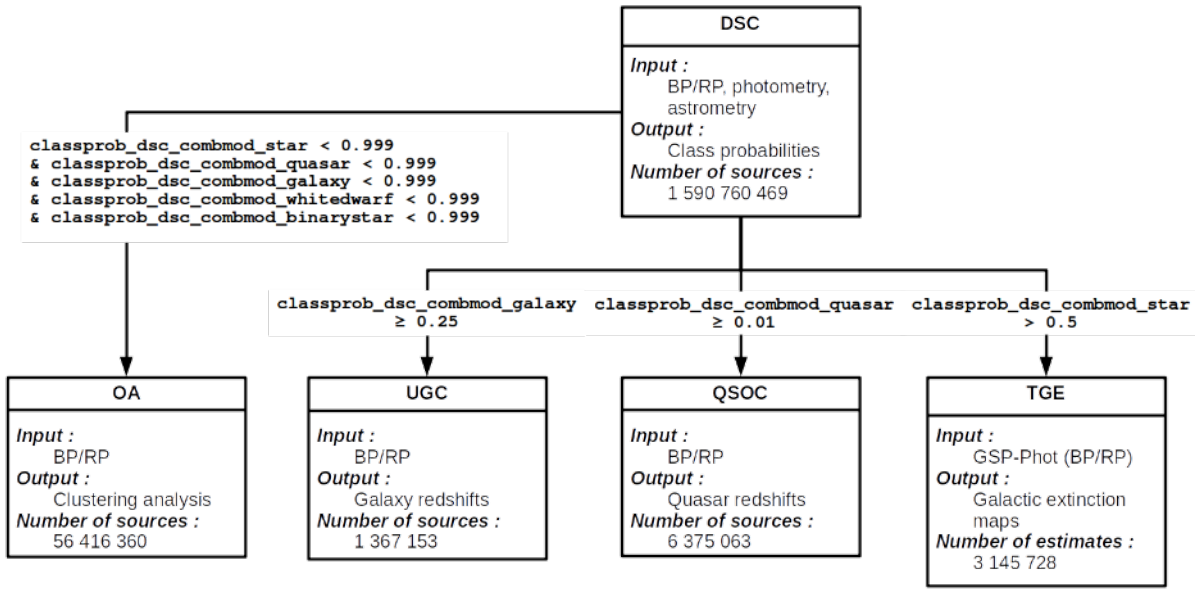

Fig. 1. Dependency of the OA, UGC, QSOC, and TGE modules on the DSC combined probabilities for the selection of the sources to be processed (`classprob_dsc_combod`, see Sect. 3 for a definition). For each module, we provide a synthetic view of their input and output, and the number of sources for which the module produces results in *Gaia* DR3. In the case of TGE, we provide the number of extinction estimates that were computed in level 9 HEALPixes (see Sect. 7). Unlike the other modules described here, TGE additionally relies on the General Stellar Parametrizer from Photometry (GSP-Phot) for its source selection and processing, which is described in Andrae et al. (2022).

Table 2. DSC class prior.

	quasar	galaxy	star	white dwarf	physical binary star
∝	1/1000	1/5000	1	1/5000	1/100
=	0.000989	0.000198	0.988728	0.000198	0.009887

Notes. The first row gives these as fractions relative to the stars, and the second row gives their decimal values summing to 1.0. This is the class prior for Specmod. The prior for the star class in Allosmod is the sum of star, white dwarf, and physical binary star.

in Table 1 of Bailer-Jones et al. (2019). We address this issue in the context of our validation data in Sect. 3.3.

3.2.3. Training data

DSC is trained empirically, meaning it is trained on a labelled subset of the actual *Gaia* data it will be applied to (except for binary stars). The classes were defined by selecting sources of each class from an external database and cross-matching them to *Gaia* DR3. The sources used to construct the training sets – and which therefore define the classes – are as follows (see the online documentation and Bailer-Jones 2021 for more details):

- Quasars: 300 000 spectroscopically confirmed quasars from the fourteenth release of the Sloan Digital Sky Survey (SDSS) catalogue, SDSS-DR14 (Pâris et al. 2018).
- Galaxies: 50 000 spectroscopically confirmed galaxies from SDSS-DR15 (Aguado et al. 2019).
- Stars: 720 000 objects drawn at random from *Gaia* DR3 that are not in the quasar or galaxy training sets. Strictly speaking, this is therefore an 'anonymous' class. But as the vast majority of sources in *Gaia* are stars, and the majority of those will appear in (spectro)photometry and astrometry as single stars, we call this class 'stars'.
- White dwarfs: 40 000 white dwarfs from the Montreal White Dwarf Database[1] that have coordinates and that are not known to be binaries using the flag provided in that table. This class is not in Allosmod.
- Physical binary stars: 280 000 BP/RP spectra formed by summing the two separate components in spatially-resolved binaries in *Gaia* DR3 (see the online documentation). This is only done for the BP/RP spectra, not for astrometry or photometry, so physical binaries are not a class in Allosmod.

The quasar, galaxy, and star class definitions are more or less the same as in Bailer-Jones et al. (2019).

The selected sources were filtered in order to remove obvious contaminants or problematic measurements (as described in the online documentation). The numbers above refer to what remains after this filtering. The remaining set was then split into roughly equally sized training and validation sets (per class). Generally speaking, the relative number of objects of each class – the *class fraction* – in the training data affects the output probabilities of a classifier, because it acts as an implicit prior in the classifier. However, for both Specmod and Allosmod, we remove this influence to ensure that their priors correspond to our class prior. We are therefore free to choose as many training examples in each class as we need, or can obtain, in order to learn the data distributions.

We note that for the common classes between Specmod and Allosmod, that is, quasars, galaxies, and stars, a common sample with complete input data was used to train both modules. In particular, this means that even though Specmod does not require parallaxes and proper motions as inputs, its training sample is

[1] http://www.montrealwhitedwarfdatabase.org

restricted to those sources that do have parallaxes and proper motions. This is important because Specmod is also applied to sources that lack parallaxes and proper motions, meaning that some of its results are on types of objects that are not represented in its training set. This is particularly important for galaxies.

Figure 2 (top) shows the distribution of the eight Allosmod features in the training data for the quasar and galaxy classes. As we do not want the model to learn the $\sin b$ distribution of extragalactic objects, which is just the SDSS footprint (shown in the plot), we replace this with a random value drawn from a uniform distribution in $\sin b$ (i.e. uniform sky density) when training Allosmod. This plot also shows, for comparison, the distribution of the features for the star class in the training data. Figure 3 (top) shows the distribution of the two colours of the quasars and galaxies in a colour–colour diagram.

3.2.4. Class labels

The main output from DSC is the class probabilities from all three classifiers. For convenience, we also compute two class labels from the probabilities, which appear only for sources in the `qso_candidates` and `galaxy_candidates` tables in the data release. The first label, `classlabel_dsc`, is set to the class that gets the highest posterior probability in Combmod that is greater than 0.5. If none of the output probabilities are above 0.5, this class label is `unclassified`. This gives a sample that is fairly complete for quasars and galaxies, but not very pure.

The second class label, `classlabel_dsc_joint`, identifies a purer set of quasars and galaxies. It is set to the class that achieves a probability above 0.5 in both Specmod and Allosmod. This produces purer samples because the Specmod and Allosmod probabilities are not perfectly correlated. This lack of correlation may be unexpected, but is what we want, because it means the classifiers are providing non-redundant information.

Because DSC is not the only contributor to the `qso_candidates` and `galaxy_candidates` tables, sources in the `qso_candidates` table can have either classlabel set to `galaxy`, and vice versa.

3.3. Performance: Purity and completeness

By assigning each source to the class with the largest probability, it is uniquely classified. An alternative is to additionally adopt a minimum probability threshold, in which case we can get multiple classifications if the threshold is low enough, or no classification if it is high enough. Doing this on sources with known classes (assumed to be correct), we can then compute the confusion matrix, which tells us how many sources of each true class are assigned to each DSC class. From this, we then compute, for each class, the completeness – the fraction of true positives among all trues – and the purity – the fraction of true positives among all positives.

Here we use the largest probabilities to compute the completenesses and purities on the validation sets[2]. As the class fractions in this validation set are not representative of what they are in *Gaia*, the raw purities are meaningless. Specifically, stars are far less common in the validation data than they are in a random sample of *Gaia* data, and so there are too few potential contaminants of the other classes in the validation data, resulting in significantly overestimated purities. This fact is sometimes overlooked in the validation of classification results in the literature. Fortunately, we can easily correct for this. As explained in section 3.4 (especially Eq. (4)) of Bailer-Jones et al. (2019), we can modify the confusion matrix to correspond to a validation set that has class fractions equal to the class prior. The purity computed from this modified confusion matrix is then appropriate for any randomly selected sample of *Gaia* sources (this modification does not affect the completeness). We note that this modification is independent of the fact that DSC probabilities are already posterior probabilities that take into account this class prior (i.e. both modifications must be done). This should also serve as a warning when assessing any classifier: if the validation data set does not have a representative fraction of contamination, or if this is not adjusted, the predicted purities will be erroneous.

Table 3 shows the completenesses and purities for the DSC classes and classifiers. This is the performance we expect for a sample selected at random from the entire *Gaia* dataset that has complete input data for both Specmod and Allosmod. It accommodates the rareness of all these classes, as specified by the global class prior (Table 2), both in the probabilities and the application data set. It is important to bear in mind that these purity and completeness measures only refer to the types of objects in the validation set. For extragalactic objects, this means objects classified as such by SDSS using the SDSS spectra. The overall population of extragalactic objects classified by DSC is of course broader than this, and so the completeness and purity evaluated on other subsets of extragalactic objects could differ.

Due to the dominance of single stars in *Gaia*, we are not really interested in the performance on this class. Indeed, it is trivial to get an excellent single-star classifier: simply call everything a single star and your classifier has 99.9% completeness and 99.9% purity.

The performance is modest overall, for reasons that are further discussed in Sect. 3.5. Results on binaries are very poor, partly because the validation set we used to compute the confusion matrix is not representative of the training set. This is because the validation set comprises only real *Gaia* objects, and so known unresolved binaries, whereas the training set was made by combining single star spectra. However, the internal performance on binaries was also poor. This suggests an intrinsic difficulty in separating binaries (as we define them) from single stars.

The performance in Table 3 refers to objects covering the full *Gaia* parameter space, in particular all magnitudes and Galactic latitudes. The purities tend to increase for brighter magnitudes, as can be seen from the plots in the online documentation and in Bailer-Jones (2021). There we see, for example, that for $G \leq 18$ mag, the purities for quasars and galaxies when using Allosmod alone is 80% or higher. However, when looking at the performance in a specific part of the parameter space, one should adopt a new prior that is appropriate for that part of the parameter space, for example fewer extragalactic objects visible at low latitudes. We then recompute the posterior probabilities (Appendix C) and the completenesses and purities (remembering that the adjustment of the confusion matrix must use the class fractions in this subset of the validation set). This we have done for sources outside of the Galactic plane, with results reported in the bottom two lines of Table 3. For $|b| > 11.54°$, we adopt a prior probability for quasars of 2.64×10^{-3} (9.9×10^{-4} globally), and a prior probability for galaxies of 5.3×10^{-4} (2×10^{-4} globally). The purities of the quasar and galaxy samples are significantly higher, as expected because there are fewer contaminating

[2] The validation data for the binaries is not the one mentioned in Sect. 3.2.3, namely synthetically-combined single stars, but instead a set of unresolved binaries directly from *Gaia*. See the online documentation for more details.

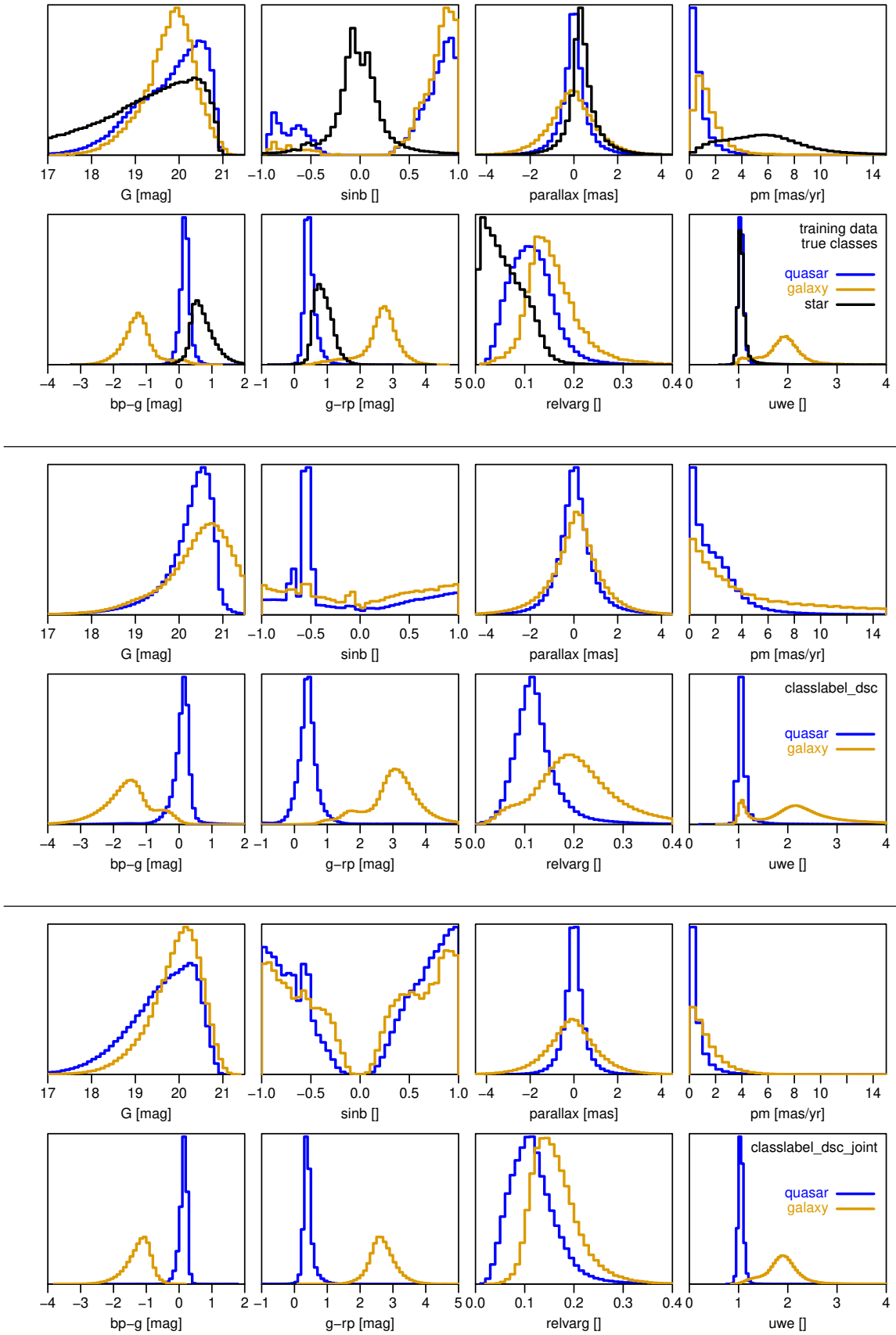

Fig. 2. Distribution (linear scale) of *Gaia* features for various samples used in DSC. *Top*: training data for quasars (blue), galaxies (orange), and stars (black). When training Allosmod, the $\sin b$ distributions for quasars and galaxies are replaced with uniform ones. *Middle*: *Gaia* sources assigned `classlabel_dsc='quasar'` (blue) and `classlabel_dsc='galaxy'` (orange). *Bottom*: *Gaia* sources assigned `classlabel_dsc_joint='quasar'` (blue) and `classlabel_dsc_joint='galaxy'` (orange).

