DSC ApsisOps3.2 Validation Report

prepared by: C.A.L. Bailer-Jones
approved by: Carlos Dafonte, Morgan Fouesneau
reference: GAIA-C8-TN-MPIA-CBJ-094
issue: 1
revision: 2
date: 2021-11-04
status: Issued

Abstract

DSC classifies objects into five classes using BP/RP spectra and various features constructed from astrometry and photometry. The main focus is on the classes "quasar" and "galaxy", as they are defined by our particular training data. ApsisOps3.2 (run 2) produces classification results for 1.8 billion sources. I summarize here these results and performance. Sample completeness and purity is computed on a validation set which is adjusted to represent realistic frequencies of occurence ("class fractions") of classes in Gaia. This reflects in particular the difficulty of identifying rare extragalactic objects among the much more dominant stars (by a factor of 10^3). I show how the performance varies with various parameters, such as magnitude and Galactic latitude. I analyse subsets of the complete set of results for sources classified with probabilities above 0.5 (and other thresholds) in the different classes, showing distributions in features, on the sky, as well as in colour–colour and colour–magnitude diagram.

Document History

Issue	Revision	Date	Author	Comment
D	1	2021-06-23	CBJ	Ready for CU8 internal review
D	2	2021-07-01	CBJ	Corrected error for "All" results in section 13 (thanks to David Teyssier). Specified quasar and galaxy selections more precisely in section 3.2.
1	0	2021-07-28	CBJ	Updated following comments from Carlos Dafonte. Added Figure 8 and appendix A.
1	1	2021-08-17	CBJ	Added minor details and corrected some typos
1	2	2021-11-04	CBJ	The last two columns of Table 3 were erroneously labelled and described as "All", i.e. also requiring the Combmod probability to be above 0.5. Corrected in table and in a few places in text.

Contents

1 Summary (for the CU8 validation report MFX-026)

Module overview DSC classifies sources probabilistically into five classes: quasar, galaxy, anonymous (essentially "star"), white dwarf, physical (unresolved) binary. DSC is, however, primarily an extragalactic classifier, so performance on white dwarfs and physical binaries is of less interest. It can be applied to every Gaia source which has the required input data.

DSC comprises three classifiers:

1. Specmod (ExtraTrees model). Input: BP/RP spectrum. Output: probabilities in the five classes.
2. Allosmod (Gaussian mixture model). Input: 8 features defined from Gaia data. Output: probabilities into the classes: quasar, galaxy, anonymous.
3. Combmod (combination of conditional probabilities). Input: probabilities from Specmod and Allosmod. Output: probabilities in the five classes.

Classes are defined by their training data (defined in CBJ-090), so you should not expect that an arbitrary quasar which does not conincide with the training selection will be classified as "quasar" with high probability, for example. This is especially true for the physical binaries.

DSC probabilities are always posterior probabilities that accommodate the class prior. The class prior is set to the expected class fraction, which is the relative frequency of objects of each class that we expect to be in the Gaia data. The class prior is listed in Table 2 and is explained in section 2.

Goals of ApsisOps3.2

✔ Produce the final results for Gaia DR3

Main results Performance was evaluated on the subset of the VST (Validation Source Table; see section 3.4) that has class labels and results from both Specmod and Allosmod. We call this the *defset*; it comprises 1 457 348 sources. The sources are representative of what was used to train DSC for all classes except physical binaries. Classes are allocated by maximum posterior probability (possibly above a minimum probability threshold) from which we can form the confusion matrix. From this we compute the completeness of each true class and the purity of each assigned class.

As the defset has arbitrary class fractions, the confusion matrix is adjusted to be representative of the performance on a sample of sources selected at random from Gaia. Specifically, it allows for the fact that most classes are very rare compared to single stars. (Note that this adjustment is in addition to using the class prior to get posterior probabilities.) Failure to make this correction gives meaningless results: it would generally show artificially good performance. See section

3.4 of Bailer-Jones et al. (2019) for a full explanation.

The performance is shown in Table 3. This is reasonable given the extreme rareness of the non-star classes. Performance is better if we limit the analysis to brighter magnitudes. The observed performance on physical binaries is poor. This is partly because it is assessed validation data (in the VST) that is not representative of the training data, for reasons that are explained in section 3.4.

The number of source classified with probabilities above 0.5 for each class in the full set of DSC results was also analyses and various plots of their properties are presented.

Known issues The flags written by the DSC processing – to `flagArray` – are not meaningful. In fact, we hacked this field in the post-processing to store the `classlabelDscJoint` (see section 4), which will be written to the Galaxy and Quasar tables.

Performance on physical binaries appears poor, for reasons just mentioned. Even if we had consistent data, separating all but the most obvious physical binaries (nearby, equal-mass systems) remains difficult.

Performance depends unavoidably on the prior. Using a uniform prior would give *apparently* better results, but worse *actual* performance than when we use a realistic prior and a corresponding validation data set. This is explained and demonstrated in Bailer-Jones et al. 2019, and will be shown in the results (see also section 2 below). It is crucially important that this dependence on the prior be explained in the documentation, as many readers may not appreciate its significance.

External feedback needs None.

Post-processing of the outputs We hacked the DSC flag field `flagArray` to store the `classlabelDscJoint` (see section 4), which will be written to the Galaxy and Quasar tables.

Mask GDR3 values if No masks or filtering.

Next developments None for GDR3.

Table 2: DSC class prior. This is the class prior for Specmod. "Anonymous" is essentially the same as "star". The prior for the anonymous class in Allosmod is the sum of anonymous, white dwarf, and physical binary.

	quasar	galaxy	anonymous	white dwarf	physical binary
$\propto$	1/1000	1/5000	1	1/5000	1/100
$=$	0.000989	0.000198	0.988728	0.000198	0.009887

2 What you must know to correctly interpret the results

Classification is done probabilistically. It depends not only on the probability of the data given the true classes (the likelihood), but also on the base rate (the prior). By way of analogy: a fever is a symptom of ebola (high likelihood), but ebola usually has a very low base rate (low prior). So what does having a fever imply? Priors are not an optional extra for inference. They are essential for achieving a meaningful classification. This is as true with Gaia as it is with diseases. It is relatively easy to build a classifier that separates stars from quasars using Gaia data with an "accuracy" of 95% or more. But if you then apply this classifier to a sample in which quasars are 1000 times rarer than stars – as they are in the whole Gaia sample – the resulting "quasar" sample would be mostly stars. The claimed "accuracy" of 95% is meaningless. We avoid this problem by using an appropriate class prior. A second adjustment is further required when assessing the performance of the classifier on a labelled test set, as this will usually not have enough potentially contaminating stars, and so would lead to a (significant) overestimation of the purity if not corrected for. Both of these steps are explained in section 3.4 of Bailer-Jones et al. (2019). That article also shows empirically what happens if you do not use sensible priors: your actual performance becomes worse (because you are solving the wrong problem), but your apparent performance becomes better (because you're using the wrong data set). I use the same terminology and notation in the present report as in that article.

To address the first issue stated above, DSC probabilities are always posterior probabilities that accommodate the class prior. This class prior is set to the expected class fraction, which is the relative frequency of objects of each class that we expect to be truly in the Gaia data (or more precisely, the subset of Gaia data that can be classified by DSC). The class prior is given in Table 2. These priors refer to the classes as we define them, i.e. by the training data we use (see below). For quasars and galaxies, the priors are derived from considerations in Bailer-Jones et al. (2019) and on the fraction of these classes found in other surveys. The value for white dwarfs is based on the 22 000 white dwarfs in the EDR3 Gaia Catalogue of Nearby Stars (Gaia Collaboration et al. 2021a) found out to 100 pc and assuming we can see all white dwarfs at the same space density out to 220 pc (assuming all white dwarfs have $M_G = 14$ mag, Gaia's magnitude limit is $G = 20.7$ mag, and there is no extinction). For binaries we just adopt a class fraction of 1%. There is of course a lot of uncertainty in these priors, and adopting different priors would lead to different samples for a given probability selection threshold. As we provide class probabilities, the user can do this.

Concerning the second issue, when assessing the performance of a classifier on an test set with an arbitrary class fraction (i.e. different from the class prior), then we must adjust the resulting confusion matrix to get the appropriate purities. The completeness is unaffected by this.

Classes are defined by their training data. Thus "galaxy" is not any galaxy, but one with Gaia data that are consistent with the galaxies in our training set. This is particularly relevant for the "physical binary" class, as this represents a very small subset of all types of physical binary one could think of and which Gaia no doubt observes. Of the five classes that DSC classifies into, only four are defined by an explicit, labelled, training set. The fifth class, "anonymous", is intended to represent everything else. This is achieved by defining the anonymous class simply as a random subset of all Gaia sources that have the required data. As such a sample is dominated by stars (by a factor of about 1000), this class is essentially synonymous with stars. (Some of these could of course be binary stars, although most could not be identified as such with the data we use.) This class is uninteresting for most purposes, because it is trivial to get an excellent single-star classifier for Gaia: just call everything a single star and your classifier has 99.9% purity and 99.9% completeness.

Beware also of assessing results on a very pure quasar sample, e.g. as defined for the Gaia reference frame. Such samples generally exclude quasars with large Gaia-measured parallaxes and proper motions. We make no such exclusions, neither in the training set or testing set. Analysis on other samples done by others often make numerous cuts on other quantities, such as ruwe, so as to increase their purity (at the cost of completeness). In our analyses we do not make such cuts.

3 DSC configuration and data

CBJ-090 provides an overview of DSC in cycle 3. Here we give the set up for this run.

3.1 Code configuration

The DSC version for this run was delivered by DPCC on 18 February 2021 via the Jira issue C8CAL-669.

Specmod is an ExtraTree model. Its input is the BP/RP spectrum. Its output is probabilities for each of the five classes (adjusted for the classprior, of course). Specmod training is done in Java.
The Specmod version is Specmod-v12-20.4.0 (folder `specmod_20201008`).

Allosmod uses Gaussian Mixture Models (GMMs) to classify Gaia sources into 3 classes: quasar, galaxy, anonymous. It uses 8 Gaia measures:
`sinb`, `parallax`, `pm`, `uwe`, `phot_g_mean_mag`, `bp_g`, `g_rp`, `relvarg`.
This is very similar to the model used in Bailer-Jones et al. (2019), so see that article for more details. The likelihood for each class is defined by 20 8-dimensional Gaussians. The

Allosmod anonymous class is the union of the anonymous, white dwarf, and physical binary classes in Specmod. The class prior for this class is the sum of their Specmod class priors, i.e. $0.988728 + 0.000198 + 0.009887 = 0.998813$.
Allosmod training is done in R. The code version used was `allosmod_20210108`. The Allosmod application is done by a separate Java code. The final update of this code is described in the DSC delivery Jira for this run, C8CAL-669.

Combmod takes the probability outputs from Allosmod and Specmod for each source, and combines them to give a new posterior probability over the five classes (with the same class prior, care being taken not to double the prior). The procedure for doing this is explained in CBJ-090. Note that all three classifiers take into account the class prior, and that Combmod only includes this once.

The DSC configuration file `dsc_global_config.properties` is shown in Table 1.

3.2 Training/testing 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 physical binaries, as will be explained). The classes were defined by selecting sources of each class from an external database and crossmatching them to Gaia edr3int2. These `source_id` was then used to extract the relevant Gaia data from DPCC. A subset of these data were then used to train, another to test. The sources used to defined each of the five classes are as follows.

Quasars SDSS-DR14 quasar catalogue[1]. This is the same catalogue as was used in Bailer-Jones et al. (2019) and is nominally the same subset. Specifically, we select the subset with
`SDSS_NAME!="00000-00000" AND ZWARNING=0`
which gave 446 492 rows. (We write "rows" here these are not necessarily unique sources.) We then positionally crossmatched this to[2] Gaia edr3int2 by taking the closest match in Gaia within a $1''$ radius using the tool on the Gaia archive web page. Propagation of coordinates (primarily of the stars) using proper motions was not done because (a) the SDSS epoch is not clearly defined for individual sources, and (b) both the Gaia archive and other codes available for doing this were too slow. All sources were retained that had both `phot_bp_n_obs` and `phot_rp_n_obs` equal to or greater than 5. This gave 349 876 sources.

Galaxies SDSS-DR15 spectroscopic catalogue[3]. This is the same catalogue as was used in Bailer-Jones et al. (2019) and is nominally the same subset. Specifically, we select the subset with
`class='GALAXY' AND subClass!='AGN' AND`

[1] `https://data.sdss.org/datamodel/files/BOSS_QSO/DR14Q/DR14Q_v4_4.html`
[2] Done 26 March 2020. Note that this is edr3int2, not dr3int2.
[3] `http://skyserver.sdss.org/CasJobs/SubmitJob.aspx)` with `context="DR15"`

```
DscSelectImplProc.generalProperties.version = 20.4
DscProcessImplProc.generalProperties.version = 20.4

##############################################################
# Only process sources within the magnitude limits
gaia.cu8.dsc.gmin=-5.0
gaia.cu8.dsc.gmax=30.0

############################################################################
# Nominal zeropoint photometric magnitudes
# Copied on 2020-06-23 from
# https://issues.cosmos.esa.int/gaia/browse/DCDREQ-414?focusedCommentId=222974&page=cor
omment-222974
# These must not be changed once Allosmod has been trained!
gaia.cu8.dsc.zeroPointMagG=25.6874
gaia.cu8.dsc.zeroPointMagBP=25.3385
gaia.cu8.dsc.zeroPointMagRP=24.7479

############################################################################
# Specify whether pixels shall be cut or not, and if so, which
gaia.cu8.dsc.cutPixels=true
gaia.cu8.dsc.BPexcluded=1,2,3,4,5,116,117,118,119,120
gaia.cu8.dsc.RPexcluded=1,2,3,4,5,116,117,118,119,120

############################################################################
# Classifier configuration
# Class order must be: quasar, galaxy, star(/anon), wd, binary
# Allosmod prior: p(anon) = p(star) + p(wd) + p(binary)
gaia.cu8.dsc.classPriors=0.000989, 0.000198, 0.988728, 0.000198, 0.009887
# In Allosmod, the star threshold will be used for anon (and wd and binary ignored)
gaia.cu8.dsc.classThresholds=0.5, 0.5, 0.5, 0.5, 0.5

# Allosmod parameters
# Number of GMM components per class. Is the same for all classes.
gaia.cu8.dsc.allosmod.n_components=20
# Names of features used.
gaia.cu8.dsc.allosmod.features=sinb,parallax,pm,uwe,phot_g_mean_mag,bp_g,g_rp,relvarg
# Surces with G < gbrightlim classified as "anon" by Allosmod, with prob [0,0,1]
gaia.cu8.dsc.allosmod.gbrightlim=14.5

############################################################################
# Class names (used in the MDB)
# These must be in the order: quasar, galaxy, star, wd, binary
# as this is assumed by all our codes.
gaia.cu8.dsc.classNames=quasar,galaxy,star,whitedwarf,physicalbinary
gaia.cu8.dsc.unclassifiedClassName=unclassified

###############################################################
# Probability softening to use when combining Allosmod and Specmod probabilities.
# This is a new field.
gaia.cu8.dsc.probSoften=1.0e-8
```

Figure 1: DSC configuration file.

`subClass!='AGN BROADLINE' AND zWarning=0`
which gave 2 312 057 rows. We then performed a crossmatch to Gaia and took the subset in the same way as done for the quasars. This gave 715 739 sources.

Anonymous Objects drawn at random from Gaia edr3int2 that are not in the quasar or galaxy training sets.

White dwarfs All white dwarfs from `http://www.montrealwhitedwarfdatabase.org/tables-and-charts.html` that have coordinates and that are not known to be binaries (using the flag provided in that table; downloaded 2020-04-22). This was cross-matched to Gaia, and sources selected, in the same way as for the quasars. This gave 48 819 sources.

Physical binaries From a set of 200 000 spatially resolved binaries identified in Gaia DR2 that have G differences less than 3.0 mag, the separate BP/RP spectrum of each pair was combined into a composite single spectrum to represent the spectrum of an unresolved binary. (In principle any two BP/RP spectra could be used for this, but components of a binary were taken to ensure they had realistic – if restricted – mass and luminosity ratios, and had the same interstellar extinction.) The original catalogue of spatially resolved binaries was taken from El-Badry & Rix (2018) who identified them using Gaia DR2. The construction of this data set is described in the Jira issue DCDREQ-445 with modifications leading to the actual set used by DSC being described in C8OTH-179. For training DSC the sample was further limited to (a) those sythnesized binaries that have flux ratios between 1.5 and 5, and (b) those for which both components have G$-$RP colours within the range 0–1.3 mag (roughly corresponding to a $T_{\rm eff}$ range of 3000-10 000 K). This gave 89 129 synthesized binaries for training/testing DSC-Specmod. These cuts were imposed to align the sample with that for which other Apsis module MSC (Multiple Star Classifier) is designed. Note that this is the only class which does not use Gaia data directly. As these sources do not exist in the Gaia data, they cannot be used for validation, and we cannot plot their Allosmod features.

3.3 Selected training/testing data and filtering

The above selection gives us a set of cycle 3 `source_ids` (except for the binaries). The actual training/testing data used were obtained from DPCC in January 2021 via the Jira issue DCDREQ-464. These are the same data as were used for ApsisVal3.5g.

To this request we add binaries that are used for validation. These are described in section 3.4. Thus any mention of "binaries" in the current subsection refers to those used for validation.

Only those quasars, galaxies, and anonymous sources that had a complete set of Allosmod features defined (no missing features) were used for training and testing both Specmod and Allosmod. It turns out that all of these also had complete BP/RP spectra. Thus the same set of quasars, galaxies, and anonymous sources were selected for training/testing Specmod and Allosmod. The training data for Specmod includes white dwarfs and physical binaries that Allosmod does not use, because Allosmod is not applied to them. The complete set of the training data (validation data for the binaries) is called the *defset*, as it defines the classes (except for the binaries).

Some minor filtering of the train/test data sets was done before training/testing (see code `dsc_cycle3_ops/code/allosmod_dr3_traintest.R`):

- Remove duplicate sources
- Remove stellar apparent stellar contaminants from the galaxies, by removing all sources from the galaxies that had

  $$\text{G–RP} < 0.3 + 1.1\,\text{BP–G} - 0.29\,(\text{BP–G})^2$$

  This is the same cut as was used in Allosmod on GDR2 public data (Bailer-Jones et al. 2019, equation 1 and Figure 3).
- Remove faint sources from the binary sample (for the validation set) using

  $$G > 16.5 + 2.3\,(\text{G–RP} - 0.55)$$

- Remove bright absolute magnitude sources from the white dwarfs using

  $$G + 5\log_{10}(\varpi/100) < 16 + 6.5\,(\text{G–RP} - 1.0) - 1.5\,(\text{G–RP})^2$$

  where ϖ is the parallax in mas.

This final set of data was then split randomly into a training set and a testing set:
`all_train_data_20210108.csv` with 696 856 sources.
`all_test_data_20210108.csv` with 696 861 sources.
These are used to train/test both Specmod and Allosmod. Allosmod of course ignores the white dwarfs and binaries in these files.

Figures 2 to 5 show the training/testing data for the first four classes, and the validation data for the physical binaries. These are plots of the data obtained in the data request Jira DCDREQ-464

Some additional modifications/selections were applied to the data before training Allosmod (see code `dsc_cycle3_ops/code/allosmod_dr3_traintest.R`).

- Allosmod assigns all sources with $G < 14.5$ mag the probability vector (quasar, galaxy, anonymous) = $(0, 0, 1)$, i.e. it classifies them as anonymous. We therefore removed all such bright sources from the Allosmod training data.
- The feature `sinb` – sine of Galactic latitude – was randomized for quasars and galaxies in the training data, i.e. set to a uniform distribution between -1 and $+1$. This is so that Allosmod does not learn the sky distribution of SDSS.

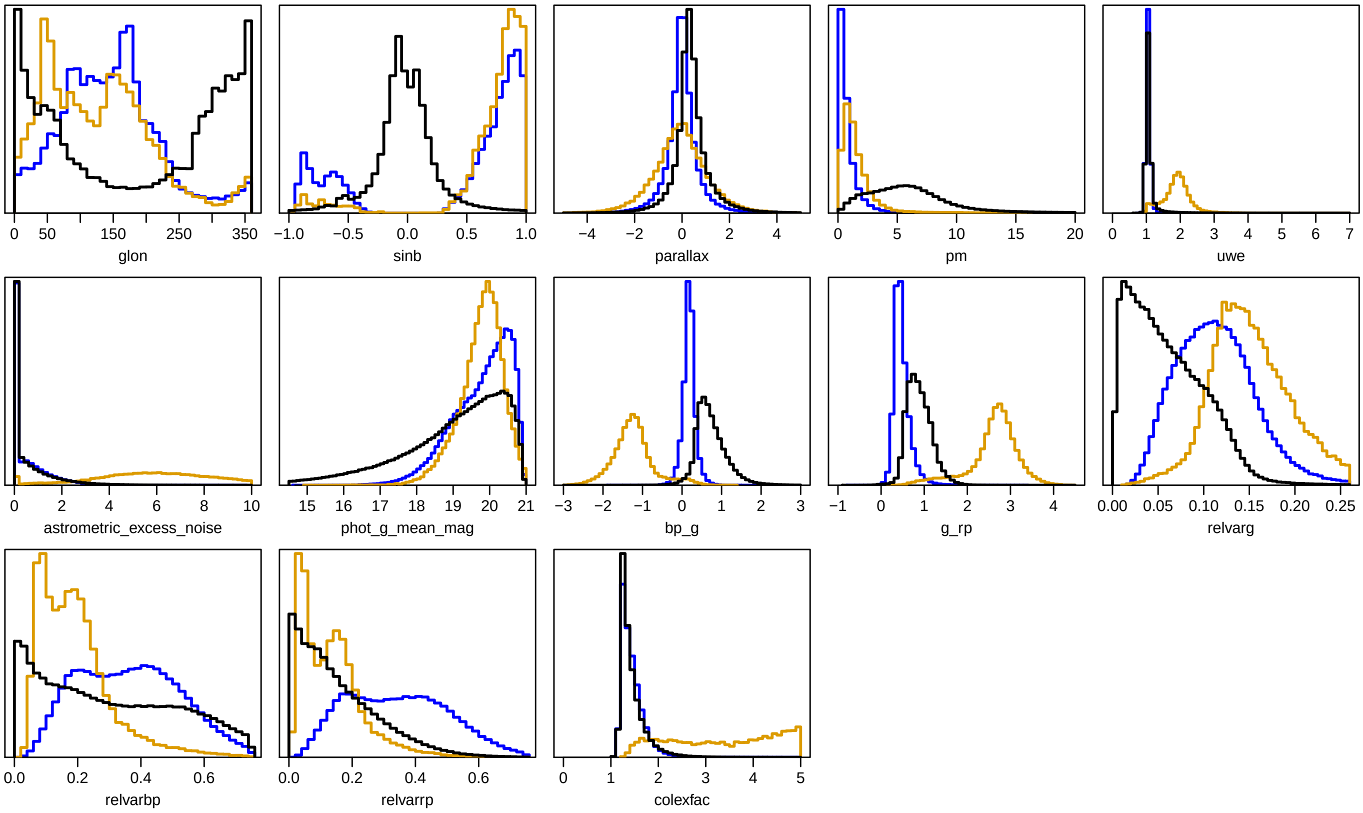


Figure 2: Distribution (linear scale) of various features (including all those used for Allosmod training) of the DSC training/testing sources. Blue=quasar, orange=galaxy, black=anonymous. Units are as in EDR3.