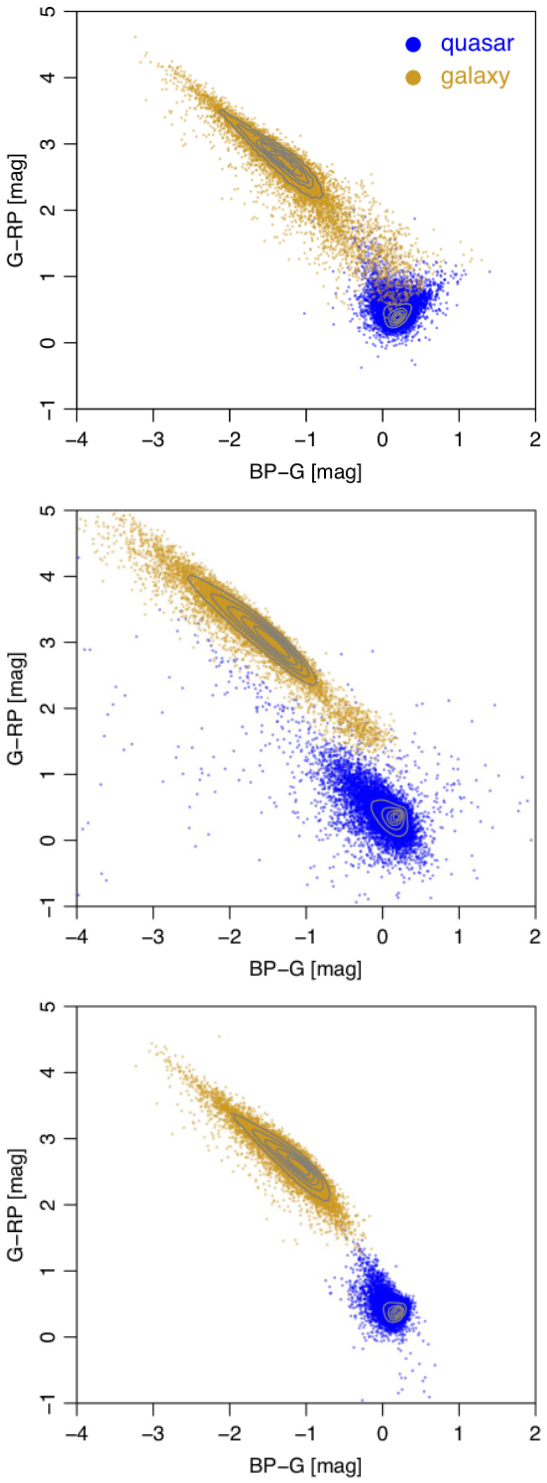

Fig. 3. Colour–colour diagrams for various samples used in DSC. *Top*: training data for quasars (blue) and galaxies (orange). *Middle*: *Gaia* sources assigned `classlabel_dsc='quasar'` (blue) and `classlabel_dsc='galaxy'` (orange). *Bottom*: *Gaia* sources assigned `classlabel_dsc_joint='quasar'` (blue) and `classlabel_dsc_joint='galaxy'` (orange). The differences in the distributions are due to the various levels of completeness and purity in the two types of class label.

stars per square degree. Using a probability threshold increases the purities even further, albeit at the expense of completeness (see online documentation for more plots). Clearly, if we were willing and able to push the prior for extragalactic objects higher, we would obtain higher purities.

3.4. Results

DSC was applied to all *Gaia* sources that have the required input data. Its results were not filtered in any way. In particular, we did not remove sources with lower quality input data, or that have input data lying outside the range of the training data. By including all results, we allow the user to apply their own filters according to their own goals and needs.

DSC produces outputs for 1 590 760 469 sources. All of these have probabilities from Combmod and Specmod, whereas 1 370 759 105 (86.2%) have probabilities from Allosmod[3]. This lower number from Allosmod is due to missing input data, usually missing parallaxes and proper motions (or missing colours in a few cases). That is, sources must have 5p or 6p astrometric solutions from the *Gaia* Astrometric Global Iterative Solution (AGIS) in order to have Allosmod results. This can be seen in Fig. 4, which shows the fraction of sources (per HEALPix) that have 5p/6p solutions, for those with `dsc_classlabel='quasar'` (left) and `dsc_classlabel='galaxy'` (right). While most objects classified as quasars have measured parallaxes (i.e. 5p or 6p solutions), most sources outside of the Galactic plane classified as galaxies do not. Those objects that lack parallaxes and proper motions (the 2p solutions) also lack Allosmod results, and so their Combmod results (and hence `dsc_classlabel`) are determined only by Specmod. We explore the differences between the 5p/6p and 2p solutions at the end of this section.

The vast majority of sources have high probabilities of being stars, and because the purities of the white dwarf and physical binary classes are low (see the online documentation), we focus here on the results for the quasar and galaxy classes.

The label `classlabel_dsc` (defined in Sect. 3.2.4) classifies 5 243 012 sources as quasars and 3 566 085 as galaxies. Their sky distributions are shown in the top two panels of Fig. 5. The analysis in Sect. 3.3 suggests that these samples are not very pure (see Table 3). In these sky plots, we see large overdensities of supposed quasars in several regions, in particular the LMC and SMC, suggesting that this sample is not very pure. However, such overdensities are expected when we have a constant misclassification rate over the whole sky, because any high-density region will have a high density of both correctly and incorrectly classified objects. However, it turns out that the fraction of sources classified as quasars is also higher than average in these regions (see below). The LMC and SMC are so dense that 38% of all the quasar identifications using `classlabel_dsc` are in the LMC, and 6.4% are in the SMC[4]. These percentages are much smaller for galaxies: just 3% for the LMC and 1% for the SMC.

The bottom row of Fig. 5 shows the distribution of the 547 201 sources classified as quasars and the 251 063 sources classified as galaxies by the purer class label `classlabel_dsc_joint`. The overdensities of quasars in the LMC and SMC regions are now greatly reduced, to 4% and 1% of all sources respectively.

[3] It so happens that all sources which have Allosmod results also have Specmod results, but not vice versa.

[4] For this purpose, the LMC is defined as a circle of 9° radius centred on RA = 81.3°, Dec = −68.7°, and the SMC as a circle of 6° radius centred on RA = 16.0°, Dec = −72.8°.

Table 3. DSC performance evaluated on the validation data set.

	Specmod		Allosmod		Combmod		Spec&Allos	
	Compl.	Purity	Compl.	Purity	Compl.	Purity	Compl.	Purity
Quasar	0.409	0.248	0.838	0.408	0.916	0.240	0.384	0.621
Galaxy	0.831	0.402	0.924	0.298	0.936	0.219	0.826	0.638
Star	0.998	0.989	0.998	1.000	0.996	0.990	–	–
White dwarf	0.491	0.158	–	–	0.432	0.250	–	–
Physical binary star	0.002	0.096	–	–	0.002	0.075	–	–
Quasar, $\lvert \sin b \rvert > 0.2$	0.409	0.442	0.881	0.603	0.935	0.412	0.393	0.786
Galaxy, $\lvert \sin b \rvert > 0.2$	0.830	0.648	0.928	0.461	0.938	0.409	0.827	0.817

Notes. Classification is done by assigning the class with the largest posterior probability. Performance is given in terms of completeness (compl.) and purity, for each classifier and for each class. Purities have been adjusted to reflect the class prior (given in Table 2). Results on the 'binary' class are largely meaningless due to the incongruity of the class definitions in the training and validation data sets. These results reflect performance for sources drawn at random from the entire *Gaia* data set, in particular for all magnitudes and latitudes. The final two columns labelled 'Spec&Allos' refer to samples obtained by requiring a probability larger than 0.5 from both Specmod and Allosmod for a given class: this is identical to `classlabel_dsc_joint` in the `qso_candidates` and `galaxy_candidates` tables. The bottom two rows refer to extragalactic sources at higher Galactic latitudes ($|b| > 11.54°$), where the prior is more favourable for detecting quasars and galaxies. These are conservative estimates, accounting only for reduced numbers of stars, not the better visibility of extragalactic objects on account of less interstellar extinction and source confusion.

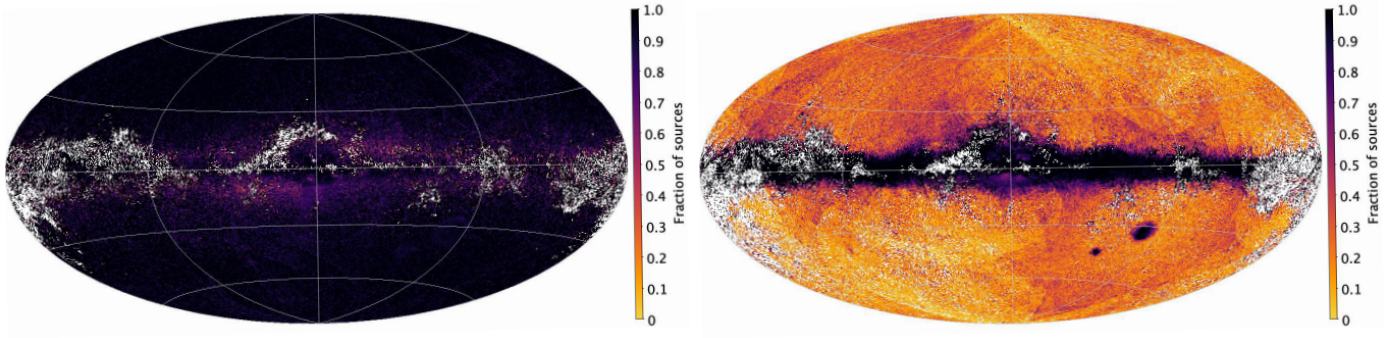

Fig. 4. Galactic sky distribution of the fraction of sources that have 5p/6p astrometric solutions (i.e. have parallaxes and proper motions) for sources that also have `dsc_classlabel='quasar'` (*left*) and `dsc_classlabel='galaxy'` (*right*). The plot is shown at HEALPix level 7 (0.210 deg^2) in a Hammer–Aitoff equal area projection with the Galactic centre in the middle, north up, and longitude increasing to the left. White indicates no sources.

Figure 6 shows the same sky distribution as before, but now expressing the numbers as a fraction of the total number of sources in that HEALPix[5] (classified by DSC as anything). As most of the sources are stars, these plots essentially show the ratio of extragalactic to Galactic objects per HEALPix, albeit with varying degrees of contamination. The four rows of the plot correspond to four possible ways of classifying extragalactic sources: the top three rows are for probabilities above 0.5 for Specmod, Allosmod, and Combmod, respectively, whereby the latter is identical to `classlabel_dsc`. The bottom row is `classlabel_dsc_joint`. Looking at the third row – for `classlabel_dsc` – we see a higher fraction of extragalactic sources (plus contamination) has been discovered outside of the Galactic plane than at lower latitudes. This we expect, as high extinction from Galactic dust obscures extragalactic objects, and also there are far more stars in the Galactic plane. However, we also see a higher fraction of supposed quasars (left) in the LMC and SMC – clear misclassifications – indicating a higher contamination in these regions. Looking at the top two left panels in Fig. 6 for Specmod and Allosmod, respectively, we see that this contamination comes from Specmod, that is, misclassification of the BP/RP spectra, but not from Allosmod, which uses photometry and astrometry. It is probably not due to crowding in the LMC/SMC corrupting the BP/RP spectra, because we do not see such high contamination in the crowded Galactic plane; it is more likely due to faint blue sources in the LMC/SMC being confused with quasars, something which does not occur as much in the Galactic plane due to the higher reddening there.

The top three rows of the right column of Fig. 6 show the corresponding plots for galaxies. The stripes are artefacts of the *Gaia* scanning law. They are much more prominent in Allosmod than in Specmod, and we see in Table 3 that Allosmod is expected to have a lower purity for galaxies than Specmod (the opposite is true for quasars).

When we use `classlabel_dsc_joint` for classification, we get smaller but purer samples (see Gaia Collaboration 2023). The sky distributions for these samples (bottom row of Fig. 6) show that low-latitude regions are excluded. In other words, only sources at higher latitudes were classified with probabilities above 0.5 by both Specmod and Allosmod. We also note that the overdensities in the LMC and SMC are greatly reduced with `classlabel_dsc_joint`.

The middle panels of Fig. 2 show the distributions of various *Gaia* features for the sources classified as quasar (in blue) and galaxy (in orange) by `classlabel_dsc`. The middle panel of Fig. 3 shows the two colours as a colour–colour diagram. These may be compared to the distributions of the training data

[5] For details on the HEALPix scheme used by *Gaia*, see Bastian & Portell (2020).

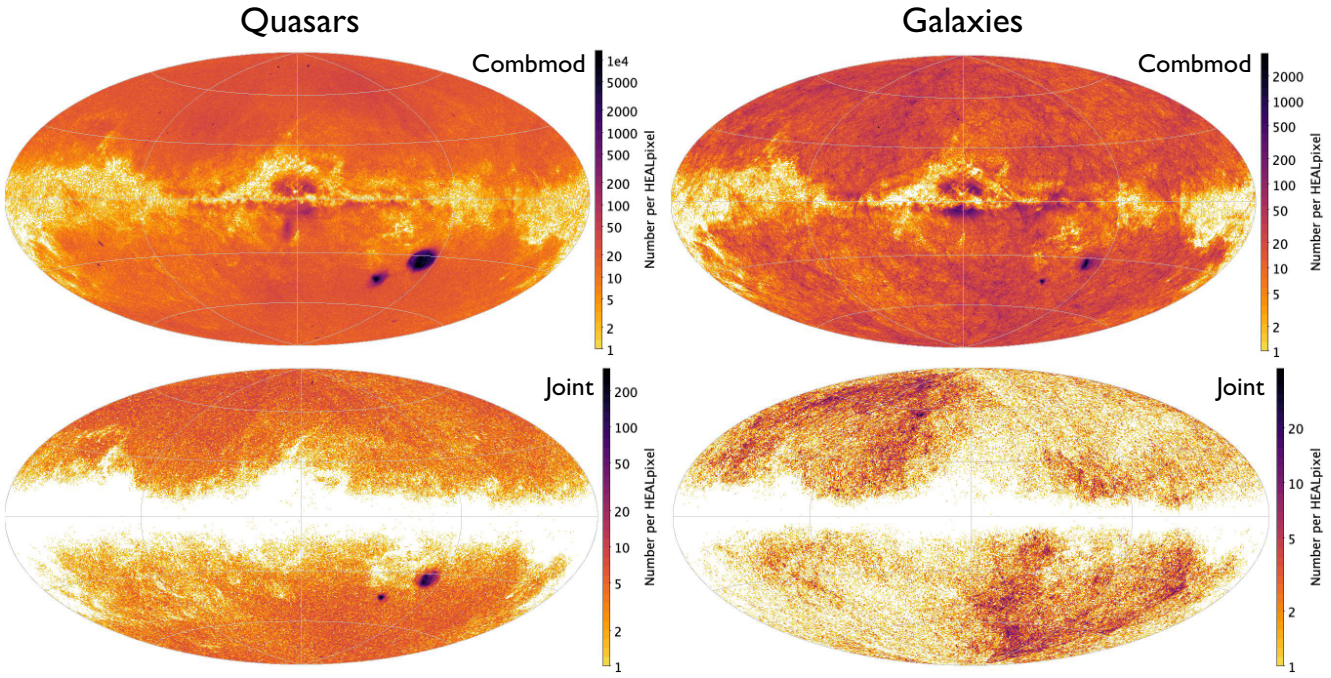

Fig. 5. Galactic sky distribution of the number of DSC sources classified as quasars (*left*) and galaxies (*right*) according to `classlabel_dsc` (*top*) and `classlabel_dsc_joint` (*bottom*) (see Sect. 3.2.4 for the label definition). The plot is shown at HEALPix level 7 (0.210 deg^2). The logarithmic colour scale covers the full range for each panel, and is therefore different for each panel.

in the upper panels in both cases. There are some noticeable differences. The most obvious is the spike in the latitude distribution for (apparent) quasars at the LMC. Recall that, when training Allosmod, we used a flat $\sin b$ distribution (see Sect. 3.2). We also see that the objects classified – galaxies in particular – extend to fainter magnitudes than the training data. This is not surprising given that the training sample had to have SDSS spectroscopic classifications, whereas we apply DSC to all *Gaia* sources, which extend to fainter magnitudes, where misclassifications are more frequent. The observed galaxies also show larger (anomalous) proper motions, plus more (anomalous) photometric variability according to the relative variability, `relvarg`, parameter. Finally, we also see differences in the colour distributions compared to the training data for both classes (Fig. 3). Some of this is due to the different populations being sampled (the training objects are brighter), as well as contamination.

The bottom panels of Figs. 2 and 3 show the features and colour–colour diagrams for objects classified using the purer `classlabel_dsc_joint` label. These show tighter distributions that are more similar to the training data. We note in particular the reduction of faint galaxies.

We now return to the issue of the 5p/6p and 2p solutions. Figure 7 shows the colour–colour diagram for all sources with `classlabel_dsc='quasar'`, excluding those in the regions around the LMC and SMC, for sources with (5p/6p) and without (2p) parallaxes and proper motions. The DSC-Comdmod probabilities for 5p/6p solutions come from both Specmod and Allosmod, whereas for the 2p solutions they only come from Specmod. Of the objects classified here as quasars, 95% have 5p/6p solutions. We see that the 5p/6p solutions are confined to a smaller range of colours than are the 2p solutions. That is, demanding the existence of parallaxes and proper motions yields a slightly different population of objects in colour space. We reiterate the fact that there is significant stellar contamination in the `classlabel_dsc='quasar'` sample as a whole. The (purer) subset defined by `classlabel_dsc_joint='quasar'` has a distribution (not shown) similar to that of the 5p/6p solutions in the bottom left panel of Fig. 7.

Figure 8 shows the colour–colour diagram for the galaxies. Again we see a difference in the colour distribution of the two types of astrometric solution, but now it is the 2p solutions that cover a narrower range of colours. Galaxies are partially resolved by *Gaia*, and their structure can induce a spurious parallax and proper motion in AGIS (which DSC-Allosmod tries to exploit). Many of these astrometric solutions are rejected by AGIS, turning them into 2p solutions, and these sources can only be classified by Specmod. Of the objects classified here as galaxies, 72% have 2p solutions, compared to 5% for the quasars. Thus, the Specmod and Allosmod results reported in *Gaia* DR3 are not for identical populations of objects, because of the different input data requirements of these classifiers.

As Specmod and Allosmod use different data, it is interesting to see how their classification probabilities differ for a common set of sources. We investigate this by selecting sources that have results from both Specmod and Allosmod, and have `classlabel_dsc` set. This is shown for the quasar candidates in the left column of Fig. 9. These plots do not convey the number of sources in each part of the diagram, and should therefore be interpreted with that in mind. Nonetheless, although we see regions where Specmod and Allosmod have similar probabilities, there are also regions where their probabilities are quite different. Because `classlabel_dsc_joint` is only set to 'quasar' when both Specmod and Allosmod probabilities are above 0.5, these figures explain why that set is comparatively small. The right column of Fig. 9 shows the same for the galaxy candidates, and again we see a significant lack of correlation between Specmod and Allosmod. This shows that the different data used by these two classifiers convey rather different information.

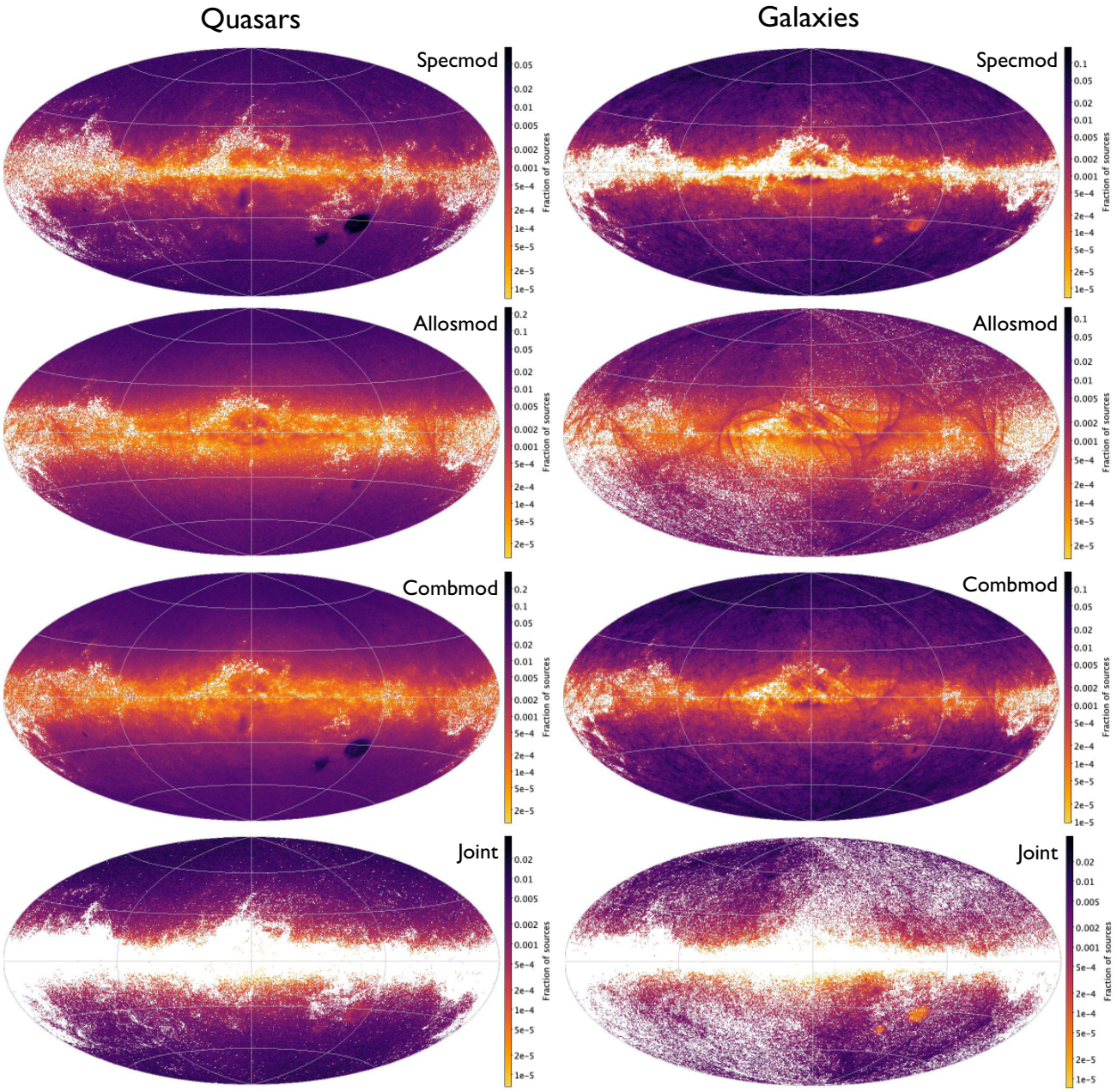

Fig. 6. Galactic sky distribution of the fraction of DSC sources classified as quasars (*left*) and galaxies (*right*) according to Specmod (*top*), Allosmod (*second*), Combmod (*third*), and Specmod and Allosmod (*bottom*) probabilities being greater than 0.5 for that class. *Bottom two rows* are identical to `classlabel_dsc` and `classlabel_dsc_joint` (respectively) being set to the appropriate class (see Sect. 3.2.4). The plot is shown at HEALPix level 7 (0.210 deg^2) with each cell showing the ratio of the sources classified to the total number of sources with DSC results (1.59 billion over the whole sky). The logarithmic colour scale covers the full range for each panel, and is therefore different for each panel.

3.5. Use of DSC results

The DSC class probabilities exist primarily to help users identify quasars and galaxies. The performance on white dwarfs and binaries is rather poor. These probabilities will be of limited use to the general user and we do not recommend their use to build samples. One could add these probabilities to the star probability for each source, and thereby end up with a three-class classifier.

Classification can be done by selecting sources with class probabilities above a given threshold. A threshold of 0.5 gives a selection (and performance) very similar to what would be obtained when taking the maximum probability. A threshold of 0.5 applied to the Combmod outputs is identical to the `classlabel_dsc` label (Sect. 3.2.4). With this choice of threshold, the purities for galaxies and quasars are rather modest, as we can see from Table 3. This is unsurprising, because with a threshold of 0.5 we expect up to half of the objects to be incorrectly classified even with a perfect classifier. Increasing the threshold does increase the purity at the cost of decreased completeness, but because the DSC probabilities tend to be rather extreme (see plots in Bailer-Jones 2021), this does not help as much as one might hope. The fact that the purities are often lower than the limit expected from the threshold may be due not only to an imperfect

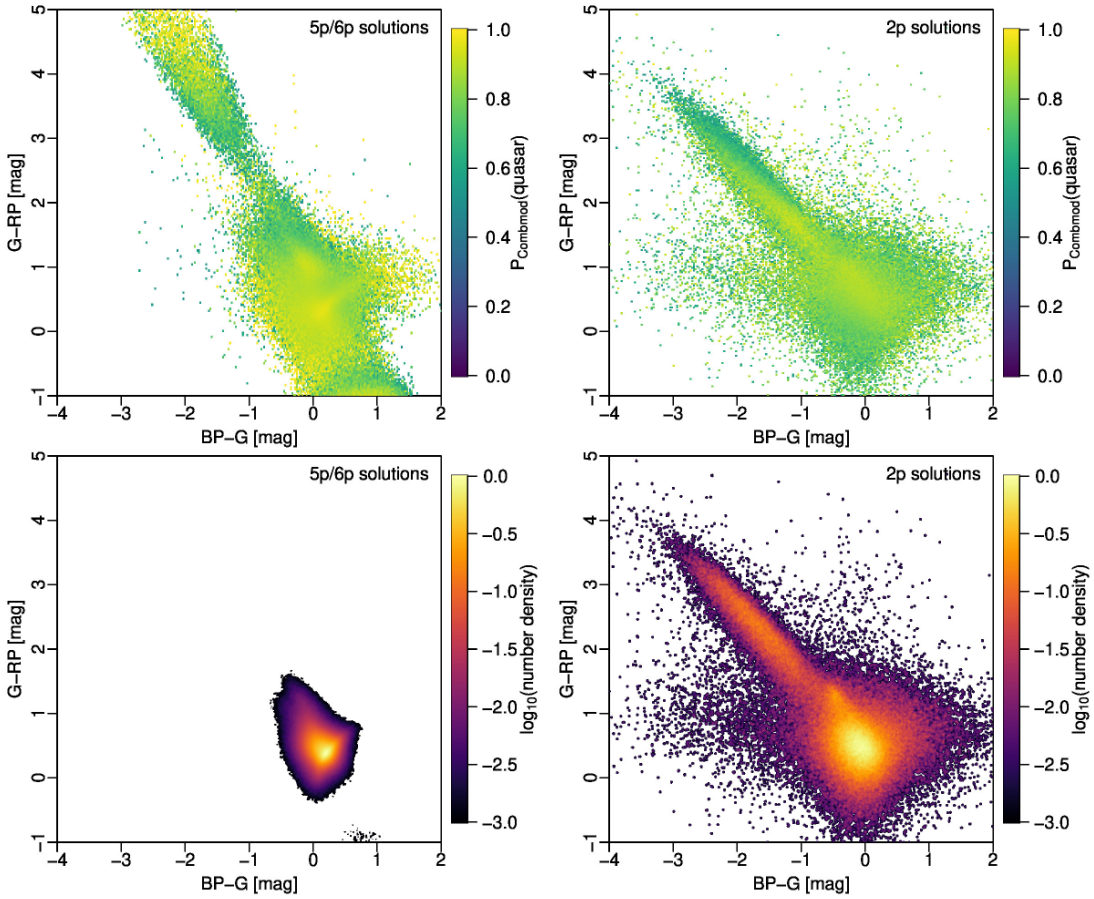

Fig. 7. Colour–colour diagram for sources in the `qso_candidates` table with `classlabel_dsc='quasar'`, excluding regions around the LMC and SMC. The left column shows sources with 5p/6p solutions (2.64 million sources), the right column shows sources with 2p solutions (0.14 million sources). These numbers refer to plotted sources, i.e. that have all *Gaia* bands. The colour coding in the *upper panel* shows the mean DSC-Combmod probability for the quasar class (the field `classprob_dsc_combmod_quasar`). The colour coding in the *lower panel* shows the density of sources on a log scale relative to the peak density in that panel.

classifier, but also to an imperfect calibration of the probabilities in Specmod and Combmod (although not Allosmod)[6].

The DSC completenesses, especially with Combmod, are quite good, but the purities are rather modest, as discussed earlier. This is a consequence of primarily two factors.

The first factor is the intrinsic rareness of the quasars and galaxies. If only one in every thousand sources were extragalactic, then even if our classifier had 99.9% accuracy, the resulting sample would only be around 50% pure. This is the situation we have: the intrinsic ability of DSC to separate the classes is actually very good, with purities of the order of 99% on balanced test sets. However, when it is then applied to a randomly selected set of *Gaia* data there are so many stars that even though a small fraction of these are misclassified, this is still a large number. We cannot overcome this problem by adopting a different prior. If we used uniform priors, for example, this would classify many more sources – both true and false – as extragalactic. This would increase the completeness of this class. It is not immediately obvious what happens to the purity, but Bailer-Jones et al. (2019) found that for Allosmod in *Gaia* DR2, the purities for quasars and galaxies were actually significantly reduced.

The extreme rareness of the extragalactic objects places high demands on the classifiers, and the performance may be limited by the second factor, namely the ability of the data to distinguish between the classes. We experimented with using different or additional *Gaia* features (e.g. colour excess factor) as inputs to Allosmod, but this did not help. Performance might improve if we define synthetic filters from the BP/RP spectra instead of using the entire spectrum, or by generating other features from the *Gaia* data, but this has not been explored[7]. The inclusion of non-*Gaia* data, such as infrared photometry, should help but was beyond the scope of the activities for *Gaia* DR3.

A third potential limiting factor is the set of training examples we use. Although the SDSS spectroscopic classifications are believed to be very good, they may have errors, and they may also not provide the clearest distinction between galaxies and quasars.

The fact remains that the classification performance depends unavoidably on the intrinsic rareness, that is, on the prior. Users may want to adopt a different prior from ours (Table 2), which would be particularly appropriate if they focus on a subset of parameter space. To recompute the DSC probabilities with a new prior we do not need to re-train or re-apply DSC. The fact

[6] The issue of expected sample purity is discussed in Sect. 5.2 of Bailer-Jones et al. (2008). Even with an imperfect classifier, it is possible to infer the expected number of true sources from the inferred numbers by inverting the confusion matrix, as shown by Bailer-Jones et al. (2019).

[7] One obvious example is to compute the absolute magnitude, because this together with colour – i.e. the HRD – clearly separates out white dwarfs when the parallax uncertainties are not too large.

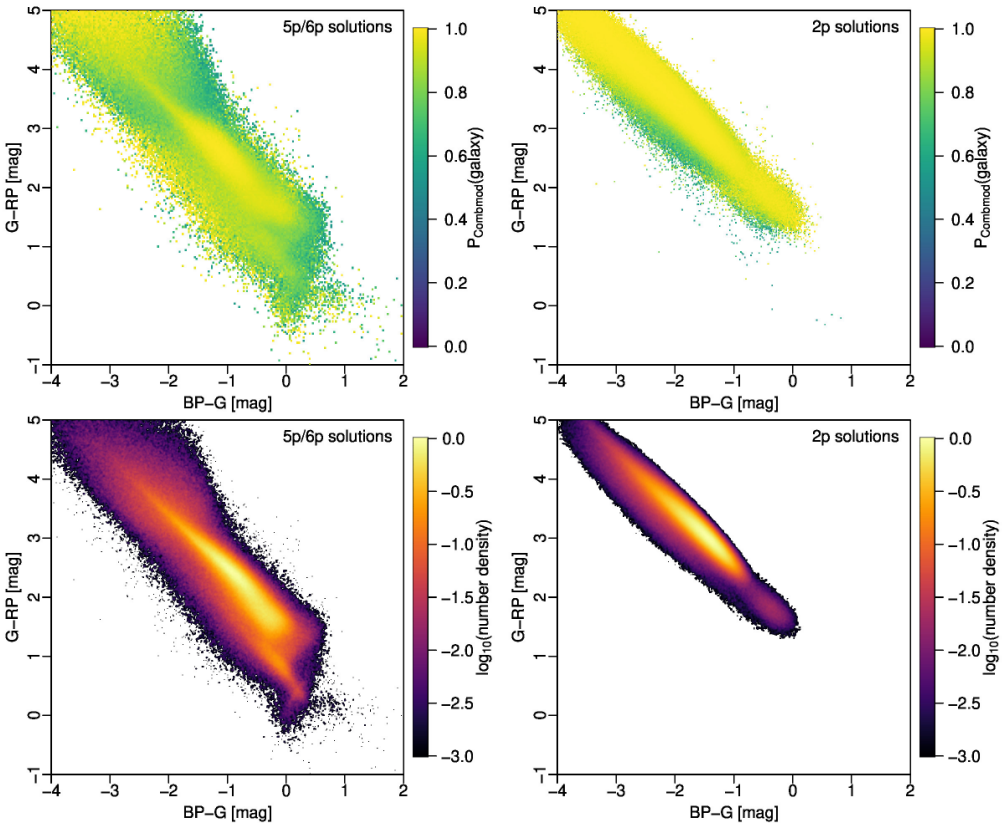

Fig. 8. As in Fig. 7 but for sources in the `galaxy_candidates` table with `classlabel_dsc='galaxy'`, excluding regions around the LMC and SMC. The left column shows sources with 5p/6p solutions (0.91 million sources), and the right column shows sources with 2p solutions (2.32 million sources). These numbers refer to plotted sources, i.e. that have all *Gaia* bands.

that DSC provides posterior probabilities as outputs makes it simple to strip off our prior and apply a new one, as shown in Appendix C.

It is important to realise that the performances in Table 3 are (a) only for the classes as defined by the training data and (b) an average over the entire *Gaia* sample, and are therefore dominated by faint sources with lower quality data. Our galaxy class in particular is a peculiar subset of all galaxies, because *Gaia* tends not to observe extended objects, and even then may not measure them correctly (see Sect. 3.2).

DSC misclassifies some very bright sources that are obviously not extragalactic, for example. As these are easily removed by the user, we chose not to filter the DSC results in any way. One may likewise wonder why there are some objects classified as quasars with statistically significant proper motions. We do use proper motion as a classification feature, but in a continuous fashion, not as a hard cut. A more conservative approach to classification is to apply a series of necessary conditions, that is, a simple decision tree. This could increase the purity – and could be tuned to guarantee that certain known objects come out correctly – but at the expense of completeness. We do nevertheless provide the class label `classlabel_dsc_joint` as a means to select a purer subsample of extragalactic sources (Sect. 3.2.4), as can be seen from the last two columns of Table 3.