- For the actual training, Allosmod then selects at random, from the remaining sources, 100 thousand quasars, 25 thousand galaxies, and 100 thousand anonymous sources (and likewise for the testing data).

3.4 Validation data

The VST (Validation Source Table) contains all of the sources defined in section 3.2, i.e. both the training and testing data and the remaining sources not selected for either. This statement does not apply to the physical binaries, however, because the train/test physical binaries are not actual Gaia sources. For validating physical binaries we instead use a set of physical binaries identified in LAMOST-DR5 by Xiang et al. (2019) using the "Payne" software. Note that this set itself does not have a high purity. We apply some further selections to retain just higher probability binaries with good signal-to-noise ratio data. The resulting set was crossmatched to edr3int2 (in the same way as explained for the quasars above) and only those sources with `phot_bp_n_obs` and `phot_rp_n_obs` equal to or greater than 5 retained. This gave 332 448 sources. It must be stressed that this set is not very appropriate for validating our binary classifications, however, because it has a completely different definition from our training data. This remains a dilemma

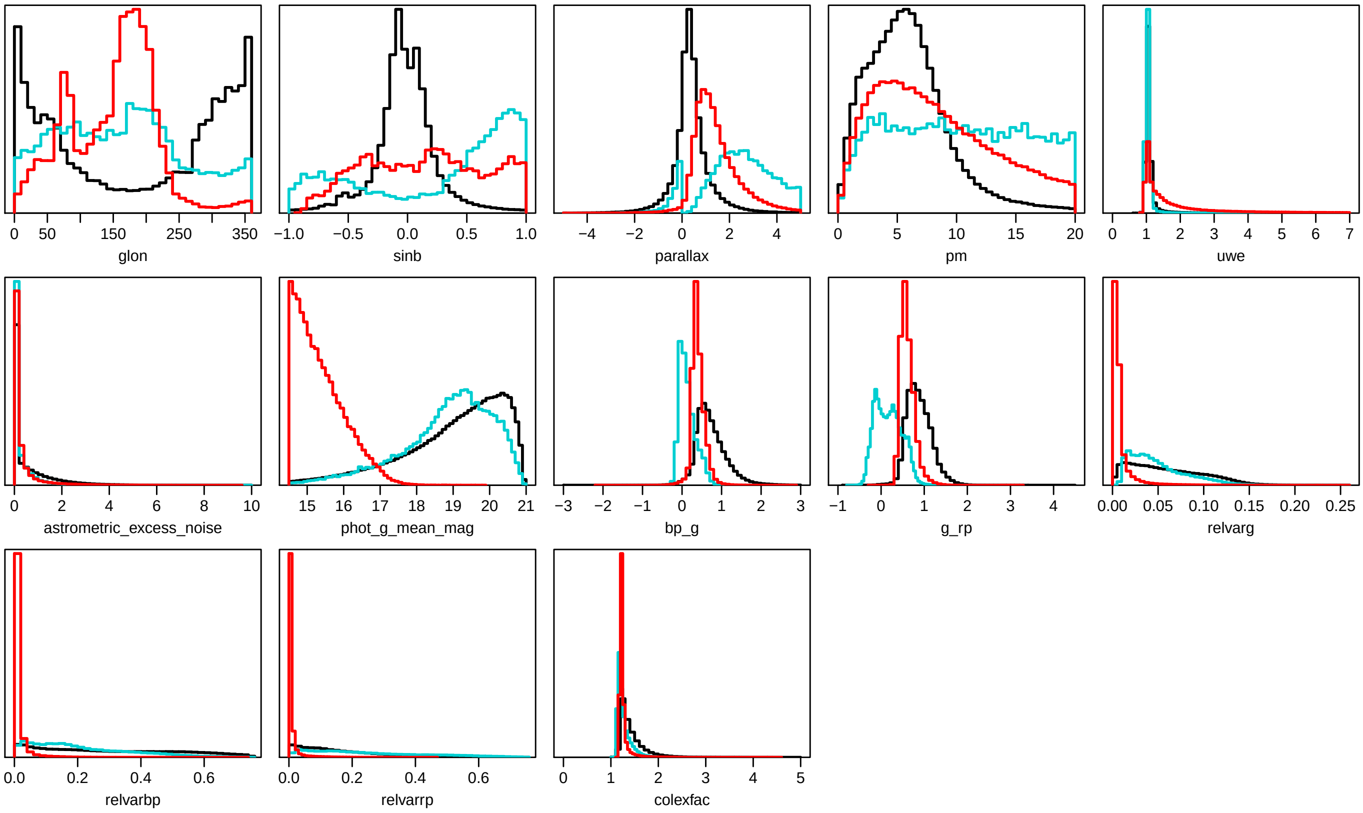


Figure 3: Distribution (linear scale) of various features (those used for Allosmod training, plus others) of the DSC training/testing sources for anonymous sources (black) and white dwarfs (cyan). The red line shows the binaries in the validation data, not in the train/test data (because the binary training set is not a set of Gaia sources). Units are as in EDR3.

which we have not solved for this cycle.

The VST additionally contains 1 612 593 sources that are nominally AGNs as identified by Sergei Klioner and colleagues for the purpose of defining the astrometric reference frame in Gaia EDR3 (Gaia Collaboration et al. 2021b). The original list, `AgnCrossId_3.2_sourceid.csv`, contained 1 614 173 sources, but not all were found in the data used in ApsisOps3.2. The set used here I will refer to as the “Klioner AGN”. Many of these should in principle agree with our “quasar” class, but they have been selected in very different ways and are not the same class of objects.

The VST additionally contains various sources provided by – and defined by – other groups in CU8, as well as 10 million sources selected at random, as described in MFX-003 (see C8CAL-616) but we do not use these.

The DSC VDT (Validation Data Table) for ApsisOps3.2 – i.e. the results from Apsis processing at DPCC on the VST – was provided in several zipped gbin files. These were converted into a single CSV file using a script written by Rene Andrae. The resulting file contains 15 909 324

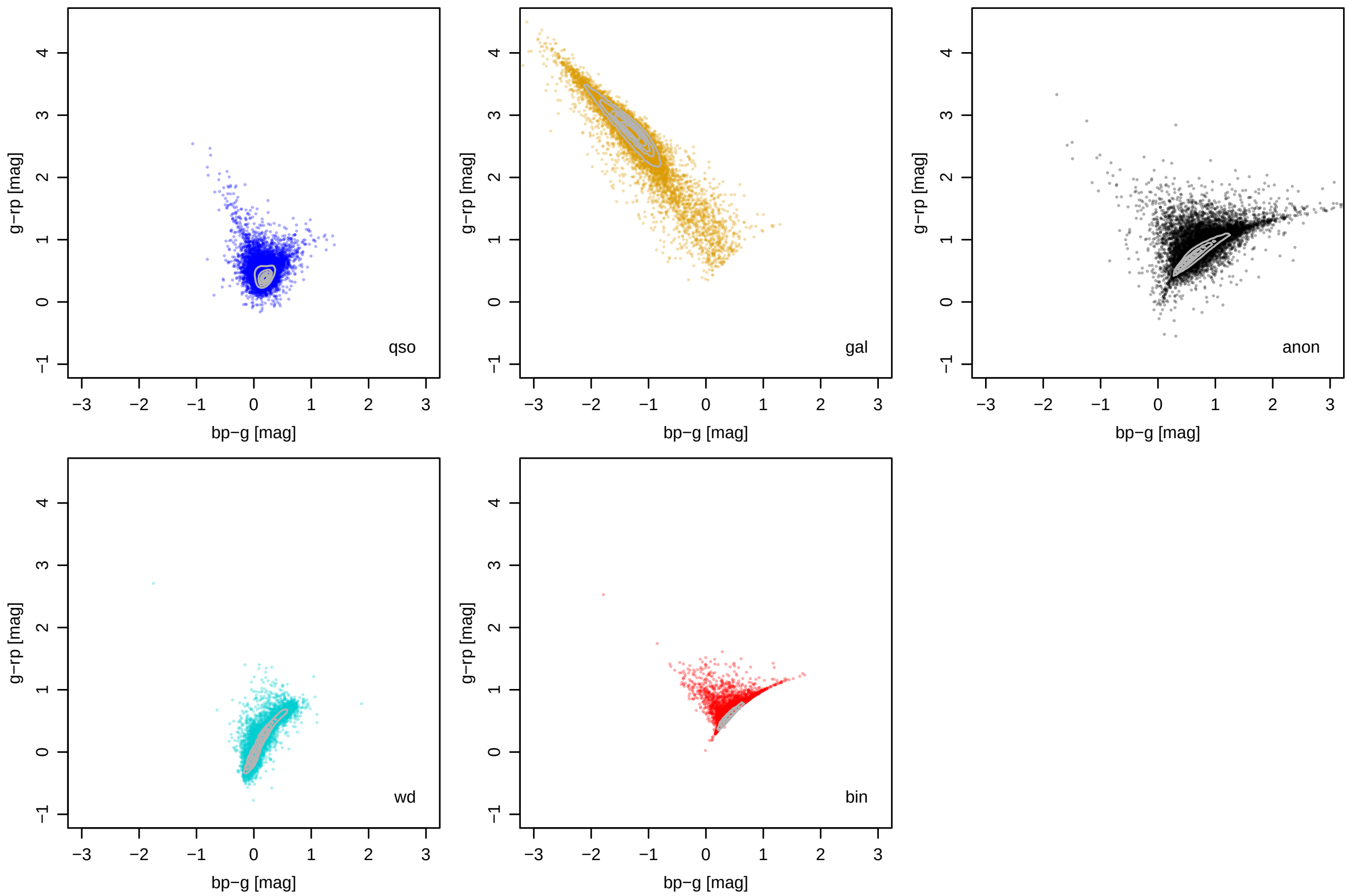


Figure 4: Colour–colour diagram (CCD) of the DSC training/testing sources. The panel for the binaries (bottom right)shows the validation data, not the train/test data.

sources. The fields include all the 13 DSC output probabilities, data from which all the Allosmod inputs can be reconstructed, plus other fields that could be used for diagnostic purposes. DSC nominally provides results on all sources in the VST.