4. Outlier analysis (OA)

4.1. Objectives

The Outlier Analysis (OA) module aims to complement the overall classification performed by the DSC module, by processing those objects with lower classification probability from DSC (see Sect. 3). OA is intended to analyse abnormal or infrequent objects, or artefacts, and was applied to all sources that received DSC Combmod probabilities below 0.999 in all of its five classes. This threshold was chosen so as to process a limited number of 134 million sources, corresponding to about 10% of the total number of sources for which DSC produced probabilities. Subsequently, a selection of the sources to be processed is carried out based on several quality criteria, the most restrictive being that the mean spectra correspond to at least five transits (see details in the online documentation). The resulting filtering leads us to process a total of 56 416 360 sources. Such sources tend to be fainter and/or have noisier data. For these objects, OA provides an unsupervised classification – where the true object types are not known – that complements the one produced by DSC, which follows a supervised approach based on a set of fixed classes.

4.2. Method

The method used by OA to analyse the physical nature of classification outliers is based on a self-organising map (SOM, Kohonen 1982), which groups objects with similar BP/RP spectra (see Sect. 4.2.1) according to a Euclidean distance measure. The SOM performs a projection of the multidimensional input space of BP/RP into a two-dimensional grid of size 30×30, which facilitates the visual interpretation of clustering results. Such a projection is characterised by its preservation of the topological order, in the sense that, for a given distance metric, similar data in the input space will belong to the same or to neighbouring neurons in the output space. Each one of these neurons

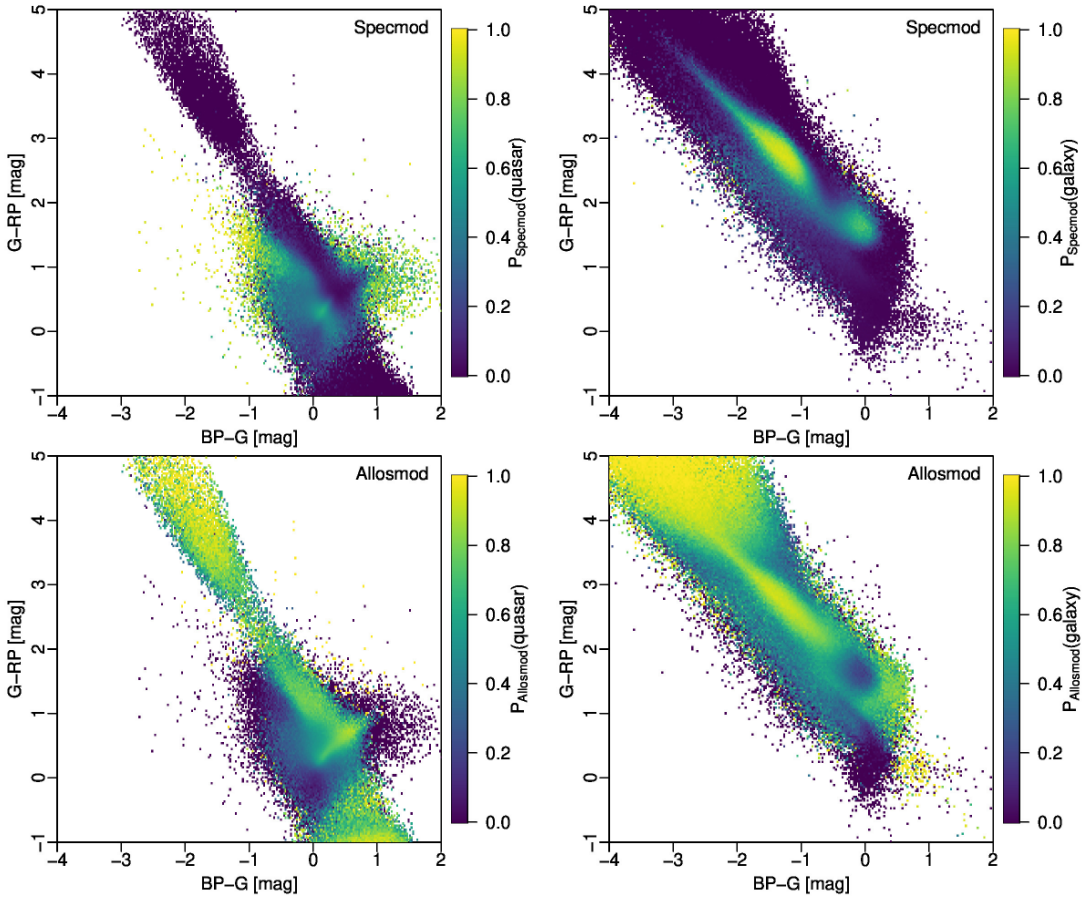

Fig. 9. Colour–colour diagram for sources in the `qso_candidates` table with `classlabel_dsc='quasar'` (left) and in the `galaxy_candidates` table with `classlabel_dsc='galaxy'` (right), in both cases excluding regions around the LMC/SMC, that have both Specmod and Allosmod results. *Upper and lower panels*: the mean DSC-Specmod probability and the mean DSC-Allosmod probability, respectively, for a common sample.

has a prototype, which is adjusted during the training phase and that best represents the input spectra that are closest to this neuron. In *Gaia* DR3, each prototype is the average spectrum of the pre-processed[8] BP/RP spectra of the sources assigned to that particular neuron, which correspond to those closest to the neuron according to the Euclidean distance between the neuron prototype and the pre-processed BP/RP spectrum of the source. Neuron prototypes are reported in the `oa_neuron_xp_spectra` table. A centroid is also identified for each neuron, which is the source whose pre-processed BP/RP spectrum is the closest to the prototype of the neuron, according to the Euclidean distance. Centroids can be found in the `centroid_id` field of the `oa_neuron_information` table along with statistics of the main Gaia observables for the sources belonging to this neuron: G, G_{BP}, and G_{RP} magnitudes, proper motions, Galactic latitude, parallax, number of BP/RP transits, renormalised unit weight error (`ruwe`), BP/RP flux excess factor, and $G_{BP} - G_{RP}$ colour.

4.2.1. BP/RP spectra preprocessing

The sampled mean BP/RP spectra produced by SMSgen are transformed in order to remove artefacts, and to improve the clustering produced by the SOMs: (a) Pixels with negative or zero flux values are linearly interpolated, provided that they do not affect more than 10% of the effective wavelength in a consecutive manner or more than 25% of the entire effective wavelength. Such a filtering was imposed because most of the spectra that did not meet such criteria were usually of low quality and had a low number of transits. These filtered spectra are not analysed; (b) BP and RP spectra are downsampled to 60 pixels each; (c) both spectra are trimmed to avoid the low transmission regions of the CCD, so that OA uses the effective wavelength ranges 375–644 nm for BP and 644–1050 nm for RP; (d) spectra are concatenated to obtain a single spectrum; and, (e) the joint spectrum is normalised so that the sum of its flux is equal to one.

4.2.2. Quality assessment

The performance of OA cannot be measured through metrics such as completeness and purity because of the unsupervised nature of the technique. Therefore, a descriptive approach based on the intra-neuron and inter-neuron distances (Álvarez et al. 2021) was followed in order to analyse the quality of the clustering. We decided to use the squared Euclidean distance as a proxy for distance because the SOM algorithm uses it as a measurement of mean quantisation error for processing elements. The intra-neuron distance of each source is then computed as the squared value of the Euclidean distance between the source and the prototype of the neuron it belongs to, whereas the inter-neuron distance is computed as the squared Euclidean distance

[8] The OA pre-processing of BP/RP spectra is later described in Sect. 4.2.1.

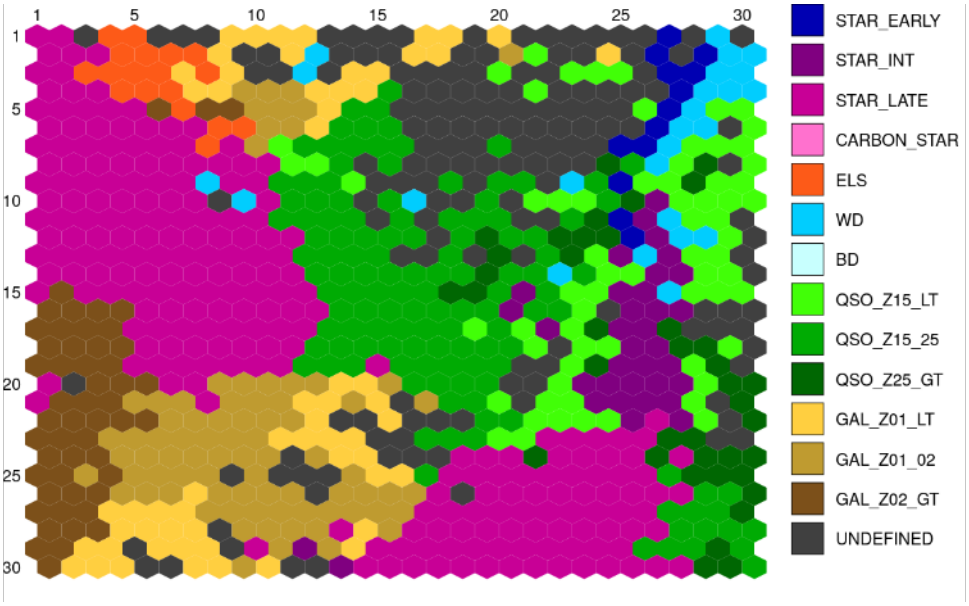

Fig. 10. SOM grid from the OA module visualised through the GUASOM tool (Álvarez et al. 2021). Each cell corresponds to a neuron from the SOM, most of which were assigned a class label. Those neurons that did not meet the quality criteria defined to establish a class label remain 'undefined', as explained in Sect. 4.2.3

between two different neuron prototypes. In order to assess the quality of the clustering, we selected the three parameters that we thought best describe the distribution of the intra-neuron distances: (a) the width of the distribution according to the value of the full width at half maximum ($FWHM$); (b) the skewness (S), which measures its asymmetry; and, (c) the kurtosis excess (K), which measures the level of concentration of distances. A high-quality clustering will result from neurons with low values of the $FWHM$ parameter, and large positive values of both skewness and kurtosis. Finally, in order to facilitate the interpretation of such quality measurements, a categorical index named QC was derived based on the values obtained for S, K, and a normalised version of $FWHM$ (which is reversed in order for the higher quality neurons to take larger values). To this purpose, seven quality categories were established, according to the values taken by such parameters with respect to six arbitrarily chosen percentiles (95th, 90th, 75th, 50th, 32th, and 10th), which are computed independently for each one of the parameters listed above over the entire map. For each neuron, we determine the lowest percentile in which the three parameters are above their respective percentile values. Thus, if a value is above the 95th percentile, then QC will take the value of zero; if it is in the 90th percentile, then QC will correspond to category one, and so on up to category six, which will correspond to those neurons whose poorest quality indicator is outside the lowest percentile that has been considered, 10th. Accordingly, the best-quality neurons will have $QC = 0$ and the worst ones $QC = 6$. It should be emphasised here that QC only assesses the quality of the clustering (i.e. how closely the pre-processed BP/RP spectra in a neuron match their prototype) compared to the overall intra-neuron distances, such that no assumption should be made on the quality of the spectra they contain, nor on the labelling of the individual neurons described below.

4.2.3. Neuron labelling

Unsupervised methods do not directly provide any label to the samples that are being analysed. For this reason, a set of reference BP/RP spectra templates for prototypical astronomical objects was built by taking into account validation sources from the various Apsis modules (see the online documentation). These reference templates are used to label the neurons in *Gaia* DR3 by identifying the closest template to the neuron prototype according to the Euclidean distance. In addition, to guarantee the suitability of the assigned templates (and class labels), two conditions were imposed: (a) the squared Euclidean distance between a template and the neuron prototype must not exceed a threshold of 3.58×10^{-2}; and, (b) the neuron must have $QC < 6$. Figure 10 shows the SOM built by OA for *Gaia* DR3, where around 80% of the neurons were assigned a template, and hence a class label. The limit of 3.58×10^{-2} on the squared distance was set during the template-building process and is detailed in the online documentation.

4.2.4. GUASOM visualisation tool

To help the user to analyse and visualise the clustering results, we designed an application called Gaia Utility for the Analysis of Self-Organising Maps (GUASOM) (Álvarez et al. 2021). It can be run over the internet, and contains several visualisation utilities that allow an interactive analysis of the information present on the map. The tool provides both classical and specific domain representations such as U-matrix, hits, parameter distributions, template labels, colour distribution, and category distribution.

4.3. Performance and results

OA processed 56 416 360 objects in *Gaia* DR3. Figure 11 displays their G magnitude distribution, demonstrating that OA covers a wide range of G magnitudes with a significant fraction of faint objects.

Figure 12 shows the histogram of neuron quality categories, QC, where the total number of sources belonging to such neurons is superimposed. Approximately 35% of the neurons have $0 \leq QC \leq 3$ and are hence referred to as 'high-quality neuron': these comprise around 55% of the sources processed. The rest of the neurons can be considered as low-quality neurons. Figure 13 shows how the quality categories are distributed over the SOM.

It is worth mentioning that the SOM does not directly label neurons, nor does it provide quality measurements on the clustering they produce, which means that we have to apply the

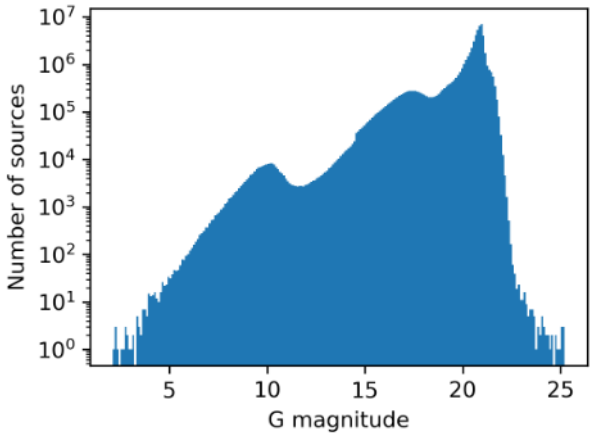

Fig. 11. *G* mag distribution of the 56 416 360 sources processed by the OA module in *Gaia* DR3 (bin width of 0.1).

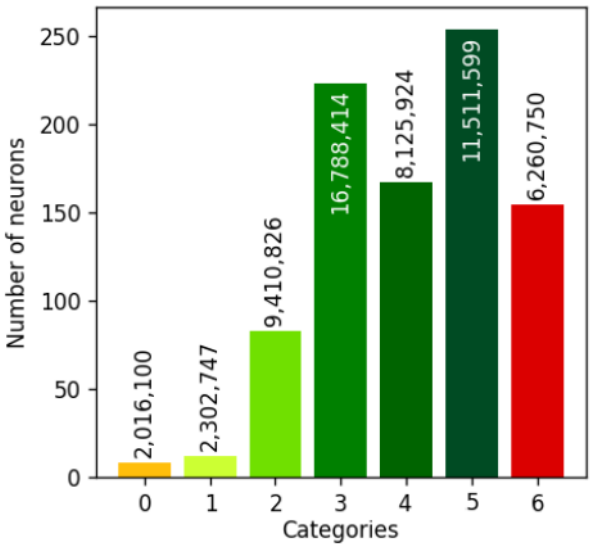

Fig. 12. Histogram of neuron quality categories for the sources processed by the OA in *Gaia* DR3. The number of sources per category is superimposed along with the bars. Those neurons with $0 \leq QC \leq 3$ are considered high-quality neurons.

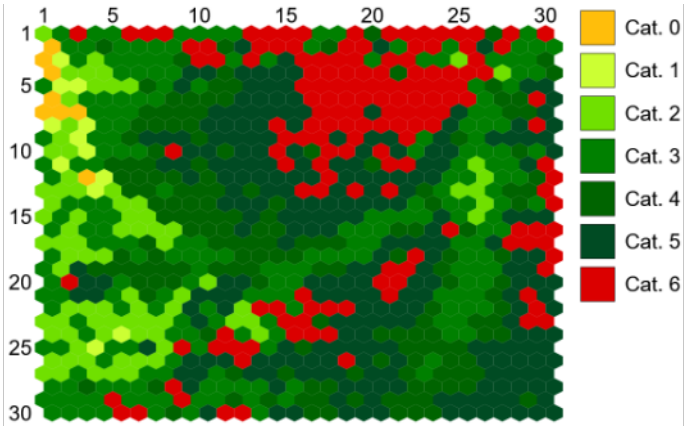

Fig. 13. SOM grid visualised through the GUASOM tool (Álvarez et al. 2021) to represent the quality category (*QC*) assigned to each neuron.

procedures described in Sects. 4.2.2 and 4.2.3 after we build the map. As a result, Fig. 13 shows the quality category associated with each neuron in our grid of 30×30 neurons. These quality categories assess how well the sources fit to the prototype of the neuron they belong to: neurons with the lowest quality category are composed of sources whose spectra are the most homogeneous (i.e. neurons of highest quality). Similarly, in Fig. 10, the label assigned to each neuron provides a hint as to the astronomical type of the sources they contain. Comparing Figs. 10 and 13, we can see that high-quality neurons mostly correspond to stars and galaxies, while quasars are usually associated with low-quality neurons. The reason for this mostly stands in the wide range of cosmological redshifts that is observed amongst those objects, in their different continuum shapes and emission-line equivalent widths.

Table 4 represents the contingency table between DSC Combmod and OA class labels. DSC labels are determined according to the class with the highest DSC Combmod probability, except for those that take a probability below 0.5, which are labelled as 'unknown'. Sources with DSC 'binary star' class are considered as 'star' as the former class is not present in OA. Similarly, OA class labels are aggregated into more generic ones in order to enable comparison with the DSC class labels. Recalling that OA only processes sources with all DSC Combmod probabilities below 0.999, the OA results can be summarised as follows.

- Galaxies: There is close agreement for galaxies, as around 80% of the galaxies identified by DSC are also confirmed by OA.
- Quasars: The agreement with DSC decreases to 35%. A large fraction of those quasars identified by DSC are considered as stars or white dwarfs by OA.
- Stars: Around 40% of those identified by DSC were also confirmed by OA. However, a large fraction of them were considered as extragalactic objects by OA.
- White dwarfs: In this case, the agreement between both modules is around 50%. Most of the remaining objects are considered as stars by OA.

Around 11% of the sources are assigned to a neuron that was not labelled by OA because of their poor quality (category six). In particular, approximately 2510 sources could not be classified by OA and have `classlabel_dsc = 'unclassified'`, meaning that studying their nature may require a deeper analysis.

4.4. Use of OA clustering

The analysis performed by the OA module can be useful for different purposes. For instance, high-quality neurons can help to assess the physical nature of some sources with DSC combmod probabilities below the chosen threshold (0.999) in all classes or to identify objects that were potentially misclassified. As OA provides an unsupervised classification based on a normalised SED comparison, for a given neuron there are sources with different degrees of similarity to the prototype. For that reason, we encourage the user to isolate clean samples for each neuron through the quality measurements provided in the online documentation. In particular, we suggest combining both the categorical quality index (*QC*) and the classification distance in order to retrieve the best classified sources from OA. Table 5 shows the number of sources per class that are assigned to a high-quality neuron (from category zero to three), and whose classification distance between the pre-processed BP/RP spectrum of the source and the neuron prototype is below 0.001 (i.e. what

Table 4. Contingency table between DSC taken from predominant probabilities produced by DSC Combmod and OA classifications, grouped into generic types.

		OA class label					
		STAR	WD	QSO	GAL	UNDEFINED	Total
DSC	STAR	40%	3%	22%	24%	11%	53 295 527
	WD	42%	51%	3%	0%	4%	92 186
	QSO	29%	21%	35%	2%	13%	2 158 916
	GAL	4%	0%	9%	83%	4%	851 127
	UNKNOWN	22%	7%	35%	22%	13%	18 604
	Total	21 763 876	2 240 195	12 680 763	13 470 776	6 260 750	

Notes. Unknown means that the DSC predominant probability was below 0.5, whereas for OA it means that no template was assigned due to quality constraints. Fractions are computed with respect to the total number of sources in each DSC class.

Table 5. Number of sources in each OA class that belong to a high-quality neuron while having a classification squared Euclidean distance below 0.001 (i.e. what we consider here as reliable).

Class label	Number of sources
STAR_LATE	8 966 955
GAL_Z01_02	3 917 749
STAR_INT	3 158 041
GAL_Z02_GT	2 952 297
GAL_Z01_LT	2 355 895
WD	1 561 204
QSO_Z15_LT	1 138 832
QSO_Z15_25	1 020 337
STAR_EARLY	914 470
ELS	489 551
QSO_Z25_GT	92 460

Notes. There may be considerable contamination in these class assignments.

we consider here as reliable predicted classes). As can be seen, around 13 million stars, 9 million galaxies, 2 million quasars, and 1.5 million white dwarfs meet these criteria.

5. Quasar classifier (QSOC)

5.1. Objectives

The quasar classifier (QSOC) module is designed to determine the redshift, z, of the sources that are classified as quasars by the DSC module (see Sect. 3 for more details). In order to produce redshift estimates for the most complete set of sources, we considered a very low threshold on the DSC quasar probability of `classprob_dsc_combmod_quasar` $\geq$0.01, meaning that we expect a significant fraction of the processed sources to be stars or galaxies. Users interested in purer sub-samples may then require that `classlabel_dsc_joint = 'quasar'`, as explained in Sect. 3.2.4, or may use more sophisticated filtering, as explained in Gaia Collaboration (2023, Sect. 8).

5.2. Method

5.2.1. Overview

QSOC is based on a χ^2 approach that compares the observed BP/RP spectra sampled by SMSgen (see Creevey et al. 2023, and the online documentation) to quasar rest-frame templates in order to infer their redshift. The predicted redshifts take values in the range $0.0826 < z < 6.12295$. As the effective redshift is not necessarily the one associated with the minimal χ^2 (see Sect. 5.2.3), it is complemented by an indicator of the presence of quasar emission lines (Z_{score} from Eq. (6)) and these are converted into a redshift score, S, from Eq. (7). For a given source, the redshift with the highest score is then the one that is selected by the algorithm. Quasar templates are described in Sect. 5.2.2 while the redshift determination algorithm is described in Sect. 5.2.3.

5.2.2. Quasar templates

The quasar templates used by QSOC were built based on the method described in Delchambre (2015) and applied to 297 264 quasars[9] from the twelfth release of the Sloan Digital Sky Survey Quasar catalogue of Pâris et al. (2017, DR12Q). These spectra are first extrapolated to the wavelength range of the *Gaia* BP/RP spectro-photometer (i.e. 300–1100 nm) with a linear wavelength sampling of 0.1 nm using a procedure similar to the one used by Delchambre (2018). They are subsequently converted into BP/RP spectra through the use of the BP/RP spectrum simulator provided by CU5 and described in Montegriffo et al. (2023). An artificial spectrum with a uniform SED (i.e. of constant flux density per wavelength) was also converted through the BP/RP spectrum simulator in order to produce the so-called 'flat BP/RP spectrum'. We then divided each simulated BP/RP spectrum by its flat counterpart before subtracting a quadratic polynomial that is fitted to the observations in a least absolute deviation sense (i.e. ℓ-1 norm minimisation), leaving pure emission line spectra. We note that, in order to avoid fitting emission lines, a second-order derivative of the flux density was estimated around each sampled point, $\mathrm{d}^2 f_i/\mathrm{d}\lambda_i^2$, and later used to scale the associated uncertainties by a factor of $\max\left(\left|\mathrm{d}^2 f_i/\mathrm{d}\lambda_i^2\right|/M, 0.01\right)$, where M is a normalisation factor equal to the maximal absolute value of the second-order derivatives evaluated over all the sampled points. As the continuum regions often have very low curvatures compared to the emission lines, they are usually overweighted by a factor of up to 100 in the ℓ-1 norm minimisation. A logarithmic wavelength sampling of $\log L = 0.001$ was then used for both the BP and RP templates, ensuring that the resolution of the BP/RP spectra, as sampled by SMSgen, is preserved.

[9] We note that for 37 of the 297 301 quasars originally contained in the DR12Q catalogue, the ℓ-1 norm fit of the continuum to the observed spectrum (later described) did not converge and these were accordingly not included in the final sample we used.

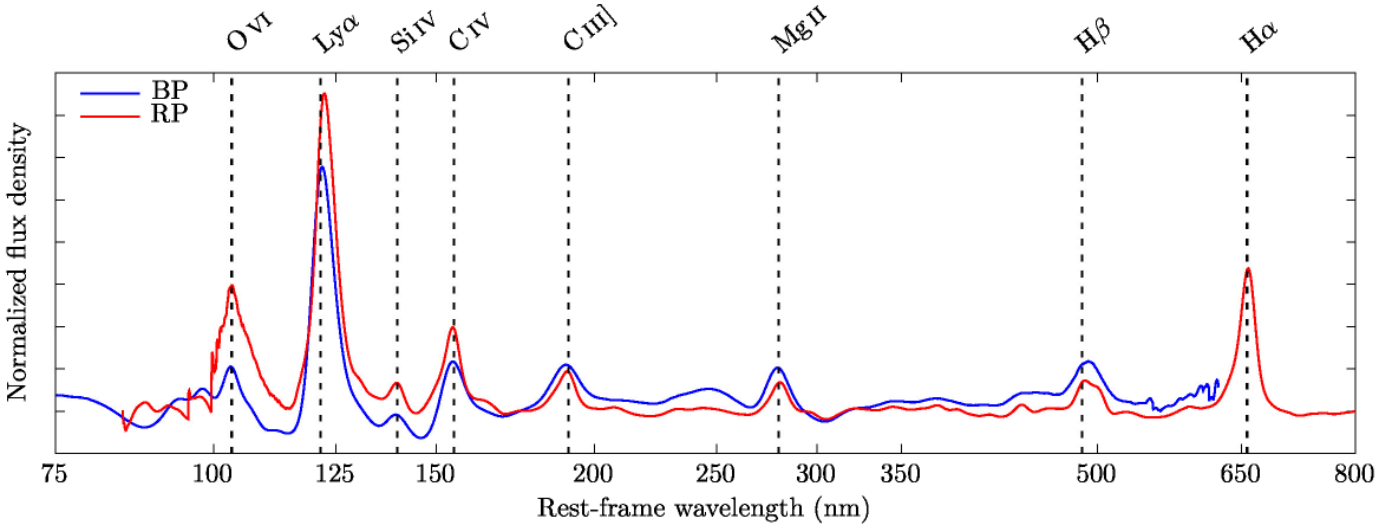

Fig. 14. Rest-frame quasar templates used by QSOC. These correspond to the dominant templates taken over the 32 templates that are computed based on the method described in Delchambre (2015) and applied to 297 264 quasars from the DR12Q catalogue that are converted into BP/RP spectra through the use of the BP/RP spectrum simulator provided by CU5.

We extracted 32 BP/RP templates based on these 297 264 simulated spectra using the weighted principal component analysis method described in Delchambre (2015); nevertheless, only the dominant BP/RP templates – corresponding to the mean of the weighted principal component analysis method – were used because cross-validation tests performed on the simulated spectra show that a larger number of templates significantly increases the degeneracy between redshift predictions.

The resulting templates, illustrated in Fig. 14, closely match the typical composite spectra of quasar emission lines (see e.g. Gaia Collaboration 2023, Sect. 7), although they are convolved by the *Gaia* line spread function which is averaged over the entire set of rest-frame wavelengths. The templates cover the rest-frame wavelength range from 45.7 nm to 623.3 nm in BP and from 84.6 nm to 992.3 nm in RP. These limits, along with the observed wavelength coverage imposed by SMSgen of 325–680 nm in BP and 610–1050 nm in RP allow QSOC to predict redshifts in the range $0.0826 < z < 6.1295$[10].

5.2.3. Algorithm

The determination of the redshift of quasars by QSOC is based on the fact that the redshift, z, turns into a simple offset once considered on a logarithmic wavelength scale:

$$Z = \log(z+1) = \log \lambda_{\rm obs} - \log \lambda_{\rm rest}, \tag{1}$$

where we assume that a given spectral feature located at rest-frame wavelength $\lambda_{\rm rest}$ is observed at wavelength $\lambda_{\rm obs}$. Consider such a logarithmic sampling $\lambda_i = \lambda_0 L^i$, where λ_0 is a reference wavelength and L is the logarithmic wavelength sampling we use, here $\log L = 0.001$ (or $L \approx 1.001$). Then for a given set of n rest-frame templates, T, and an observation vector, $\boldsymbol{s}$, which are both logarithmically sampled with L, the derivation of the optimal shift, k, between T and $\boldsymbol{s}$ can be formulated as a χ^2 minimisation problem through

$$\chi^2(k) = \sum_i \frac{1}{\sigma_i^2} \left(s_i - \sum_{j=1}^{n} a_{j,k} T_{i+k,j} \right)^2, \tag{2}$$

where σ_i is the uncertainty on s_i and $a_{j,k}$ are the coefficients that enable the fit of T to $\boldsymbol{s}$ in a weighted least squares sense while considering a shift k that is applied to the templates. The redshift that is associated with the shift k is then given by $z = L^k - 1$. A continuous estimation of the redshift is then obtained by fitting a quadratic polynomial to $\chi^2(k)$ in the vicinity of the most probable shift.

Despite its appealing simplicity, Eq. (2) is known to have a cubic time complexity on N, as shown in Delchambre (2016), where N is the number of samples contained in each template. In the same manuscript, it is shown that the computation of the cross-correlation function (CCF), defined as

$$\mathrm{ccf}(k) = \left(\sum_i \frac{s_i^2}{\sigma_i^2} \right) - \chi^2(k) = C - \chi^2(k), \tag{3}$$

requires only $O(N \log N)$ floating point operations. Furthermore, given that C is independent of the explored shift, k, maximising $\mathrm{ccf}(k)$ is equivalent to minimising $\chi^2(k)$.

However, some features of the BP/RP spectra complicate the computation of the CCF. First, the BP and RP spectra are distinct such that the effective CCF is actually composed of the sum of two CCFs associated with the BP and RP spectra and templates, $\mathrm{ccf}_{\rm bp}(k)$ and $\mathrm{ccf}_{\rm rp}(k)$, respectively:

$$\mathrm{ccf}(k) = \mathrm{ccf}_{\rm bp}(k) + \mathrm{ccf}_{\rm rp}(k). \tag{4}$$

Secondly, the BP/RP spectra have bell shapes (i.e. their flux smoothly goes to zero at the borders of the spectra), and have spectral flux densities that are integrated over wavelength bins of different sizes, as explained in Creevey et al. (2023). Equation (3) is therefore not directly applicable to these spectra. In order to overcome these difficulties, we divided each BP/RP spectrum by the previously mentioned flat BP/RP spectrum (i.e. BP/RP spectrum coming from a constant flux density and converted through the BP/RP spectrum simulator) and updated their uncertainties accordingly. This solution enables us to solve both

[10] As the cross correlation function computed by QSOC is extrapolated by $\pm \log L$ at its border, the range of the QSOC redshift predictions is slightly wider than one would expect from a straight comparison of the observed and rest-frame wavelengths.

Table 6. The QSOC parameters used to compute the redshift score of quasars from Eq. (7) and the $Z_{\rm score}$ from Eq. (6).

Parameters of the redshift score								
$w_0 = 0.71413$			$w_1 = 0.28587$			$p = 0.24365$		
Parameters of the $\mathbf{Z_{score}}$ for BP spectra								
	O IV	Lyα	Si IV	C IV	C III]	Mg II	Hγ	Hβ
$\boldsymbol{\lambda}$ [nm]	103.202	121.896	139.349	154.658	189.957	279.259	437.904	491.899
$\boldsymbol{I_\lambda}$	0.017	1.0039	0.01	0.13202	0.31359	0.94396	0.23848	0.93124
Parameters of the $\mathbf{Z_{score}}$ for RP spectra								
	O IV	Lyα	Si IV	C IV	C III]	Mg II	Hγ	Hβ	Hα
$\boldsymbol{\lambda}$ [nm]	103.353	122.388	139.563	154.588	190.398	280.470	435.600	488.952	657.736
$\boldsymbol{I_\lambda}$	0.062484	0.10984	0.18982	0.07023	0.1409	0.22011	0.4101	0.25137	0.59948

Notes. The rest-frame wavelengths, λ, of each emission line were retrieved from the quasar templates described in Sect. 5.2.2. Theoretical emission line intensities, I_λ, and score parameters, w_0, w_1, and p, were computed based on a global optimisation procedure that is designed to maximise the score of the redshift predictions with $|\Delta z| < 0.1$ amongst 88 196 randomly selected sources with a redshift estimate from DR12Q. We note that another set of 89 839 observations was then kept as a test set, though the two sets provide a similar distribution of scores.

the bell shape issue and the varying wavelength size of each pixel, passing from units of flux to units of flux density. Finally, most of the quasar flux resides in its continuum, which we model here as a second-order polynomial, concatenated to the set of templates, T, and subsequently fitted to the observations in Eq. (3).