4 Postprocessing: classlabelDscJoint

In June 2021 during the postprocessing we introduced a new class label, via Jira C8CAL-728 (see also the long comment by CBJ on C8MNG-177 on 2021-06-05). This label is defined as:

```
if (P_Specmod(quasar)>0.5 & P_Allosmod(quasar)>0.5)
  then classlabelDscJoint="quasar"
if (P_Specmod(galaxy)>0.5 & P_Allosmod(galaxy)>0.5)
  then classlabelDscJoint="galaxy"
otherwise classlabelDscJoint = "unclassified"
```

This class label is designed to give higher purity subsamples of quasars and galaxies compared to just using Combmod probabilities. The reason this works is that Specmod and Allosmod

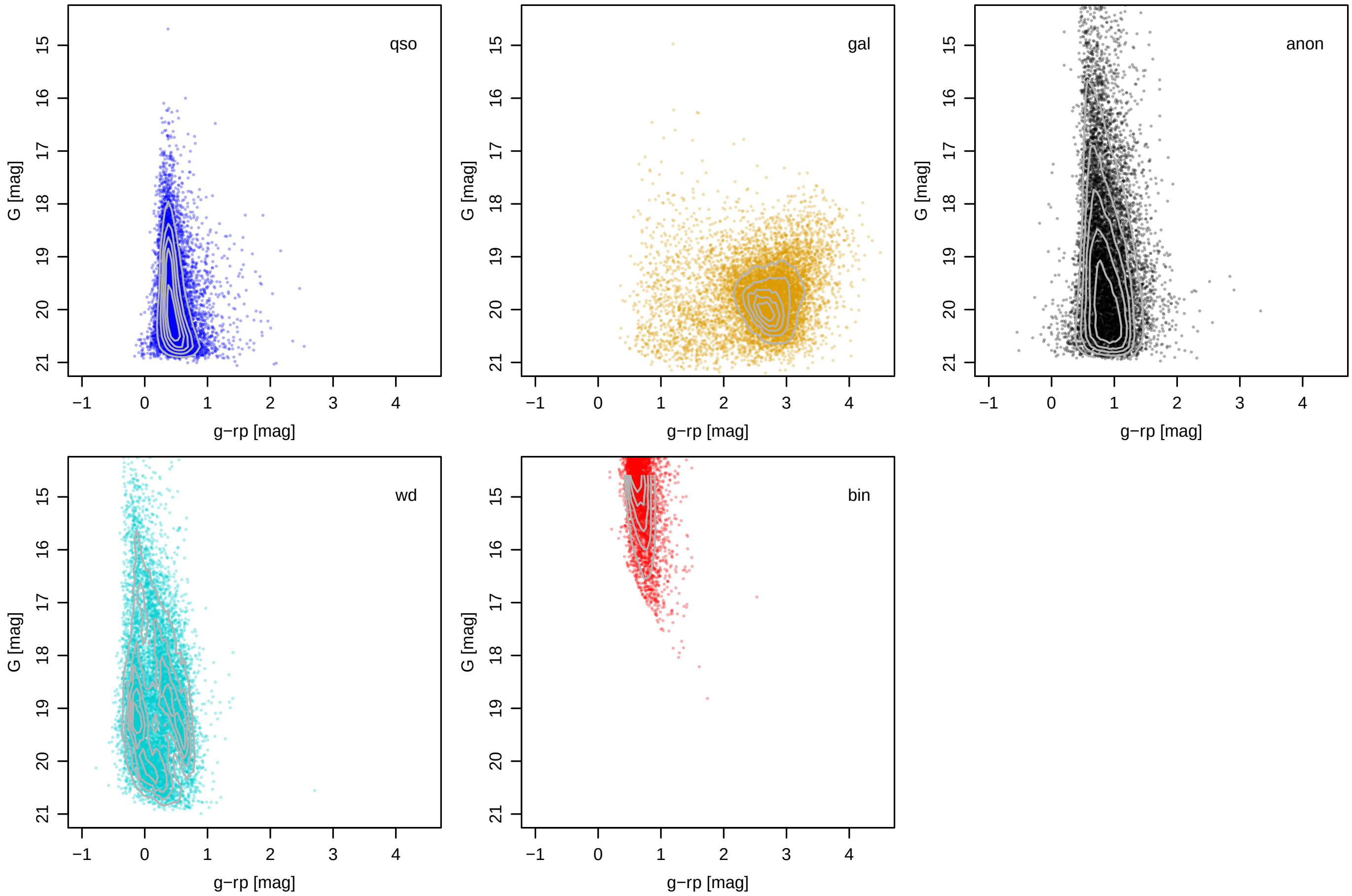


Figure 5: Colour–magnitude diagram (CMD) of the DSC training/testing sources. The panel for the binaries (bottom right)shows the validation data, not the train/test data.

seem to have rather different opinions about what constitutes a quasar or a galaxy, as the following validation will shown. This may be counterintuitive. Yet although they are defined by the same set of training sources (more or less), these two classifiers are based on different data.

In the DSC code, this`classlabelDscJoint` was written to
`SourceProb.getClassifierResults()[2].getFlagArray()`
It will only appear in the QSO and Galaxy tables in GDR3.

5 Known issues

DSC writes a flagArray field for each of the three classifiers. The flag was never intended to be used. DSC appears to write them as follows (see C8CAL-659; not thoroughly checked):

```
For Specmod:
= 1 if missing Allosmod features
= -1 otherwise
```

```
For Allosmod:
= "null" if missing Allosmod features
= "-1" otherwise
For Combmod:
= 1 if missing Allosmod features
= 0 otherwise
```

However, during postprocessing in June 2021 we hacked the Combmod flag to store `classlabelDscJoint`, as described in section 4.

6 Performance metrics

When we have a labelled validation data set, we can define and compute the following two metrics.
Completeness: the number of true positives as a fraction of all objects of that true class.
Purity: the number of true positives as a fraction of all objects of that assigned class.

7 Validation with labelled data: Validation Source Table (VST)

Local directory:
`dsc_cycle3_ops/apsis_ops_3_2_run2/`

The subset of the VST with DSC class labels comprises 2 225 945 sources. Of these, all have valid Specmod results, and 1 457 348 have valid Allosmod results. This latter subset therefore forms the "defset" for the validation (see section 3.3), as it is characteristic of the training data (except for binaries).

Basic checks performed and passed:

- Probabilities range between 0 and 1.
- Probabilities normalized to within 2×10^{-7}.
- Allosmod probabilities are (quasar, galaxy, anonymous) = $(0, 0, 1)$ when $G < 14.5$ mag.
- Allosmod probabilities are missing – (quasar, galaxy, anonymous) = (`NA, NA, NA`) – when (parallax or proper motion is missing) and ($G \geq 14.5$ mag).
- Probability combination (to make Combmod probabilities) done correctly.
- Specmod, Allosmod, and Combmod labels assigned correctly, i.e. to that class with the largest probability that is above 0.5.

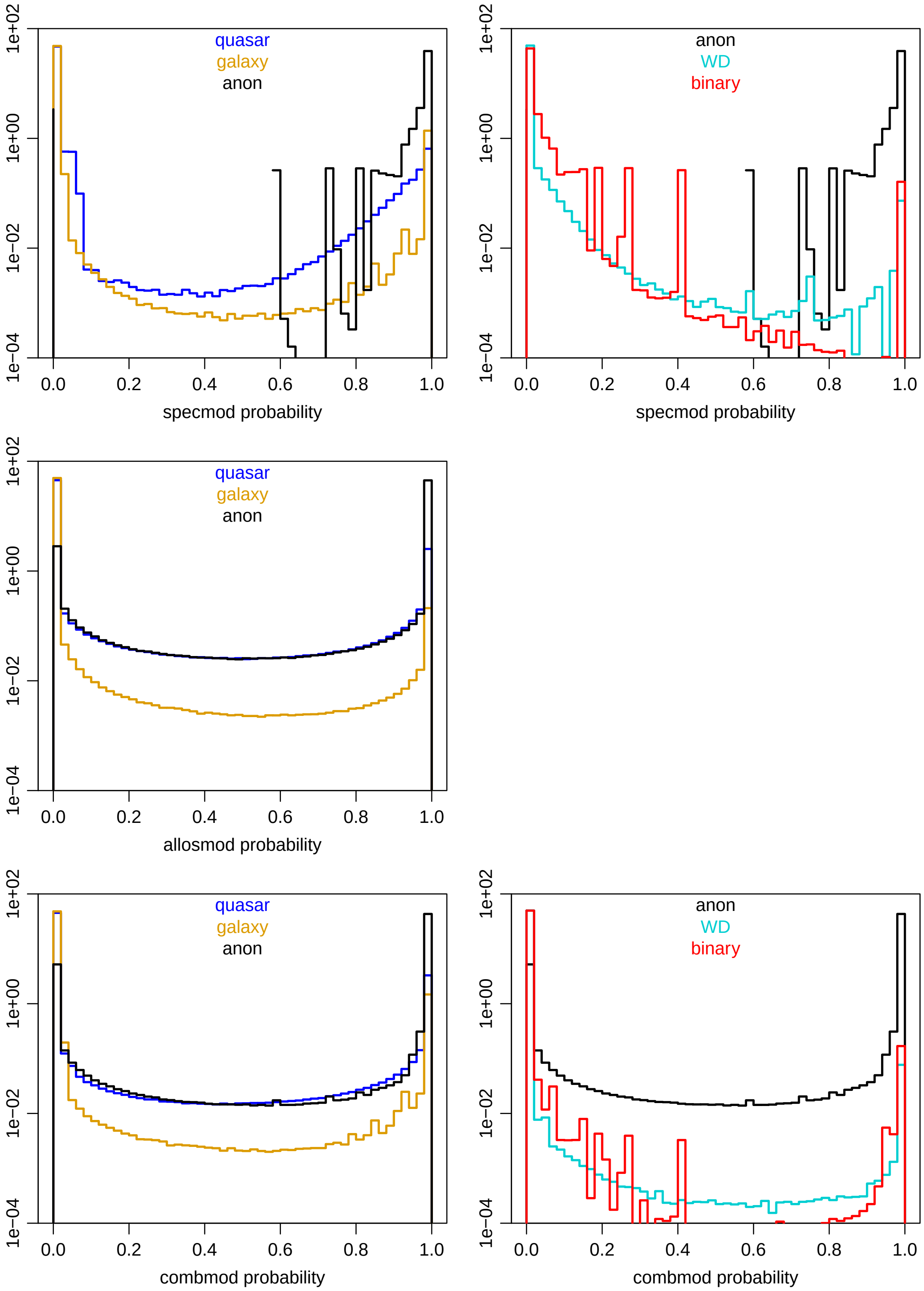


Figure 6: DSC output probabilities for all sources in the VST, where colours indicate the output classes. Top=Specmod, middle=Allosmod, bottom=Combmod. Note the log scale.

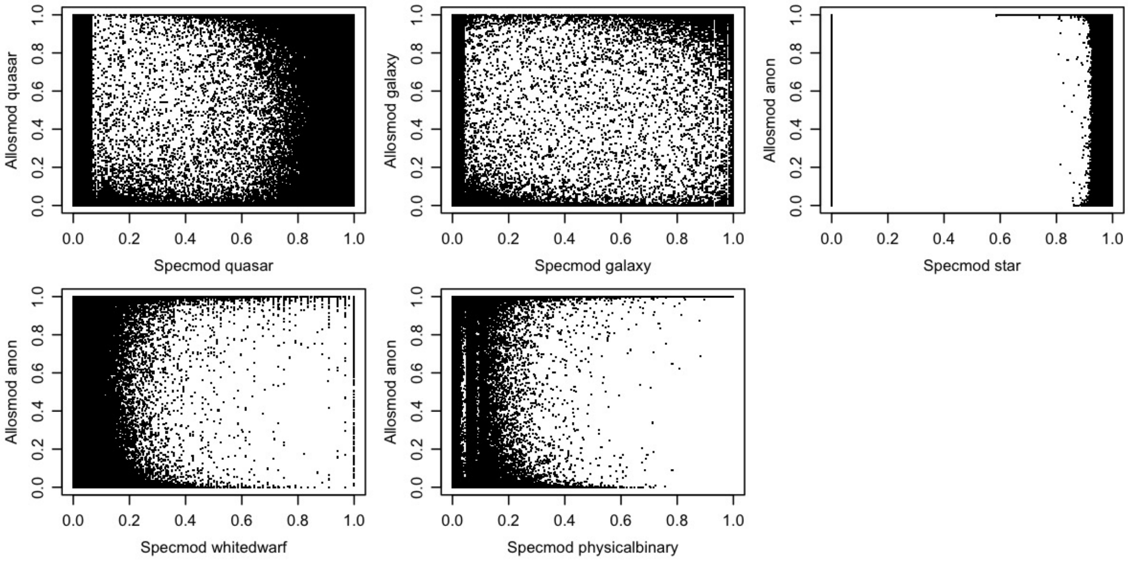


Figure 7: Relationship between Specmod and Allosmod output probabilities for all sources in the VST, one panel per class.