As highlighted in Delchambre (2018), the global maximum of the CCF may not always lead to a physical solution as, for example, some characteristic emission lines of quasars (e.g. Lyα, Mg II, or Hα) may be omitted from the fit while some emission lines can be falsely fitted to absorption features. This global maximum may also result from the fit of noise in the case of very low signal-to-noise-ratio (S/N) spectra. In order to identify these sources of error, we define a score, $0 \leq S(k) \leq 1$, that is associated with each shift; the shift associated with the highest score is the one that is selected by the algorithm. This score is computed as a weighted p-norm of the chi-square ratio defined as the value of the CCF evaluated at k over the maximum of the CCF,

$$\chi_r^2(k) = \frac{\mathrm{ccf}(k)}{\max_k(\mathrm{ccf})} \quad \text{where} \quad 0 \leq \chi_r^2(k) \leq 1, \tag{5}$$

and of an indicator of the presence of quasar emission lines,

$$Z_{\rm score}(k) = \prod_\lambda \left[\frac{1}{2}\left(1 + \mathrm{erf}\,\frac{e_\lambda}{\sigma(e_\lambda)\sqrt{2}}\right)\right]^{I_\lambda}, \tag{6}$$

where e_λ is the value of the BP/RP flux of the continuum-subtracted emission line at rest-frame wavelength λ if we consider the observed spectrum to be at redshift $z = L^k - 1$; $\sigma(e_\lambda)$ is the associated uncertainty and I_λ is the theoretical intensity[11] of the emission line located at λ, which is normalised so that the total intensity of all emission lines in the observed wavelength range is equal to one. Equation (6) can then be viewed as a weighted geometric mean of a set of normal cumulative distribution functions of mean zero and standard deviations $\sigma(e_\lambda)$ evaluated at e_λ. A $Z_{\rm score}$ close to one indicates that all the emission lines that we expect at redshift z are found in the spectra while missing a single emission line often leads to a very low $Z_{\rm score}$. The final formulation of the score is then given by

$$S(k) = \sqrt[p]{w_0\,[\,\chi_r^2(k)\,]^p + w_1\,[\,Z_{\rm score}(k)\,]^p}, \tag{7}$$

where w_0, w_1, and p are parameters of the weighted p-norm, as listed in Table 6.

Table 6 summarises the various parameters used in the computation of the redshift score, $S(k)$. Also, in order to facilitate the filtering of these potentially erroneous redshifts by the final user, we define binary processing flags, `flags_qsoc`, which are listed in Table 7. As later highlighted in Sect. 5.4, most secure predictions often have bits 1–4 unset (i.e. `flags_qsoc` = 0 or `flags_qsoc` = 16).

Finally, the uncertainty on the selected redshift, σ_z, is derived from the uncertainty on the associated shift, σ_k, using the asymptotic normality property of the χ^2 estimator, which states that k is asymptotically normally distributed with a variance that is inversely proportional to the curvature of the CCF around the optimum. In particular, the variance on k is asymptotically given by $\sigma_k^2 = -2\,\mathrm{d}k^2/\mathrm{d}^2\,\mathrm{ccf}(k)$, and as $Z = k\log(L)$, the logarithmic redshift, $Z = \log(z+1)$, is also normally distributed with a variance of

$$\sigma_Z^2 = 2\left|\frac{\mathrm{d}^2\,\mathrm{ccf}(k)}{\mathrm{d}k^2}\right|^{-1} \log^2(L)\,. \tag{8}$$

Furthermore, as $z = \exp Z - 1$, the redshift that is reported by QSOC is distributed as a log-normal distribution of mean Z and variance σ_Z^2, although this distribution is shifted by -1. Accordingly, the squared uncertainty on the computed redshift is given by

$$\sigma_z^2 = (z+1)^2\left(\exp\sigma_Z^2 - 1.0\right)\exp\sigma_Z^2, \tag{9}$$

while its lower and upper confidence intervals, taken as its 0.15866 and 0.84134 quantiles, respectively, are given by

$$z_{\rm low} = \exp(Z - \sigma_Z) - 1 \quad \text{and} \quad z_{\rm up} = \exp(Z + \sigma_Z) - 1. \tag{10}$$

5.3. Performance and results

The QSOC contributions to *Gaia* DR3 can be found in the `qso_candidates` table and consist of: `redshift_qsoc`, the quasar redshift, z;

[11] Theoretical emission line intensities should be regarded as weights. They do not refer to a particular theoretical model of the emission lines of quasars but to the values inferred in Table 6.

Table 7. Binary warning flags used in the QSOC redshift selection procedure and reported in the `flags_qsoc` field.

Warning flag	Bit	Value	Condition(s) for rising
`Z_AMBIGUOUS`	1	1	The CCF has more than one maximum with $\chi^2_r(k) > 0.85$, meaning that at least two redshifts lead to a similar χ^2 and the solution is ambiguous.
`Z_LOWCHI2R`	2	2	$\chi^2_r(k) < 0.9$
`Z_LOWZSCORE`	3	4	$Z_{\rm score}(k) < 0.9$
`Z_NOTOPTIMAL`	4	8	The selected solution did not correspond to the global maximum (i.e. $\chi^2_r(k) < 1$)
`Z_BADSPEC`	5	16	The BP/RP spectra upon which this prediction is based are considered as unreliable. An unreliable spectrum has a number of spectral transits in BP, $N_{\rm bp}$ or RP, $N_{\rm rp}$ that is lower than or equal to ten transits or $G \geq 20.5$ mag or $G \geq 19 + 0.03 \times (N_{\rm bp} - 10)$ mag or $G \geq 19 + 0.03 \times (N_{\rm rp} - 10)$ mag (see the online documentation for more information on the derivation of these limits).

Notes. Sources with `flags_qsoc` = 0 encountered no issues during their processing and are based on reliable spectra which means that they are more likely to contain reliable predictions.

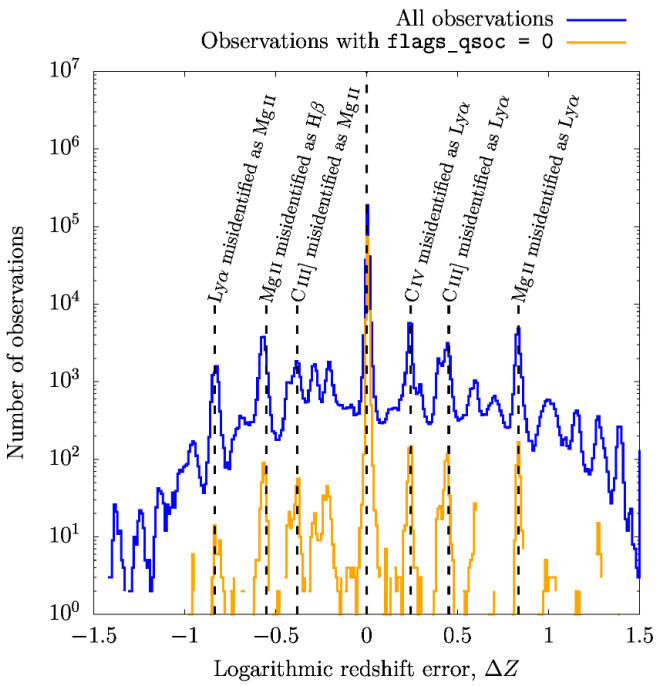

Fig. 15. Histogram of the logarithmic redshift error, $\Delta Z = \log(z + 1) - \log(z_{\rm true} + 1)$ between QSOC redshift, z, and literature redshift, $z_{\rm true}$, for 439 127 sources contained in the Milliquas 7.2 catalogue. A bin width of 0.01 was used for both curves.

`redshift_qsoc_lower`/`redshift_qsoc_upper`, the lower and upper confidence intervals, $z_{\rm low}$ and $z_{\rm up}$, corresponding to the 16% and 84% quantiles of z, respectively, as given by Eq. (10); `ccfratio_qsoc`, the chi-square ratio, χ^2_r, from Eq. (5); `zscore_qsoc`, the $Z_{\rm score}$ from Eq. (6), and `flags_qsoc`, the QSOC processing flags, $z_{\rm warn}$, from Table 7.

We quantitatively assess the quality of the QSOC outputs by comparing the predicted reshifts against values from the literature. For this purpose, we cross-matched 6 375 063 sources with redshift estimates from QSOC with 790 776 quasars that have spectroscopically confirmed redshifts in the Milliquas 7.2 catalogue of Flesch (2021) (i.e. `type = 'Q'` in Milliquas). Using a 1″ search radius, we found 439 127 sources in common between the two catalogues. It should be emphasised here that the distributions of the redshifts and G magnitudes of the cross-matched sources are not representative of the intrinsic quasar population as they inherit the selection and observational biases that are present in both the Milliquas catalogue and in *Gaia*. The numbers reported here should therefore be interpreted with that in mind. A straight comparison between the QSOC predictions and the Milliquas spectroscopic redshifts, illustrated in Fig. 15 on a logarithmic scale, shows that 63.7% of the sources have an absolute error on the predicted redshift, $|\Delta z|$, that is lower than 0.1. This ratio increases to 97.6% if only `flags_qsoc = 0` sources are considered.

As most of the DR12Q quasars we use for building our templates are also contained in the Milliquas catalogue (161 278 QSOC predictions are contained in both the DR12Q and Milliquas catalogue), one may wonder whether these induce a positive bias on the fraction of sources with $|\Delta z| < 0.1$. In order to answer this question, we note that the QSOC templates were built based on a statistically significant number of 297 264 sources, and so we expect the computed templates to be representative of the whole quasar population under study while not being too specific to the particular set of spectra we used (i.e. any other set of spectra of the same size would have provided us with very similar templates). Nevertheless, 71% of the sources in the DR12Q catalogue have $|\Delta z| < 0.1$. This compares to 59.5% of the sources with $|\Delta z| < 0.1$ that are not in the DR12Q catalogue. If we consider only sources with `flags_qsoc`= 0, then these numbers are 97% and 98.8%, respectively. The observed differences can be explained primarily by the fact that, due to the selection made in the SDSS-III/BOSS survey, 31.7% of the DR12Q sources that are found among the QSOC predictions have $2 < z < 2.6$, where the presence of the Lyα+Si IV+C IV+C III emission lines allows secure determination of the redshift (81.4% of the sources in this range have $|\Delta z| < 0.1$). In contrast, the redshift distribution of the sources that are found only in Milliquas peaks in the range $1.2 < z < 1.4$ where only 50.5% of the sources have $|\Delta z| < 0.1$, owing to the sole presence of the Mg II emission line in this redshift range (see Sect. 5.4 for more information on these specific redshift ranges). However, both subsets have a comparable fraction of predictions with $|\Delta z| < 0.1$ once these are computed over narrower redshift ranges, as expected.

We further investigate the distribution of the logarithmic redshift error, defined as

$$\Delta Z = \log(z + 1) - \log(z_{\rm true} + 1), \tag{11}$$

between QSOC redshift, z, and the literature redshift, $z_{\rm true}$, in Fig. 15. If we assume that a spectral feature at rest-frame

wavelength λ_{true} is falsely identified by QSOC as another spectral feature at λ_{false}, then the resulting logarithmic redshift error will be equal to $\Delta Z = \log \lambda_{\text{true}} - \log \lambda_{\text{false}}$, such that ΔZ, besides its ability to identify good predictions, can also be used to highlight common mismatches between emission lines. In Fig. 15, we can see that most of the predicted (logarithmic) redshifts are in good agreement with their literature values while emission line mismatches mainly occur with respect to two specific emission lines: C III] and Mg II. In the most frequent case, the C IV emission line is misidentified as Lyα, because the separation between these two lines is comparable to the separation between C IV and C III] when considered on a logarithmic wavelength scale. The Lyα and C III] lines are subsequently fitted to noise or wiggles in the very blue part of BP and in RP, respectively. By requiring that `flags_qsoc = 0`, we can mitigate the effect of these emission-line mismatches without affecting the central peak of correct predictions too much.

Finally, we note that the distribution of $\Delta Z/\sigma_Z$, where $\sigma_Z = [\log(z_{\text{up}} + 1) - \log(z_{\text{low}} + 1)]/2$ is defined in Eq. (8), effectively follows an approximately Gaussian distribution of median 0.007 and standard deviation (extrapolated from the inter-quartile range) of 1.053 if observations with $|\Delta z| < 0.1$ are considered. If only observations for which `flags_qsoc = 0` are considered, $\Delta Z/\sigma_Z$ have a median of 0.002 and standard deviation of 1.14.

5.4. Use of QSOC results

In *Gaia* DR3, QSOC systematically publish redshift predictions for which `classprob_dsc_combmod_quasar` ≥ 0.01 and `flags_qsoc` ≤ 16, leading to 1 834 118 sources that are published according to these criteria (see source_selection_flags for more information on the selection procedure). Nevertheless, for the sake of completeness, we also publish redshift estimates for all sources with `classprob_dsc_combmod_quasar` ≥ 0.01 that are contained in the `qso_candidates` table, yielding 4 540 945 additional sources for which `flags_qsoc` > 16. However, these last predictions are of lower quality as, for example, a comparison with the Milliquas spectroscopic redshift shows that 39.6% of the `flags_qsoc` > 16 sources have $|\Delta z| < 0.1$, compared to 87% for sources with `flags_qsoc` ≤ 16.

Of the source parameters published in the *Gaia* DR3, the *G*-band magnitude, `phot_g_mean_mag`, has a particularly strong impact on the quality of the QSOC predictions; it shows a clear correlation with the S/N of the BP/RP spectra, as does the number of BP/RP spectral transits to a lesser extent. From the top panel of Fig. 16, we see that more than 89% of the sources with $G \leq 19$ mag have $|\Delta z| < 0.1$ (black line) while the same fraction is obtained for spectra with $19.9 < G < 20$ mag only for sources with `flags_qsoc`$= 0$ (orange solid line). However, these correspond to a very small fraction (5.5%) of the sources in this magnitude range (blue solid line). A less stringent cut, `flags_qsoc` $= 0$ or `flags_qsoc` $= 16$, where we encounter no processing issue (i.e. flag bits 1–4 are not set) even when the BP/RP spectra are unreliable (i.e. flag bit 5 can be set), still leads to 92% of the sources with $|\Delta z| < 0.1$ (orange dotted line) while retaining 36.5% of the sources in this magnitude range (blue dotted line). The same cut concurrently retains 22% of the $20.4 < G < 20.5$ mag observations where 81.5% of the predictions have $|\Delta z| < 0.1$ and is accordingly recommended for users dealing with sources at $G > 19$ mag.

Besides the aforementioned recommendations on the flags_qsoc and *G* magnitude, we should point out an important limitation of the *Gaia* BP/RP spectro-photometers regarding the identification and characterisation of quasars, namely the fact that the Mg II emission line is often the sole detectable emission line in the BP/RP spectra of $0.9 < z < 1.3$ quasars in the moderate-S/N regime of $G \gtrsim 19$ mag spectra. Indeed, despite the broad 325–1050 nm coverage of the BP/RP spectrophotometers, quasar emission lines are often significantly damped in the observed wavelength regions $\lambda < 430$ nm and $\lambda > 950$ nm, owing to the low instrumental response in these ranges (see for example Gaia Collaboration 2023, Fig. 10). As a result, the Hβ and C III] emission lines surrounding the Mg II line[12] only enter

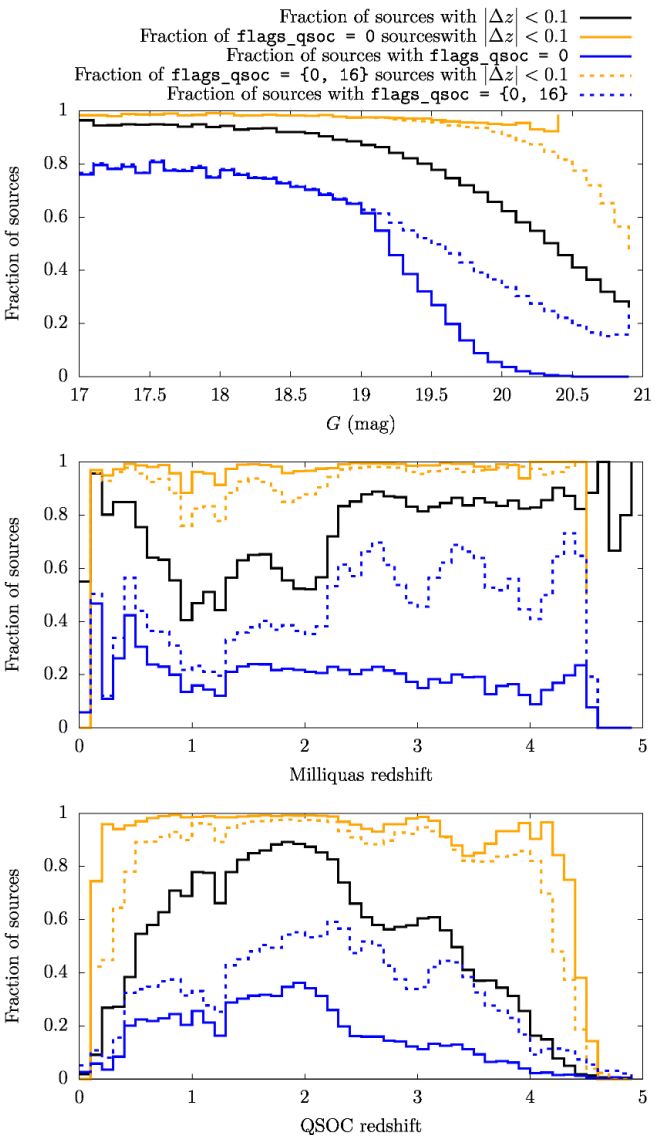

Fig. 16. Fraction of successful and reliable QSOC predictions computed over 439 127 sources contained in the Milliquas 7.2 catalogue with respect to *G* magnitude (*top*), Milliquas redshift (*middle*), and QSOC redshift (*bottom*). Black line: Fraction of observations with an absolute error of the predicted redshift, $|\Delta z|$, lower than 0.1. Orange line: Fraction of `flags_qsoc = 0` sources with $|\Delta z| < 0.1$. Blue line: Fraction of observations with `flags_qsoc = 0`. Orange and blue dotted lines correspond to their solid counterpart while considering `(flags_qsoc = 0 or flags_qsoc = 16)` observations instead of `flags_qsoc = 0` observations. Fractions are computed with respect to the number of sources in magnitude and redshift bins of 0.1.

[12] The Hγ emission line being intrinsically weak, it is often not seen in the BP/RP spectra of quasars and is accordingly not considered here.

the BP/RP spectra at $z = 0.95$ and $z = 1.25$, respectively. Nevertheless, we consider a range of $0.9 < z < 1.3$ in order to take into account low-S/N spectra where these lines, although present, are often lost in the noise. The sole presence of the Mg II emission line has the deleterious effect of increasing the rate of mismatches between this line and mainly the Lyα and Hβ emission lines, as seen in Fig. 15. Another issue also arises for $z \approx 1.3$ quasars, where the C III] emission line enters the BP spectrum while the Mg II line now lies on the peak of the BP spectrum, which complicates its detection by the algorithm leading to mismatches between C III] and the Lyα or Mg II emission lines. These effects are clearly visible in the middle panel of Fig. 16 at $0.9 < z < 1.3$, along with the previously discussed misidentification of the C IV line as Lyα at $z \approx 2$. Appropriate cuts on `flags_qsoc` allow both of these shortcomings to be alleviated, as seen in Fig. 16.

In the bottom panel of Fig. 16, we see that the fraction of sources with $|\Delta z| < 0.1$ amongst very low- and high-redshift sources, as predicted by QSOC, is low (7.25% for $z < 0.2$ sources and 2.66% for $z > 4$ sources). The explanation is that these very low- and high-z quasars are rare in our sample, such that any erroneous prediction towards these loosely populated regions is largely reflected in the final fraction of predictions (i.e. the 'purity' in these regions becomes very low). Again, cuts on the `flags_qsoc` allow us to recover about 90% of sources with $|\Delta z| < 0.1$ in the range $0.1 < z < 4.4$. Concentrating on the drop at $z < 0.1$, we note that only 69 sources have a Milliquas redshift in this range, while only 31 have $0.0826 < z < 0.1$ (i.e. in the predictable QSOC redshift range). Amongst these 69 sources, 38 have $|\Delta z| < 0.1$ while 4 have `flags_qsoc`= 0 but these are unfortunately erroneously predicted. These low numbers, along with the fact that QSOC predicts 2 154 sources in this redshift range (i.e. 0.5% of the total predictions) explains the drop at $z < 0.1$ in the middle and bottom panels of Fig. 16, even when `flags_qsoc`= 0. Regarding the $z > 4.4$ quasars, only 76 of them have redshifts in both *Gaia* and Milliquas, while only 10 have `flags_qsoc` = `0` and 9 of these also have $|\Delta z| < 0.1$. There are 18 959 sources with QSOC redshift predictions in this range, although only 101 (i.e. 0.5%) of them have `flags_qsoc`= 0. This leads to a rather poor fraction of 9/101 of the sources with $|\Delta z| < 0.1$ and `flags_qsoc` = `0` in this redshift range.

In conclusion, we should insist first on the fact that QSOC is designed to process Type-I/core-dominated quasars with broad emission lines in the optical and accordingly yields only poor predictions on galaxies, type-II AGN, and BL Lacertae/blazar objects. Secondly, SMSgen does not provide covariance matrices on the integrated flux (Creevey et al. 2023), meaning that the computed χ^2 from Eq. (2) is systematically underestimated and is consequently not published in *Gaia* DR3. The computed redshift and associated confidence intervals, $z_{\rm low}$ and $z_{\rm up}$ from Eq. (10), though appropriately re-scaled, might also sporadically suffer from this limitation.

6. Unresolved galaxy classifier (UGC)

6.1. Objectives

The Unresolved Galaxy Classifier (UGC) module estimates the redshift, z, of the sources with $G < 21$ mag that are classified as galaxies by DSC-Combmod with a probability of 0.25 or more (see Sect. 3 for details). UGC infers redshifts in the range $0 \leq z \leq 0.6$ by using a combination of three support vector machines (SVMs, Cortes & Vapnik 1995), all taking as input the BP/RP spectra of the sources as sampled by SMSgen (Creevey et al. 2023, Sect. 2.3.2). The SVMs are trained on a set of BP/RP spectra of galaxies that are spectroscopically confirmed in the SDSS DR16 archive (Ahumada et al. 2020). UGC further applies filtering criteria for selecting redshifts to be published in *Gaia* DR3, as described in Sect. 6.2.

6.2. Method

UGC is based on the LIBSVM library of Chang & Lin (2011), from which three SVM models are built: (i) t-SVM, the total-redshift range SVM model, which computes the published redshift, `redshift_ugc`, and associated SVM prediction intervals, `redshift_ugc_lower` and `redshift_ugc_upper`, (ii) r-SVM, and (iii) c-SVM, which are respectively regression and classification SVM models applied to discretised versions of the redshift and used exclusively for the internal validation of the redshift produced by the t-SVM model. All SVM models use common training and test sets, which we describe below.

6.2.1. Training and test sets

The sources in the training and test sets were selected from the SDSS DR16 archive (Ahumada et al. 2020), which provide position, redshift, magnitudes in the *u*-, *g*-, *r*-, *i*-, *z*-bands, photometric sizes (we used here the Petrosian radius), and interstellar extinction for each spectroscopically confirmed galaxy. There are 2 787 883 objects in SDSS DR16 that are spectroscopically classified as galaxies, but we rejected sources with poor or missing photometry, size, or redshift, thus reducing the number of galaxies to 2 714 637. Despite the known lack of uniformity of the SDSS DR16 redshift distribution due to the BOSS target selection[13], this survey still provides the largest existing database of accurate spectroscopic redshifts of galaxies that can be used as target values in the SVM training and test sets.

The selected galaxies were cross-matched to the *Gaia* DR3 sources prior to their filtering by CU9 using a search radius of 0.54″, which resulted in 1 189 812 cross-matched sources. Amongst these, 711 600 have BP/RP spectra, though not all of them are published in *Gaia* DR3. Because the inclusion of high-redshift galaxies would lead to a very unbalanced training set (i.e. very few high-redshift galaxies), we further imposed an upper limit on the SDSS DR16 redshift of $z \leq 0.6$, leaving 709 449 sources that constitute our *base set*.

For the preparation of the training set, a number of conditions were further imposed on the sources in the base set: (i) $G \leq 21.0$ mag; (ii) BP/RP spectra must be composed of a minimum of six epochs of observations; (iii) the mean flux in the blue and red parts of the BP/RP spectra, as computed by UGC, must lie in the ranges $0.3 \leq bp\text{SpecFlux} \leq 100\,\mathrm{e^-\,s^{-1}}$ and $0.5 \leq rp\text{SpecFlux} \leq 200\,\mathrm{e^-\,s^{-1}}$, respectively, in order to exclude potentially poor-quality spectra; (iv) the image size, as characterised by the Petrosian radius, must be in the range $0.5'' \leq \text{petroRad50\_}r \leq 5''$ in order to exclude suspiciously compact or significantly extended galaxies; (v) the interstellar extinction in the *r*-band must be below the upper limit of $\text{extinction\_}r \leq 0.5$ mag to avoid highly reddened sources; and (vi) the redshift must be larger than 0.01 in order to exclude nearby extended galaxies. After applying all these cuts, 377 875 sources remained, which we refer to as the clean set. Of these, 6000 sources were randomly selected in order to construct the training set, the redshift distribution of which is given in Table 8.

[13] `https://www.sdss.org/dr16/algorithms/boss_target_selection/`

Table 8. Distribution of the sources in the UGC data sets according to their SDSS redshifts.

	Redshift ranges						
Data set name	0.0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	Total
Base set	224 264	292 968	118 248	65 912	7 055	1 002	709 449
Clean set	152 564	192 675	29 145	2 490	724	327	377 875
Clean test set [(a)]	150 964	191 025	28 045	1 590	224	27	371 875
Training set	1 600	1 600	1 100	900	500	300	6 000
Base test set [(a)]	222 664	291 368	117 148	65 012	6 555	702	703 449

Notes. [(a)]The base test set and clean test set are respectively composed of sources in the base set and clean set that are not contained in the training set.

Table 9. Galactic coordinates and colour–colour regions from which UGC results are filtered out.

Area	Galactic coordinates range		Colour-colour box A	Colour-colour box B
	longitude [°]	latitude [°]	[mag]	[mag]
CNT	0.0 ± 15.0	-5.0 ± 5.0	$-0.5 < G - G_{BP} < 0.5$ $0.4 < G_{BP} - G_{RP} < 1.3$	$-0.5 < G - G_{BP} < 3.0$ $-0.2 < G_{BP} - G_{RP} < 1.4$
LMC	279.5 ± 4.0	-33.25 ± 3.25	$-3.0 < G - G_{BP} < -1.5$ $-0.4 < G_{BP} - G_{RP} < 1.0$	$-0.7 < G - G_{BP} < 2.0$ $-0.8 < G_{BP} - G_{RP} < 1.4$
SMC	303.0 ± 1.0	-44.0 ± 1.0	$-3.0 < G - G_{BP} < -1.5$ $-0.4 < G_{BP} - G_{RP} < 1.0$	$-0.7 < G - G_{BP} < 2.0$ $-0.8 < G_{BP} - G_{RP} < 1.4$

Notes. Those correspond to regions where extragalactic objects are not expected: Magellanic clouds (LMC, SMC) and an area (CNT) close to the Galactic centre.

The imbalance of this training set is clearly visible in this table, and is caused by the small number of high-redshift galaxies present in the clean set.

The conditions described in the previous paragraph were not imposed for the test set. Instead, all 703 449 spectra in the base set that were not used for training were included in the base test set, whose redshift distribution is shown in Table 8. Additionally, a purest test sample, the clean test set, was derived from the clean set by removing the training data it contains.

6.2.2. Support vector machine models

The input of all SVM models are BP/RP spectra. The spectra are first truncated by removing the first 34 and the last 6 samples in BP, and the first 4 and the last 10 samples in RP, in order to avoid regions of low S/N. These cuts result in the definition of the usable wavelength ranges for the BP and the RP parts of the spectrum, namely 366–627 nm and 620–996 nm, respectively. Each pair of truncated spectra is then concatenated to form the SVM input vector of 186 fluxes.

A common setup was implemented for the SVM model preparation (see LIBSVM[14] for details): The Standardization Unbiased method was selected to scale the target data and the vector elements to the range $[-1.0, 1.0]$; the radial basis function (RBF) $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp(-\gamma|\boldsymbol{x_i} - \boldsymbol{x_j}|^2)$ was chosen as the kernel function, and the tolerance of the termination criterion is set to $e = 0.001$; shrinking heuristics are used to speed up the training process; a four-folded tuning (cross-validation) is applied to determine the optimal γ kernel parameter and the penalty parameter C of the error term in the optimisation problem.

The UGC redshifts are estimated by t-SVM, which implements a ϵ-SVR regression model trained for redshifts in the range $0.0 \leq z \leq 0.6$. The two other SVM models, c-SVM and r-SVM, use the BP/RP spectra as input but are trained to predict a discretised version of the redshifts and are used solely for the purpose of redshift validation (Sect. 6.2.3). The c-SVM model is a C-SVC classification model trained on six different classes corresponding to the redshift ranges $0 \leq z < 0.1$, $0.1 \leq z < 0.2$, $0.2 \leq z < 0.3$, $0.3 \leq z < 0.4$, $0.4 \leq z < 0.5$, and $0.5 \leq z < 0.6$. The output of the c-SVM model is a class-probability vector. The element of the vector with the highest value above 0.5 is taken as the selected class. If there is no element with probability larger than 0.5, then the source is marked as unclassified. The r-SVM model implements the ϵ-SVR regression model of LIBSVM – similarly to the t-SVM model – but it is trained on six discrete target values $(0.05, 0.15, \ldots, 0.55)$. As only the first decimal is retained for the predictions, the output of the r-SVM model is directly comparable to the classes used by the c-SVM model.

6.2.3. Source filtering

Two sets of criteria are used to select the UGC outputs to be published in *Gaia* DR3. The first set applies to specific properties of the processed sources, while the second concerns the redshift validity. An output is included in *Gaia* DR3 only if all the criteria of the two sets are satisfied.

Although UGC processes all $G < 21$ mag sources for which the DSC Combmod galaxy probability is higher than or equal to 0.25, additional criteria were imposed for selecting the purest sample of results. First, we require that the number of spectral transits in both BP and RP is higher than or equal to ten. Second, we require that the mean flux in the blue and red parts of the BP/RP spectra lies in the ranges set in Sect. 6.2.1. Third, we decided to only publish redshifts for sources with $G > 17$ mag, so as to exclude bright and possibly extended sources, for which it is likely that only part of the galaxy has been recorded. Fourth, we require $G - G_{BP} > 0.25$ mag in order to reduce the number of sources with true $z > 0.6$ (which lie outside the range of

[14] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

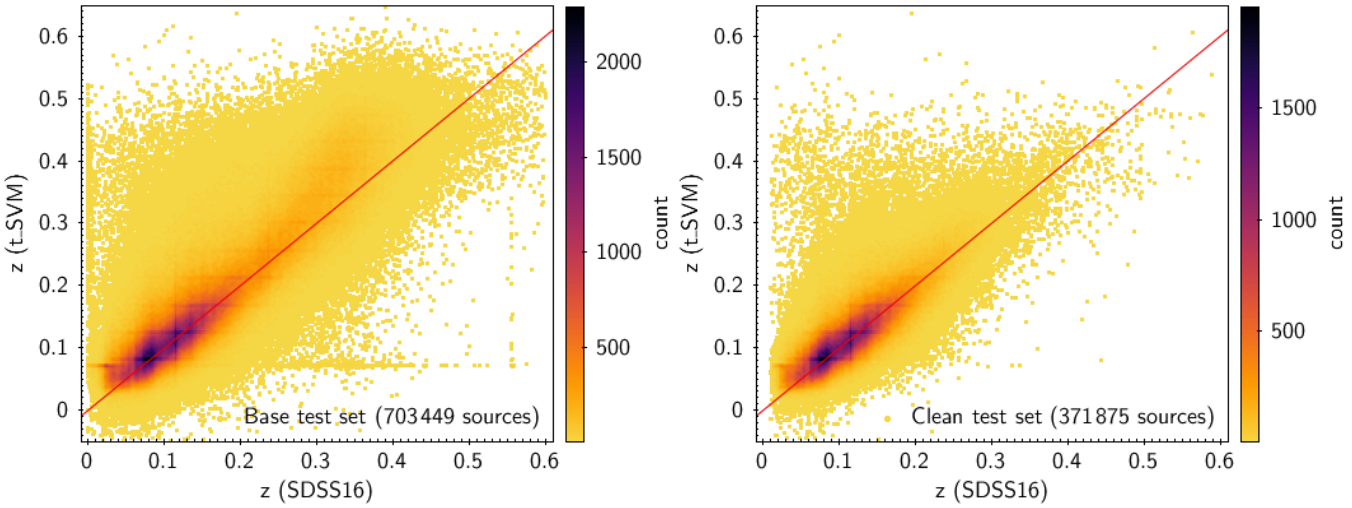

Fig. 17. Comparison of the UGC redshifts, as estimated from the t-SVM model with SDSS DR16 redshifts for the base test set (*left*) and for the clean test set (*right*), as identified in Sect. 6.2.1.

the training data) by as much as possible. The fifth and final condition is related to the location of blended sources that are erroneously classified as galaxies in high-density regions in the sky (see also Sect. 3.4). Indeed, the positional distribution of the sources processed by UGC shows a high concentration of galaxies in three small areas where extragalactic objects are not expected in large numbers: a region below the Galactic centre, and two areas centred on the Magellanic Clouds (see Table 9). Almost 9% of the total number of processed sources originate in these three areas. Sources in these areas also occupy a specific region of the $G - G_{\rm BP}, G_{\rm BP} - G_{\rm RP}$ colour–colour diagram that is distinct from the locus of the remaining sources. This distinction has been used to define colour cuts (shown in Table 9) which, in combination with the coordinates of the three areas, allowed us to clean the suspicious clumps of galaxies and to remove a large number of potentially misclassified sources in these three areas. Nonetheless, conditions listed in Table 9 are not applied if the DSC Combmod probability for the source to be a galaxy is equal to one.

The comparison of the redshifts produced by the t-SVM model to those of the r-SVM and c-SVM models allows us to internally validate the UGC redshifts. The implementation of the filtering involves first the rejection of sources for which at least one of the SVM models has not produced an output (either because there is no prediction or because the source is marked as unclassified). Second, the three computed redshifts are required to span at most two adjacent bins of redshift, similar to those defined for the c-SVM and r-SVM models. The largest absolute difference between the t-SVM redshift and the central value of the c-SVM and r-SVM redshift bins is 0.08. The redshifts of sources not satisfying one of these criteria are not published in *Gaia* DR3.