Figure 6 shows the distribution of the output probabilities of each module of DSC into each of its output classes for all sources in VST. Figure 7 shows the relationships between the probabilities of Specmod and Allosmod for each output class. There are no strong correlations, showing that the two types of classifiers have different different opinions about what source should be assigned to which class.

All of the following performance results are just on the defset, i.e. those sources that have both Specmod and Allosmod results.

Figure 8 shows the distribution of the output probabilities of each module of DSC into each of its output classes, just for those sources in the defset with the corresponding true class. Thus the blue line in the top left panel shows the distribution of the Specmod quasar probabilities just for the true quasars in the defset, for example. Note that this plot does not make use of any class assignments by DSC. If we now classified sources using a probability threshold, then the part of the plot above that threshold would show the true positive classifications for that class, and those below would show the false negative classifications.

Note that Figures 6, and 8 use the actual number of objects in the VST/defset, and are normalized histograms (unit area), so they do not represent the class fractions in the real Gaia data.

Classes may be assigned using the highest probability for each source. From this we can construct the confusion matrix (not shown), from which we can in tern can compute the completeness of each true class and the purity of each assigned class (see Bailer-Jones et al. 2019 for a full explanation). These are shown in Table 3. No probability threshold is imposed here (except for the columns “Spec&Allos”, as explained in the caption). If we additionally required the highest probability to be greater than 0.5, then we can get unclassified sources. In practice

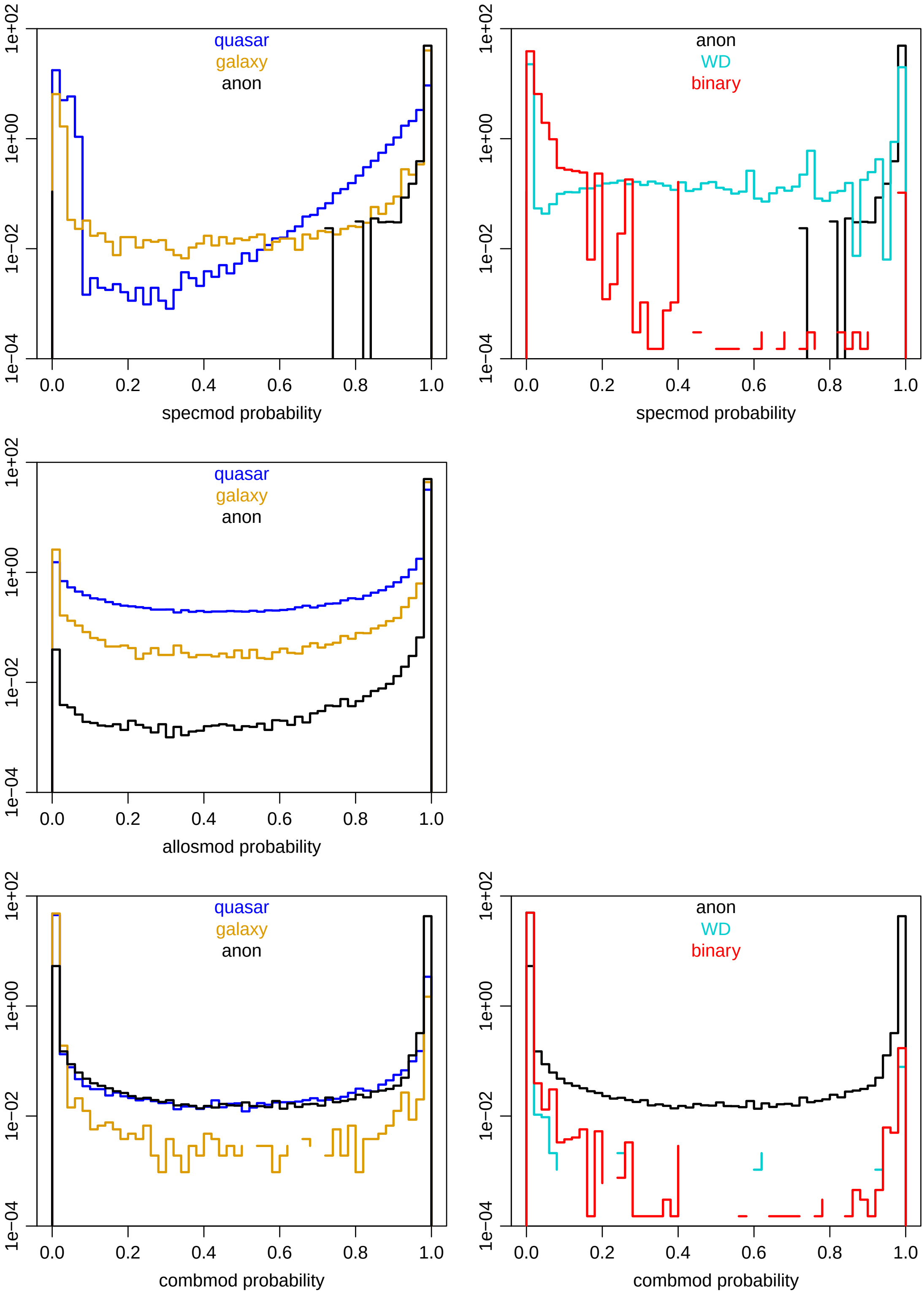


Figure 8: As Figure 6, but the sources plotted for each class are now restricted to those sources in the defset that have that true class. Top=Specmod, middle=Allosmod, bottom=Combmod.

Table 3: DSC performance on the defset, i.e. the sources in the VST that have complete features and so were classified by both Specmod and Allosmod. 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 for each class. As usual, purities have been adjusted to reflect the class prior. Results on the "binary" class are largely meaningless due to the incongruity of the class definitions in the training and validation data sets. The final two columns labelled "Spec&Allos" refer samples obtained by requiring a probability larger than 0.5 from both Specmod and Allosmod for a given class: this is identical to `classlabelDscJoint` defined in section 4.

	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
anonymous	0.998	0.989	0.998	1.000	0.996	0.990	–	–
white dwarf	0.491	0.158	–	–	0.432	0.250	–	–
binary	0.002	0.096	–	–	0.002	0.075	–	–

Table 4: As Table 3, but where the purities have *not* been adjusted to take account of the class prior. This produces spuriously large purities for the non-anonymous classes, because in the VST the anonymous class is severely underrepresented compared to a random Gaia sample. This table is the only place in this report where we do not correct the purities, and is just for demonstration purposes.

	Specmod		Allosmod		Combmod	
	compl.	purity	compl.	purity	compl.	purity
quasar	0.409	0.970	0.992	0.408	0.916	0.976
galaxy	0.831	0.009	0.961	0.298	0.936	0.956
anonymous	0.998	0.569	0.955	1.000	0.996	0.654
white dwarf	0.491	0.975	–	–	0.432	0.984
binary	0.002	0.485	–	–	0.002	0.259

this leads to very few unclassified sources, and if we recompute Table 3 with this threshold the purities and completenesses change by less than 0.001. This means that when classifying by highest probability, the probabilities are almost always above 0.5.

Recall that the purities in these tables have been adjusted to correspond to a sample in which the class fraction equals the class prior (see section 2). Table 4 shows what happens if we do not do this: The purities for all classes except anonymous are significantly larger. This because there are unrealistically few anonymous sources in the VST that can contaminate the other classes. Using a balanced VST would also be wrong here, because anonymous sources are far more common than everything else in reality. Thus Table 4 shows the erroneously optimistic results we would present if we ignored the class imbalance problem, which unfortunately is all too

Table 5: As Table 3, but additionally requiring the class probabilities to be above 0.99.

	Specmod		Allosmod		Combmod	
	compl.	purity	compl.	purity	compl.	purity
quasar	0.127	0.741	0.603	0.540	0.769	0.316
galaxy	0.795	0.464	0.858	0.550	0.905	0.301
anonymous	0.965	0.994	0.993	1.000	0.991	0.990
white dwarf	0.398	0.274	–	–	0.405	0.273
binary	0.002	0.111	–	–	0.002	0.111

Table 6: As Table 3, but only for sources with $G \leq 19.0$ mag.

	Specmod		Allosmod		Combmod	
	compl.	purity	compl.	purity	compl.	purity
quasar	0.706	0.769	0.965	0.770	0.981	0.773
galaxy	0.787	0.595	0.831	0.444	0.857	0.425
anonymous	0.999	0.990	1.000	1.000	0.999	0.990
white dwarf	0.617	0.389	–	–	0.576	0.409
binary	0.002	0.050	–	–	0.002	0.048

common in the literature. All other tables and results in this report make the correction.

Table 3 shows rather modest purities and completeness for all classes except anonymous (which, as pointed out in section 2 is easy to get good results on and is uninteresting). This is the performance we expect for sources selected at random from the entire Gaia dataset that has complete data for Specmod and Allosmod. It accommodates the rareness of all these classes, as specified by the class prior (section 2), both in the probabilities and in the test set.

It is important to appreciate that the rareness of the non-single star classes is a strong driver of the performance of both Specmod and Allosmod. In contrast, the intrinsic ability of these classifiers to separate the classes is quite good: if we had balanced data sets *and* adopted a uniform class prior and balanced test set, then we easily get completenesses and purities in excess of 95% for all classes. But this tells us nothing about our ability to classify sources over the whole Gaia data set.

DSC is intended primarily as a quasar and galaxy classifier. The performance on white dwarfs and especially binaries is not as good, nor as easy to quantify, because (a) the class priors are very uncertain, and (b) the binary training sample that defines this class was created by combining Gaia data on single sources (section 3.2). As our validation data is made up of sources taken directly from Gaia, this means that our binary validation set is likely very different from our training set, so it's not surprising that the performance on binaries in Table 3 is poor.

The final two columns of Table 3 show the performance when we require a probability larger than 0.5 from Specmod and Allosmod for the quasar class, and similarly for the galaxy class. This is equivalent to setting `classlabelDscJoint` (see section 4). We see that this gives us a purer but less complete sample than any of the other selections. The size of the extragalactic samples obtained in this way from all the sources in cycle 3 is explored in section 10.

It may be of interest to identify where, in feature space, we get correct or incorrect classifications. Figure 9 shows the colour–colour diagram (CCD) for the false positives in orange, overplotted over the true positives in black (for orientation), for each class for the three DSC classifiers. Figure 10 shows the same but on the colour–magnitude diagram (CMD).

7.0.1 Performance variation with probability threshold for class assignment

We can trade off completeness and purity by changing the probability threshold that we must exceed in order to assign a class. The left column of Figure 11 shows how the completeness and purity varies for each class (denoted by colours) as a function of that probability threshold, for each of the three classifiers. For Specmod in particular there is little change in performance over a wide range of thresholds between 0.1 and 0.9. When the threshold is above $(1/K)$, where K is the number of classes, a source can remain unclassified. The fraction of unclassified sources is shown in the right column of Figure 11. Changing the threshold has a similar affect as changing the class prior, in the sense that both raising the threshold and lowering the prior will classify fewer objects into a given class. So both generally reduce the completeness and increase the purity.

A tabulation of the completeness and purity when using a threshold of 0.99 is shown in Table 5. Note that the resulting samples for quasars and galaxies are not as pure as when we also require `classlabelDscJoint` to be set, which corresponds to the “Spec&Allos” column in Table 3.