6.3. Performance

The overall performance of the t-SVM model is given by the mean (μ) and the standard deviation (σ) of the difference between the estimated and the real (target) redshifts. The internal test, applied to the training set itself, yields $\sigma = 0.047$ and $\mu = -0.003$. The external test, which is performed on all 703 449 spectra in the base test set, yields $\sigma = 0.053$ and $\mu = 0.020$ (Fig. 17, left panel). These values indicate that the performance is worse for the base test set, as expected. If the clean test set of 371 875 spectra is used the performance is improved significantly, with $\sigma = 0.037$ and $\mu = 0.008$ (Fig. 17, right panel).

The performance varies with redshift. To quantify this, the base test set was divided into SDSS redshift bins of size 0.02. The mean, μ_i, and the standard deviation, σ_i, of the differences between the redshift predicted by t-SVM and the real (SDSS) redshifts were determined for each one of these bins, as shown in Fig. 18 (left panel). Generally, there are three regions with different performance. For $z < 0.02$, the error and the bias are relatively large indicating that the t-SVM is ineffective for redshifts close to zero. The performance is good in the range of $0.02 < z < 0.26$; however, for larger redshifts, the bias changes significantly from almost zero to positive and then to negative values, while the error progressively increases. For $z > 0.5$, both μ_i and σ_i show large scatter, probably due to the fact that large redshifts are under-represented in the t-SVM training set.

In addition, the performance of the t-SVM model as a function of redshift was investigated by constructing a confusion matrix, as in classification problems. To this effect, a different class has been assigned to each redshift bin, $z_{\rm bin}$, both for the real (SDSS) and the predicted (t-SVM) redshifts. In this case, the bin size was 0.1. The confusion matrix presents the total number of cases for each real and each predicted class (see for details the online documentation).

For a given redshift bin, $z_{\rm bin}$, the numbers of true-positive TP, false-negative FN, and false-positive FP predictions are used to evaluate the sensitivity, or completeness, $TP/(TP+FN)$, and the precision, or purity, $TP/(TP+FP)$. Figure 18 (middle and right panels) show the t-SVM completeness and purity for the redshift bins of the base and clean test sets in bins of redshift. Both completeness and purity for the base and clean test sets are very good up to a redshift of $z = 0.2$. The purity is moderate (~0.5) for the two test sets for the redshift bin 0.2–0.3 and fails at larger redshifts. The completeness is moderate in the 0.3–0.5 bin and fails for the last bin. Generally, good performance can be expected for redshifts $z \leq 0.2$.

6.4. Results

The UGC output is included in the `galaxy_candidates` table. There are 1 367 153 sources for which UGC provides a redshift value as estimated by t-SVM (Sect. 6.2.2), `redshift_ugc`,

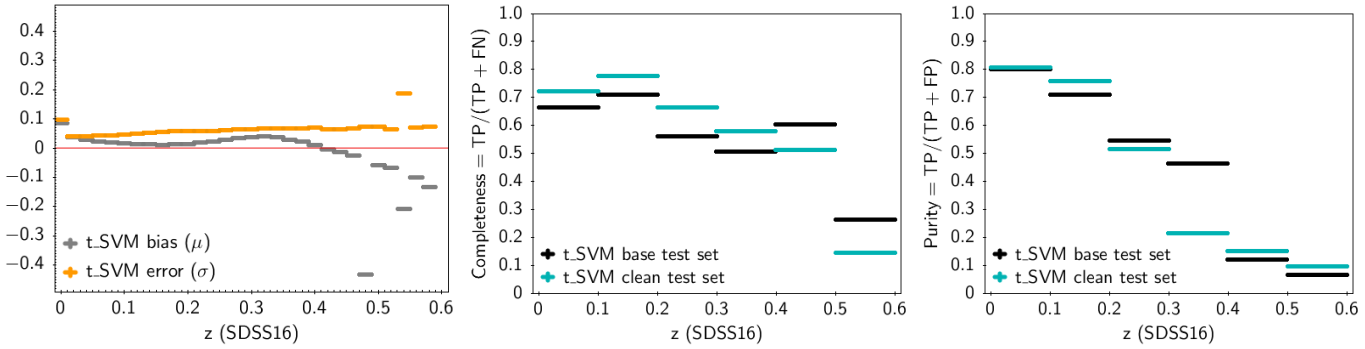

Fig. 18. *Left panel*: mean (μ_i) and standard deviation (σ_i) of the difference between the UGC redshifts, from the t-SVM model, and associated SDSS redshifts for sources contained in the UGC base test set and averaged over redshift bins of size 0.02. Completeness (*middle panel*) and purity (*right panel*) as a function of redshift, evaluated on the UGC test set (black) and clean set (cyan). The bin size is equal to 0.1.

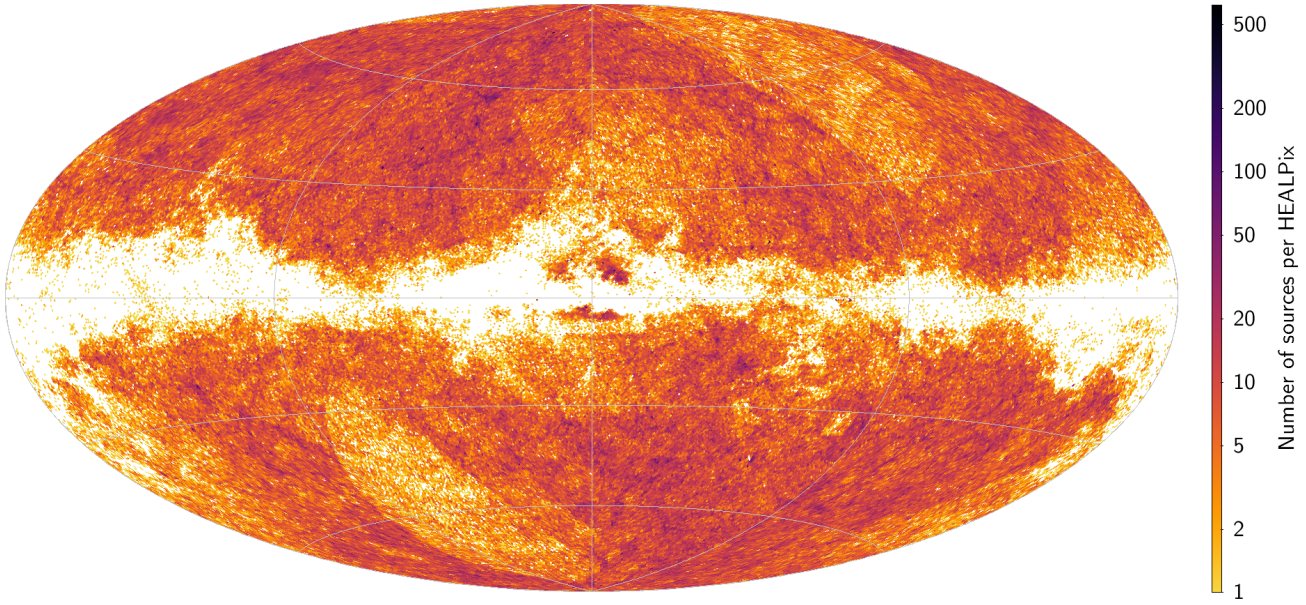

Fig. 19. Galactic sky distribution of the number of sources with redshifts estimated by UGC. The plot is shown at HEALPix level 7 (0.210 deg^2).

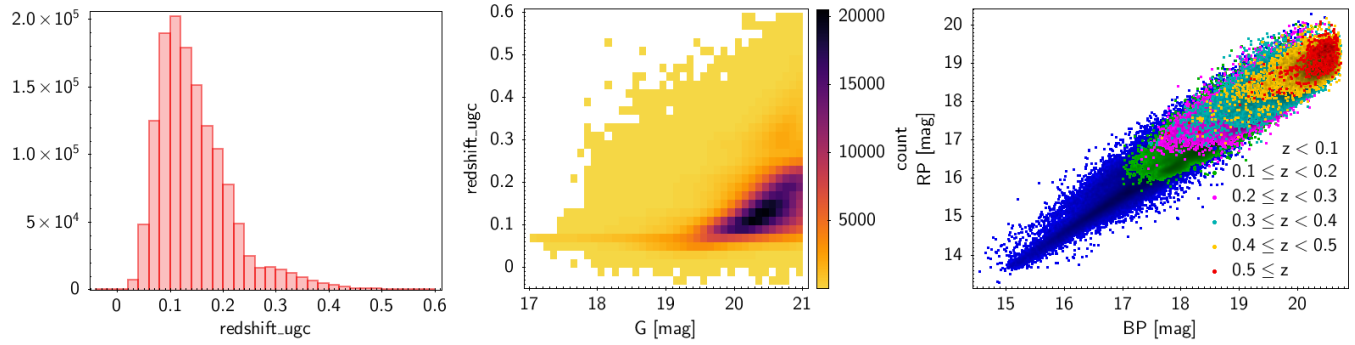

Fig. 20. Distribution of the UGC redshifts. *Left*: histogram of the estimated redshift in bins of size 0.02. *Middle*: UGC redshifts as a function of G magnitude. *Right*: distribution of the sources with UGC redshifts on a BP/RP magnitude diagram where different colours correspond to different redshift ranges.

along with the corresponding lower and upper limits of the SVM prediction interval, `redshift_ugc_lower` and `redshift_ugc_upper`, respectively. The parameter `redshift_ugc_lower` is defined as `redshift_ugc`$-\mu_i - \sigma_i$, where i corresponds to the ith redshift range identified in the previous section, and μ_i and σ_i are the associated bias and standard deviation computed on the base test set. Similarly, the parameter `redshift_ugc_upper` is defined as `redshift_ugc`$-\mu_i+\sigma_i$. The value of (`redshift_ugc_upper`$-$`redshift_ugc_lower`)/2 can therefore be used as an estimate of the 1-σ uncertainty on `redshift_ugc`.

Apart from the Galactic plane, the sources with UGC redshifts are almost uniformly distributed on the sky, as seen in Fig. 19, although there are two strips (lower-left and upper-right)

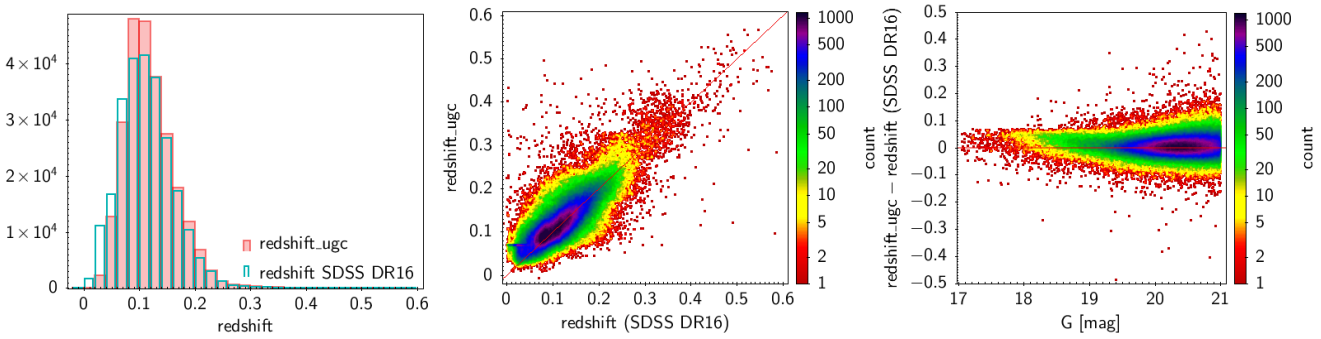

Fig. 21. Comparison of the UGC estimated and the actual (SDSS DR16) redshifts for the 248 356 sources in common (not shown are 67 sources with actual redshift greater than 0.6). *Left panel*: distributions of the UGC redshifts and SDSS DR16 redshifts indicates that UGC tends to overestimate the small redshifts. *Middle panel*: comparison of the UGC redshifts and SDSS DR16 redshifts. The unit line is shown in red. A small horizontal branch at `redshift_ugc`=0.07 is discussed in the text. *Right panel*: differences between the UGC and SDSS DR16 redshifts as a function of G magnitude. The red horizontal line designates perfect agreement.

of relatively lower density displaying residual patterns. These are regions that have been observed fewer times by *Gaia* and thus many of the sources in them do not appear in the UGC output because of the filters applied on the number of transits (see Fig. 5).

The distribution of the estimated `redshift_ugc` values shown in the left panel of Fig. 20 has a maximum at $z \simeq 0.1$, while almost 91% of the redshifts are within $0.05 \leq z < 0.25$. About 7% of the sources have redshifts larger than 0.25. The lowest and the highest redshifts reported are $z_{\rm min} = -0.036$ and $z_{\rm max} = 0.598$, respectively. There are 33 sources with negative redshifts, although most of these values are very close to zero (with median value of −0.0054).

The dependence of the `redshift_ugc` values on G magnitude is shown in the middle panel of Fig. 20. As expected, sources with higher redshift are fainter (e.g. $z > 0.4$ sources are mostly found at $G > 19$ mag, while $z > 0.5$ sources are found at $G > 20$ mag). The dependence of the estimated redshift on the source magnitude is also evident in the BP/RP magnitude–magnitude diagram shown in the right panel of Fig. 20, where different redshift ranges are represented with different colours.

There are 248 356 sources with published `redshift_ugc` in common with those spectroscopically classified as `'GALAXY'` or `'QSO'` in the SDSS DR16 (using a radius of 0.54″, as before). The differences between the `redshift_ugc` and the SDSS redshifts have a mean and standard deviation of $\mu = 0.006$ and $\sigma = 0.054$, respectively. If the 67 sources with SDSS redshifts greater than 0.6 are excluded, the standard deviation is reduced to 0.029. Figure 21 (left panel) compares the distributions of the two redshift estimates. There is a clear excess in the number of sources with UGC redshifts around 0.1 compared to the SDSS redshifts. At the same time, there is a deficit in the lower redshift bins for UGC. The observed differences are probably due to an overestimation by UGC of lower SDSS redshifts. These effects are better demonstrated in Fig. 21 (middle panel). Most of the sources follow the unit line, albeit with significant scatter. However, there is a small bias which tends to be positive for $z \approx 0.1$.

We also see in Fig. 21 (middle panel) a short dense horizontal feature of sources with `redshift_ugc` around 0.07, while the corresponding SDSS redshifts span a range of values from ≃0 to 0.07. We see that the majority of these problematic values occur at 0.07 < `redshift_ugc` < 0.071, with 5178 sources with redshift values in the range 0.070822–0.070823. Detailed analysis (see the online documentation) indicates that this peak contains a relatively large fraction of very bright sources (with $G < 17.5$, $G_{\rm BP} < 16$ and $G_{\rm RP} < 15$ mag), suggesting that the SVM models, which are not trained at all for bright, nearby galaxies, tend to make constant redshift predictions for such objects.

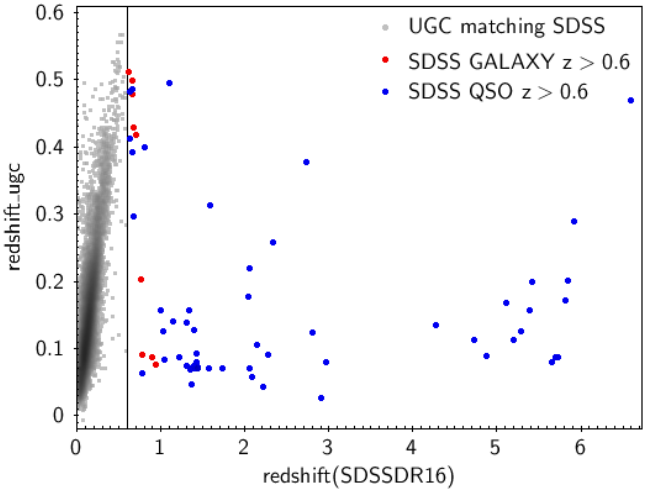

Fig. 22. UGC sources with high redshift from the SDSS DR16. Blue and red points are sources that are spectroscopically classified as `'QSO'` and `'GALAXY'` in the SDSS DR16, respectively.

Figure 21 (right panel) shows the difference between `redshift_ugc` and the actual SDSS redshift, as a function of G magnitude. As expected, the performance of the UGC redshift estimator is poorer for fainter sources as indicated by the larger dispersion seen at faint G magnitudes. The positive bias of the very bright and nearby galaxies is also clearly seen.

6.5. Use of UGC results

UGC selects sources that have a DSC probability of being a galaxy of `classprob_dsc_combmod_galaxy` ≥ 0.25. This is a relatively low threshold, and so the final UGC galaxy catalogue is expected to include some misclassified quasars. Indeed, 5170 sources, or ≃2% of the sources in common with the SDSS DR16, have a SDSS spectroscopic class `'QSO'` while 58 of them also have SDSS redshifts $z > 0.6$, i.e. higher than the UGC limit. There are also 9 high-redshift sources spectroscopically classified as `'GALAXY'` by the SDSS. Figure 22 shows a comparison between `redshift_ugc` and SDSS redshifts for