7.0.2 Performance variation with magnitude

Fainter sources generally have poorer data (lower signal-to-noise), so we might expect performance to be better when examining brighter sources. A tabulation of the completeness and purity for just those sources with $G \leq 19.0$ mag is given in Table 6. This has of course involved a different adjustment to the purities than used before, because the number of sources we actually have – and so the class fractions – varies with magnitude. However, both this adjustment (and the posterior probabilities themselves) are for the class prior in Table 2, which was set for the entire DSC data set of 1.79 billion sources over all magnitudes. If we confine our analysis to a specific magnitude range, then we would probably want to use a different class prior. In particular, for these brighter sources we probably want to use lower prior probabilities for quasars and galaxies. We have not changed the prior here, however.

We investigate this dependence on magnitude in more detail by computing completeness and purity for sources in bins of width of 1 mag. Classes are assigned by maximum probability above a threshold of 0.5. This is shown in Figure 12. The bright bins in particular have very

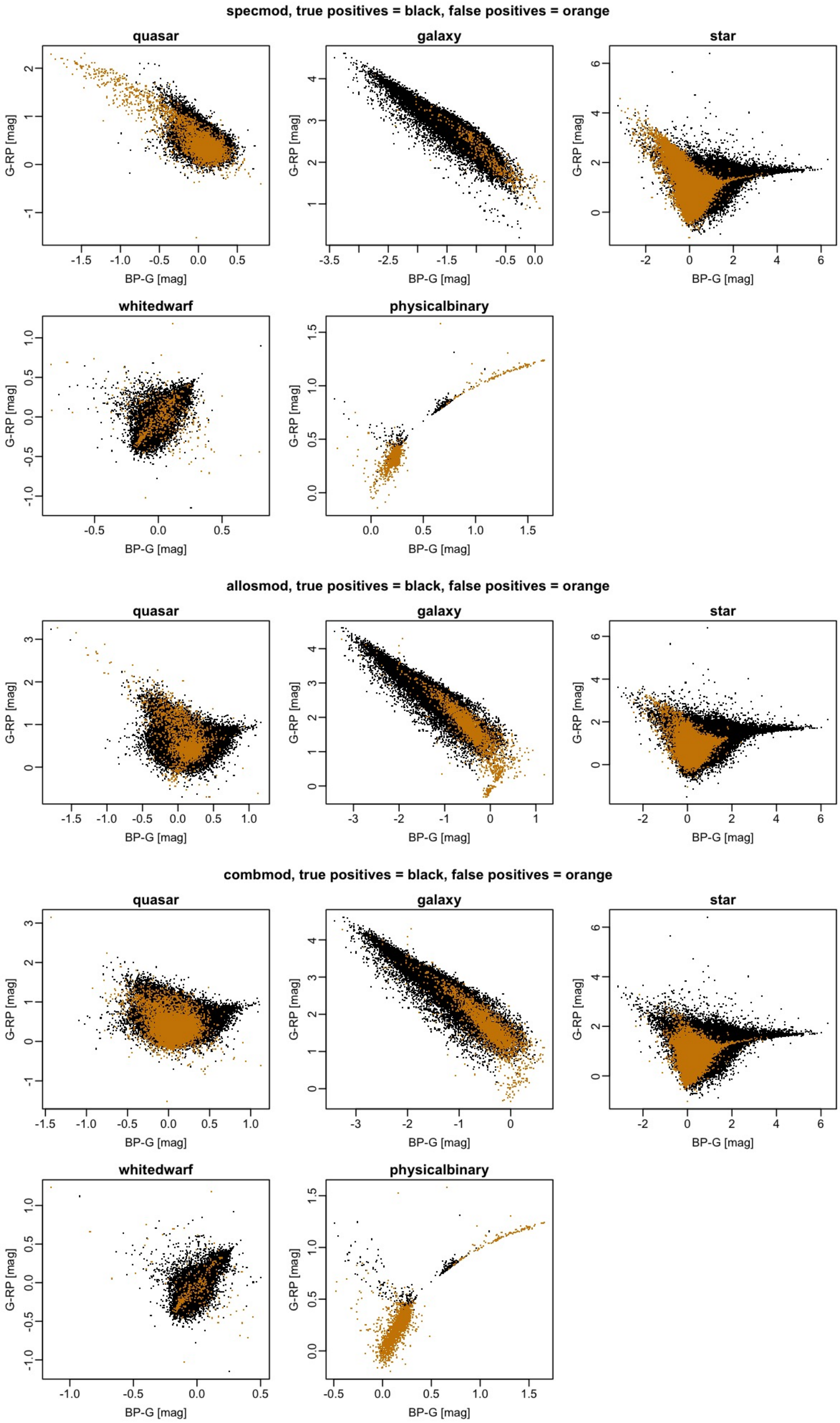


Figure 9: Dataset: defset. Colour–colour diagrams of true positives (black) and false positives (orange) for each class (one per panel) in the defset for Specmod (top), Allosmod (centre), and Combmod (bottom). Classes are assigned by the largest posterior probability.

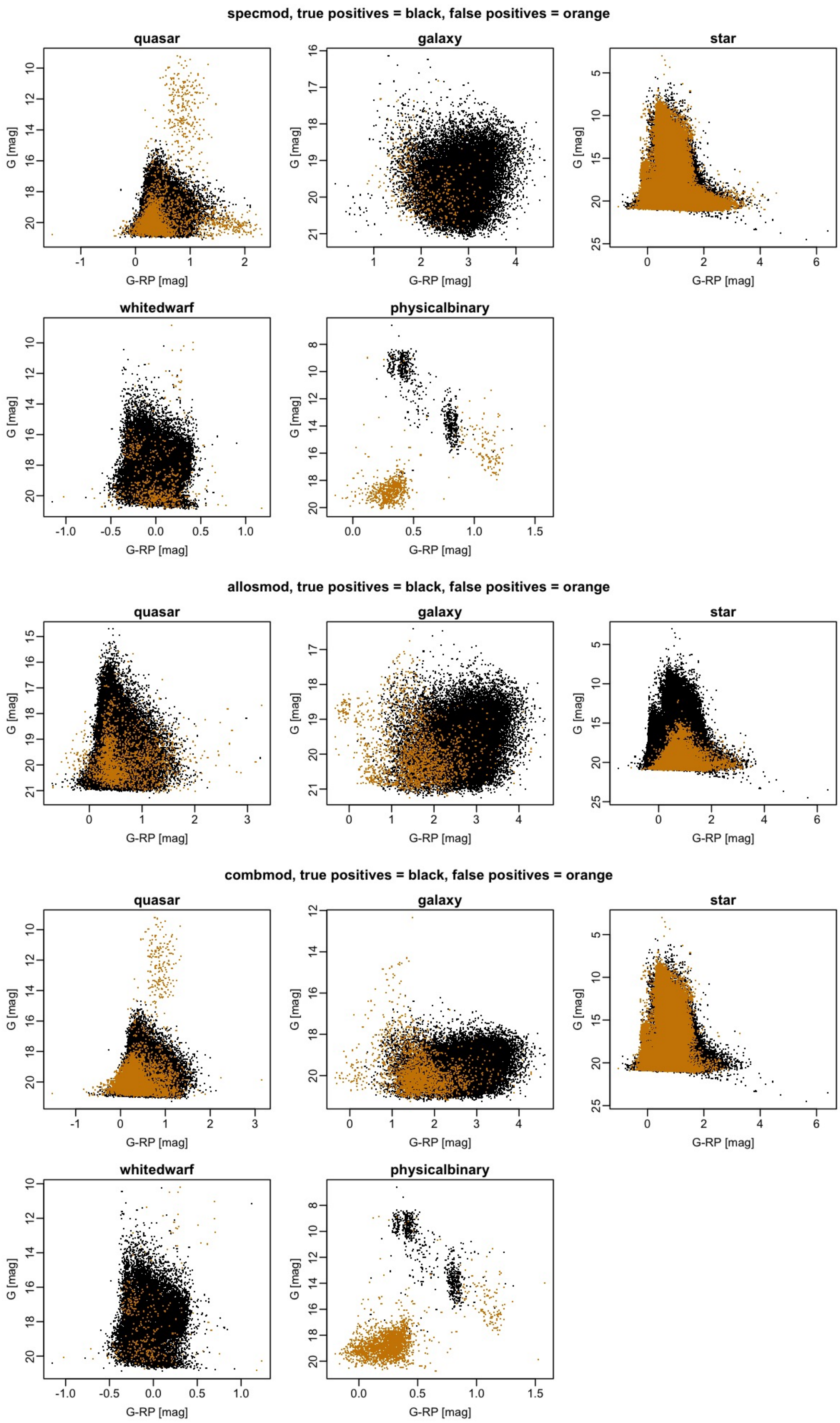


Figure 10: As Figure 9, but showing colour–magnitude diagrams.

Table 7: Number of sources with probabilities above 0.5 for each class in each of the classifiers. “Any” refers to any of the classifiers being above 0.5. “Spec&Allos” means the probabilities from both Specmod and Allosmod must exceed 0.5 (this sets `classlabelDscjoint` for quasars and galaxies). “All” means the probabilities from all three classifiers must exceed 0.5. For these last two rows, no numbers are given for white dwarfs and binaries because Allosmod does not give results for them.

	quasar	galaxy	anonymous	white dwarf	binary
Specmod	2 883 624	2 839 487	1 780 439 858	1 069 114	457 858
Allosmod	1 934 875	999 258	1 368 541 759	–	–
Combmod	5 454 722	3 663 012	1 777 400 591	601 502	647 107
Any	5 756 678	3 823 580	1 783 890 223	1 107 965	654 634
Spec&Allos	547 374	251 091	1 365 091 394	–	–
All	547 245	251 091	1 363 970 127	–	–

few sources in them for some classes, as can be seen from Figure 13, so the completenesses and purities are quite noisy in some of those. Generally speaking we see a drop in purity towards fainter magnitudes, as expected.

The same plots but using a probability threshold of 0.99 are shown in in Figure 14 and 15.

7.0.3 Performance variation with the minimum number of transits

Figure 16 shows how the completeness and purity vary with the minimum number of transits used in the data (i.e. only retaining sources with a minimum number). More transits generally implies better (higher signal-to-noise ratio) data, and so we would expect the purity to increase with minimum number of transits. We see that this is the case once we have more than 8 or 16 transits, except sometimes for a very large number of transits due to very low number statistics. As usual, the purity has been adjusted taking into account the actual class fraction in the test set for the sample in question. Note that sources with more transits will tend to be brighter, and for brighter sources we may want to adopt a different class prior. This is not done here. The main point of these plots is to show that performance does not degrade significant for a very low number of transits – below 8 – and therefore that we should not filter DSC results based on the number of transits.

8 Analysis of entire set of DSC outputs

Analysis done using the CU8 VDB using table `ops32_dsc`.

DSC produces outputs for 1 787 815 254 sources. The numbers of sources that achieve probabilities above 0.5 for each class in each of the classifiers are given in Table 7. Note how the selection using Combmod gives a much larger number of quasars and galaxies than either Specmod or Allosmod alone, or indeed than their sum. This is because the combination of Specmod and Allosmod probabilities that form the Combmod probabilities is not a simple logical "and" or "or", but rather a proper combination that respects the priors (see CBJ-090, and the examples in appendix A. We also see that the number of quasars and galaxies obtained when we require both Specmod and Allosmod probabilities to be greater than 0.5 (row "Spec&Allos") is far lower than what we get in Combmod, or indeed in either Specmod or Allosmod. This implies that Specmod and Allosmod only have a small intersection of extragalactic objects for which they agree on the class (at a probability of 0.5). This is curious, because they are trained on the same set of sources (more or less), although the data are different.

9 Analysis of unlabelled data: random subset of full data

Local directory:
`dsc_cycle3_ops/apsis_ops_3_2_run2/cu8_validation_database/`

We randomly select 0.1% of all results from the CU8 VDB:

```
SELECT * FROM (SELECT * FROM ops32_dsc
  TABLESAMPLE BERNOULLI(0.1) REPEATABLE (20210506)) AS tmp
JOIN ops32_asp ON tmp.sourceid = ops32_asp.sourceid
```

This returns 1 788 509 sources.

Basic checks performed and passed:

- Probabilities range between 0 and 1.
- Probabilities normalized to within 2×10^{-7}.
- Allosmod probabilities are (quasar, galaxy, anonymous) = $(0, 0, 1)$ when $G < 14.5$ mag.
- Allosmod probabilities are missing – (quasar, galaxy, anonymous) = (`NA, NA, NA`) – when (parallax or proper motion is missing) and ($G \geq 14.5$ mag).

Figures 18, 19, and 20 show the CCD, CMD, and CQD for these data, where the panels of course refer to the assigned classes, as the true classes are unknown.

Figure 21 shows the number and fraction of sources that are classified with a probability above 0.5 (into any class) as a function of G magnitude. The black line indicating the Combmod classifier is barely distinguishable from the magenta line showing the Specmod classifier. This is because almost all sources have BP/RP spectra, whereas some lack features required for Allosmod, typically the parallaxes and proper motions, and so have no Allosmod results.

Figure 22 shows the number and fraction of sources with a probability above 0.5 for each class that are brighter than a given magnitude. Note that at the bright end there are very few sources, as can be seen in the left panels, so the corresponding fractions in the right panels are noisy. The actual numbers for the full data set will be around 1000 times larger than those shown in the left panels.

10 Analysis of unlabelled data: the selection for the quasar and galaxy tables

Local directory:
`dsc_cycle3_ops/apsis_ops_3_2_run2/cu8_validation_database/`

Here we analyse the subset of extragalactic sources selected from the CU8 VDB according to

```
(P_Specmod(quasar)>0.5 OR P_Allosmod(quasar)>0.5
  OR P_Combmod(quasar)>0.5) OR
(P_Specmod(galaxy)>0.5 OR P_Allosmod(galaxy)>0.5
  OR P_Combmod(galaxy)>0.5)
```

which, in terms of the field names used in the CU8 VDB is

```
SELECT * FROM ops32_dsc WHERE
((classprobxpquasar>0.50        OR classprobxpgalaxy>0.50) OR
 (classproballosquasar>0.50     OR classproballosgalaxy>0.50) OR
 (classprobcombinedquasar>0.50 OR classprobcombinedgalaxy>0.50))
```

This returns 9 475 642 sources. Note that this is larger than the sum of the number of "quasar" and "galaxy" sources for selection "Any" in Table 7 – which is 9 580 258 – because those two "Any" selections can (and do) have common sources, whereas the above selection only includes sources once.

Note that some sources do not have all features, so do not appear in the all panels. In particular, the majority of galaxies do not have parallaxes and proper motions. This is shown in Figure 23. Whenever any of these features are missing there are no Allosmod results.

Figure 24 shows the distribution of 8 features (those used to train Allosmod) for the quasars and galaxies in the extragalactic sample, here classified when the corresponding Combmod probability is above 0.5. This can be compared (in Figure 25) to the purer subset obtained when demanding that both the Specmod and Allosmod probabilities also be above 0.5, i.e. the `classlabelDscJoint` label is also set (section 4). Recall that this subsample has higher purity but lower completeness (column "Spec&Allos" of Table 3). By comparing these two figures to the distribution of features in the training sample, shown in Figure 26, we see that the purer sample is closer to the training sample than the full sample. The same can be see when looking at the colour–colour diagrams (Figures 27 to 29) and colour–magnitude diagrams

(Figures 30 to 32). Figure 33 plots the CMDs and CCDs for the full extragalactic sample for quasars and galaxies in a single panel to ease comparison.

Figure 34 shows the cumulative distribution of the number of objects classified as quasars (blue) and galaxies (orange) brighter than some G magnitude. Classification is done by requiring the Combmod probability to be above 0.5 for that class. As the full sample includes sources for which the Specmod or Allosmod probabilities can also be greater than 0.5, the full sample includes a few objects for which the Combmod probability is greater than 0.5 for the non-extragalactic classes (anonymous, white dwarfs, binaries), but the lines for the classes can be ignored. There are of course very few bright extragalactic objects, although there are a few with $G < 17$ mag, as no classifier is perfect. We see in these distributions the Gaia magnitude limit at around $G = 21$ mag, although the number of galaxies continues to grow slightly to fainter magnitudes.

Figure 35 also show the G magnitude distributions, but now non-cumulatively. Note that the full sample, in black, includes a few very bright contaminants (note the log scale) which are obviously not extragalactic sources. But their presence is entirely consistent with the low purity of this sample, and they could be removed simply by imposing a magnitude cut. There are also some very faint sources, the exact nature of which will be unreliable due to the very low signal-to-noise data from Gaia. This plot also shows how the magnitude distribution changes (from the black line to the green line) as we additionally require the Specmod and Allosmod probabilities to be greater than 0.5, which leads to the purer sample. Note that this removes all sources brighter then $G = 14.5$ mag by design, because anything brighter is classified as "anonymous" by Allosmod. It also removes the faintest sources: apparently these can get a probability above 0.5 in Combmod, but not in both Specmod and Allosmod. Most importantly, we see that the magnitude distribution for this sample (in green) is very similar to the magnitude sample of the training data, which is another indication that this sample has high purity, as it follows closely the class definition defined by the training data.

In the same way as for the magnitude, Figure 36 shows the cumulative distribution of the number of objects classified above some absolute Galactic latitude. Note how the number of quasars per unit area increases away from the Galactic plane to a latitude of about 30 degrees, then drops and levels off. This probably indicates a higher contamination by stars at low latitudes. The galaxy sample does not show this feature, even though it is expected to have a similar purity overall (Table 3). This explanation for the quasars appears to be confirmed when look at the same plot for the purer sample, i.e. for which `classlabelDscJoint` has also been set, shown in Figure 37: We no longer see the sharp drop in quasar density at around 30 degrees.

Figure 38 shows the cumulative distribution of the number of objects classified above some probability threshold, p. The selection is not a very strong function of p until we get to very high probabilities. For this reason, varying p is probably not a very effective way of trading off sample completeness and purity in practice. This is why we introduced the class label `classlabelDscJoint` (Table 4).

Figure 39 shows the sky distribution for all sources with a quasar Combmod probability greater than 0.5. Figure 40 immediately underneath it shows the sky distribution for the subset with `classlabelDscJoint = quasar`, i.e. Specmod and Allosmod probabilities are also both above 0.5 for quasars. Note how this purer set, which is also about ten times smaller, has very few sources at low Galactic latitudes. This is unsurprising, because at low latitudes we have far more stellar contaminants, so a conservative selection is likely to remove them. Figures 41 and 42 show the same for galaxies. Note the overdensity in the Magellanic clouds. Obviously we don't expect quasars or other galaxies to have a true high density there: this is just an inevitable consequence of the high density of sources and therefore also of contaminants.

11 Analysis of unlabelled data: white dwarfs and binaries

Local directory:
`dsc_cycle3_ops/apsis_ops_3_2_run2/cu8_validation_database/`

Here we analyse the subset of sources selected from the CU8 VDB according to

```
P_Combmod(whitedwarf)>0.5 OR P_Combmod(binary)>0.5
```

which, in terms of the field names used in the CU8 VDB is

```
SELECT * FROM ops32_dsc WHERE
(classprobcombinedwhitedwarf>0.50 OR classprobcombinedbinary>0.50)
```

This returns 1 248 609 sources, 601 502 satisfying the white dwarf condition, 647 107 the binary condition, which are mutually exclusive (Table 7).

White dwarfs and binaries are only classified by Specmod. Their Combmod probabilities are not identical to their Specmod ones, however, because we apply probability softening to avoid numerical errors when probabilities are exactly zero (which can otherwise make normalization impossible; see CBJ-090). Thus the probabilities are similar, but usually not identical.

We were unable to assess the performance of DSC on binaries using the VST in section 7 for the reasons given there: the validation set is not even approximately representative of the training set. The sources classified here as "binaries" are therefore those Gaia sources that were closer to the binary training sample than to the training samples of any of the other classes, once the class prior had been taken into account. This may well deviate from some other definition of unresolved binaries. In general the DSC results for binaries should be used with great caution.

Figures 43 and 44 compare the distribution of the Allosmod features for the selected sources with their training data. Figures 45 and 46 show these in a CCD and Figures 47 and 48 show these in a CMD. Recall that the binary training objects are composite Gaia BP/RP spectra, so there are no Gaia sources corresponding to these and thus no Gaia features. For the white dwarf, we see some differences between the distributions in these features for the training sources and those found with Combmod probability above 0.5. The latter extend to higher parallaxes and proper motions and also include some bluer sources.

Figures 49 and 50 shows the sky distribution for all sources with Combmod probabilities greater than 0.5 for the classes white dwarf and binary respectively. Note that the density does not increase significantly towards the Galactic plane, as it does with the anonymous sources (compare to Figure 55). This suggests that the contamination from anonymous sources may not be dominant. One might interpret this to mean all these white dwarf and binaries are nearby – within a few hundred pc sources have a near uniform sky distribution – but their parallax distributions in Figure 43 suggest that they are generally as distant as the anonymous sources (compare to Figure 51), so this is not the reason.

12 Analysis of unlabelled data: anonymous

Local directory:
`dsc_cycle3_ops/apsis_ops_3_2_run2/cu8_validation_database/plots/analyse_subset/anon/`

The vast majority of objects found are of class anonymous, which are essentially stars. Achieving a star sample with high completeness and purity is of course trivial for this reason alone. They were only included as a class in DSC to enable the model to define the regions of feature space that are dominated by the other classes. (Intrinsic overlaps in feature space cannot be avoided, of course, and ultimately limit the purity that can be achieved.) An analysis of the anonymous class performance is therefore uninteresting. But for comparison purposes, it it useful to see similar plots for the anonymous class as those shown for the other classes. These are in Figures 51 to 55. These plots have been made using the 0.1% random subset obtained in section 9. We see the distributions for these results are very similar to their training data. The main difference is that the results extend to fainter magnitudes. No faint magnitude limit was imposed for the training data, although the 99th and 100th percentiles for the anonymous sources in the training data are at $G = 20.8$ and 24.5 mag respectively, whereas for this random subset of results they are at 22.3 and 25.6 mag respectively. This difference is probably because the training data was constructed from dr3int2.

13 Validation with labelled data: Klioner AGN

Local directory:
`dsc_cycle3_ops/apsis_ops_3_2_run2/`

Here we examine the DSC results on the Klioner AGN in the VST (see section 3.4. We look at which of these are classified as (a) quasars, and (b) non-quasars. Under the (incorrect) assumption that DSC should classify them all as quasars, set (a) is the true positives, and set (b) is the false negatives (their union is the total set). This assumption is incorrect because the DSC quasar training set is not a similar set of quasars as the Klioner AGN set. In particular, the Klioner set excludes extragalactic objects with large parallaxes and proper motions, yet we use precisely this "failure" of AGIS to help us identify extended extragalactic sources. We do not

Table 8: The first three columns show the fraction of the Klioner AGN assigned to each class by each DSC classifier when assigning classes by largest probability. The last two columns given the fraction classified when we require a probability greater than 0.5 in Specmod and Allosmod (penultimate columns) and in all three classifiers (last column).

	Specmod	Allosmod	Combmod	Spec&Allos	All
quasar	0.315	0.684	0.791	0.277	0.278
galaxy	0.007	0.023	0.029	0.006	0.006
anonymous	0.677	0.292	0.180	0.254	0.175
white dwarf	0.000	–	0.000	–	–
binary	0.000	–	0.000	–	–
unclassified	–	–	–	0.463	0.542

"purge" our training set in the way the Klioner AGN set has been purged. Consequently, even if our classifier were perfect at what it does, we would expect it to get an unknown fraction of "false" negatives, and also miss some of the Klioner objects. Note that we also cannot investigate quasar false positives (nor true negatives) with this set, because we are using a set which is nominally all of one class: We have no *known* contaminants in the Klioner set, although it surely does contain some.

The first three columns of Table 8 show the fraction of Klioner AGN classified into the different classes by the three DSC classifiers when assigning classes by maximum probability (so there are no unclassified objects). We see that the percentage of objects classified as quasars is 32% in Specmod, 68% in Allosmod, and 79% in Combmod. (These numbers change by less than 0.1% if we additionally use a threshold of 0.5.) Note that Combmod finds more than either Specmod or Allosmod individually. In building this table there is no adjustment for class fractions in the data set, because it is a test set of nominally all one class (we are measuring completenesses, not purities). Note that the posterior probabilities used to make the class assignment do still accommodate the class prior, as it *always* the case (this is done internally in DSC before it writes out probabilities). Assuming (incorrectly, as argued) that all the Klioner AGN should be classified as quasars, then the missed objects (false negatives) are predominantly anonymous. We might think that many should be misclassified as galaxies, given the blurred distinction between some galaxies and quasars in the optical. But galaxies have such a small prior that it's hard to classify anything as a galaxy without strong evidence. The anonymous class, in contrast, has a prior of nearly 1.0, so unless there is a good evidence of being a quasar, the source will be classified as a star.

Figures 56 and 57 show the location of these "true" positives and "false" negatives in the CCD and CMD (respectively) for the three classifiers. As before we can vary the threshold used for classification. This changes the fraction of sources assigned to each DSC class. This is shown in Figure 58.

The penultimate column of Table 8 shows the fraction classified into each class when we demand that the corresponding class probabilities are above 0.5 in both Specmod and Allosmod (equivalent to `classlabelDscJoint` be set to the corresponding class: section 4). This is a more conservative (and purer) selection and so reduces the percentage of quasars found to 28%, with 25% being classified as anonymous, 0.6% as galaxies, and the remaining 46% are unclassified. If we additionally require the Combmod probability for the class to be above 0.5, then the percentage classified as quasars drops further to 18% (final column of Table 8).

14 Conclusions

DSC is designed primarily to identify quasars and galaxies among the far more numerous stars. Consequently it expects extragalactic objects to be very rare, and this must be represented in the class prior. Performance depends strongly on the choice of prior, which is hard to fix. The prior has an uncertainty of a factor of two for quasars and galaxies, and much more (perhaps ten) for white dwarfs and physical binaries. The DSC classes are defined by the training data, so when DSC's performance on some other catalogue of objects called "quasars" or "galaxies", for example, may differ significantly from what we obtained on the VSTs.

Within this context, the performance on quasars and galaxies seems reasonable, although certainly too poor for some applications. Assuming, as we have done, that quasars have a frequency (class prior) of 1 in 1000, then the purity of a quasar sample classified by largest posterior probability is 24% (from Combmod). The completeness is 92% (also Combmod). Completeness can in principle be traded off against purity by adopting a probability threshold for classification. It would be desirable to adopt a large threshold in order to get a purer sample, because it is difficult to even call a set of objects "quasars" when we expect 3/4 of them not to be quasars. However, we find that neither completeness nor purity vary rapidly with threshold probability other than at the extremes. The threshold would have to be increased above 0.99 to get a quasar purity above 32%, for example. But it may not be wise to use the probabilities very close to their extremes, so we probably cannot create a quasar set with an arbitrarily large purity (which would anyway become asymptotically small). A higher purity can be achieved by using just Allosmod, however: 41%, compared to 25% from just Specmod.

Galaxy performance is similar to quasar performance, despite the fact that we have adopted an even small class prior of 1 in 5000 (remember that "galaxy" in Gaia is limited to unresolved sources). Completeness is 94% and purity is 22% (from Combmod). If we just used Specmod we get a lower completeness of 83% but a larger purity of 41% (and this is arguably a net gain).

Performance improves if we limit the classification to brighter sources (Figure 12), although we of course then have fewer extragalactic objects (Figure 22)

To get higher purity samples of quasars and galaxies, we introduced `classlabelDscJoint`. This is set to "quasar" if both the Specmod and Allosmod probabilities for quasars is greater than 0.5, and likewise for galaxies. This gives a sample with a purity of 62% for quasars and 64% for galaxies (the "Spec&Allos" column in Table 3). The price we pay is a significantly lower completeness for quasars (38%) although interestingly not for galaxies (still high at 83%).

`classlabelDscJoint` will be written to the Galaxy and QSO tables in GDR3 to enable users to extract this subset.

Performance is likely to improve if we used additional information, e.g. infrared photometry, or if we narrow the definition of the class so that it occupies a smaller volume of parameter space and thus overlaps less with other classes. We could also be creative with the priors – raising the prior will improve the results – but we must not fool ourselves or report "optimistic" performance.

The completeness and purity of white dwarfs are 43% and 25% respectively, although the prior adopted is very uncertain. Results for physical binaries appear to be useless, although this is hard to assess given the way this class is defined, and the fact that the validation set is not representative of the training data. Probably we could do better on both these classes if Allosmod were extended to include them, but the DPAC schedule did not permit this. Furthermore, it is very difficult to get a reliable and sizeable sample of unresolved binaries for training.

15 References

Bailer-Jones C.A.L., Fouesneau M., Andrae R., 2019, Quasar and galaxy classification in Gaia Data Release 2, *Monthly Notices of the Royal Astronomical Society*, 490, 5615

Bailer-Jones C.A.L., Fouesneau M., Andrae R., 2021, Description of DSC in Cycle 3, GAIA-C8-PL-MPIA-CBJ-090-D-7

Gaia Collaboration (R. Smart et al.), 2021a Gaia Early Data Release 3: The Gaia Catalogue of Nearby Stars, *Astronomy& Astrophysics*, 649, A6

Gaia Collaboration (S.A. Klioner et al.), 2021b Gaia Early Data Release 3: Acceleration of the Solar System from Gaia astrometry, *Astronomy& Astrophysics*, 649, A9

Xiang M., et al., 2019, Abundance estimates for 16 elements in 6 million stars from LAMOST DR5 low-resolution spectra, *Astrophysical Journal Supplement Series*, 245, 34

A Probability combination examples

CBJ-090 defines the algorithm for combining the Specmod and Allosmod probabilities to produce the Combmod probabilities. Although straight forward, the results may be counter-intuitive. There are at least three reasons for this: (1) both Specmod and Allosmod include separately the prior, whereas Combmod only includes it once; (2) we add a small value ϵ to all probabilities in the combination to avoid divide by zero in the normalization ("probability softening"); (3) results are sometimes missing (due to missing input data). The following examples illustrate this, in all cases using $\epsilon = 10^{-8}$, as in the actual results. All probabilities are rounded to four significant figures. The five elements of the Combmod probabilities reported below are always in the order: quasar, galaxy, anonymous, white dwarf, binary.

If a source has P(quasar)=P(galaxy)=0.5 in both Specmod and Allosmod (P=0 for all other classes) their corresponding Combmod probabilities are not 0.5. They are instead
`1.668e-01 8.332e-01 6.540e-20 1.310e-23 6.540e-22`
This is due to the non-uniform prior (we are using the values in Table 2), because the Specmod and Allosmod already probabilities are posterior probabilities that include this prior. If we instead use equal priors, then the posterior is 0.5 for quasar and galaxy, to within the limits set by the probability softening:
`5.000e-01 5.000e-01 2.222e-17 2.222e-17 2.222e-17.`

We can investigate this behaviour further by changing the prior. For example, if we had equal results for the first two classes from Specmod and Allosmod, but unequal priors for these

```
specmodprobs=c(0.5, 0.5, 0, 0, 0)
allosmodprobs=c(0.5, 0.5, 0)
classprior=c(2, 1, 1, 1, 1)
```

then the Combmod probabilities are
`3.333e-01 6.667e-01 2.963e-17 2.963e-17 2.963e-17`
(the prior is written unnormalized for convenience). The class with the *smaller* prior (galaxies, second position) gets the *larger* posterior after combination. This is because if the class with the smaller prior could achieve the same posterior (in both classifiers), then it must have had a larger likelihood in both classifiers, and so when combined probabilistically, these will play a larger role than the (single) prior.

Another example of this likelihood effect – which is what we want – can be seen when we have P=0.5 for galaxies and anonymous in both classifiers and zero for all others classes:

```
specmodprobs=c(0, 0.5, 0.5, 0, 0)
allosmodprobs=c(0, 0.5, 0.5).
```

With the prior
`classprior=c(1, 1, 1, 0, 0)`
the Combmod probabilities are
`2e-16 5e-01 5e-01 0e+00 0e+00`
i.e. equal for galaxies and anonymous, as we would expect. Recall that the anonymous class in Allosmod corresponds to the union of the anonymous, white dwarf, and binary classes in Specmod. So if we now change the prior to

```
classprior=c(1, 1, 1, 1, 1)
```

then for the sake of the combination the Allosmod posterior probability for anonymous is spread across the three classes in Specmod (anonymous, white dwarf, binary) in proportion to the prior. So the Combmod probabilities are now

```
3.6e-16 9.0e-01 1.0e-01 2.0e-09 2.0e-09.
```

Galaxies become more probable and anonymous less probable. The effect is even stronger when we use the actual class prior in Table 2. The Combmod probabilities are then

```
8.007e-17 9.998-01 1.962e-04 7.858e-16 3.924e-14,
```

i.e. galaxies are even more probable for the same Specmod and Allosmod probabilities. This is because this prior has galaxies 5000 times rarer than stars, so to have achieved a posterior probability of 0.5 in both Specmod and Allosmod, both likelihoods must have been very high, and these combine (with the single prior) to give a large posterior in Combmod.

If Allosmod probabilities are missing, the Combmod probabilities are equal to the Specmod ones, for any value of the prior. If Specmod probabilities are missing, then for the combination the Allosmod probability for anonymous is spread across the three classes in Specmod (anonymous, white dwarf, binary) in proportion to the prior. With $1/3$ probabilities for all Allosmod classes, and using the class prior in Table 2, the Combmod probabilities are

```
3.333e-01 3.333e-01 3.300e-01 6.608e-05 3.300e-03.
```

If Allosmod indicates a source is anonymous, and Specmod indicates that the white dwarf and binary probabilities are equal, then providing the prior is equal for these two classes, their Combmod probabilities will also be equal. For example, with

```
specmodprobs=c(0.07, 0.03, 0.2, 0,3, 0.3)
allosmodprobs=c(0, 0, 1)
classprior=c(2, 3, 4, 1, 1)
```

then the Combmod probabilities are

```
9.000e-09 2.571e-09 5.714e-01 2.143e-01 2.143e-01.
```

Of course, if the priors for those two classes are not equal then their Combmod probabilities will be different. Indeed, as their Specmod probabilities are identical in the example above, and Allosmod does not distinguish between them, their Combmod probabilities will be in proportion to their priors. If we double the (unnormalized) white dwarf prior probability in the previous example, so that the prior is

```
classprior=c(2, 3, 4, 2, 1)
```

and change nothing else, then the Combmod probabilities become

```
1.009e-08 2.882e-09 4.706e-01 3.529e-01 1.765e-01
```

i.e. the last two are in a ratio of exactly two (to within the rounding errors of the numbers shown).

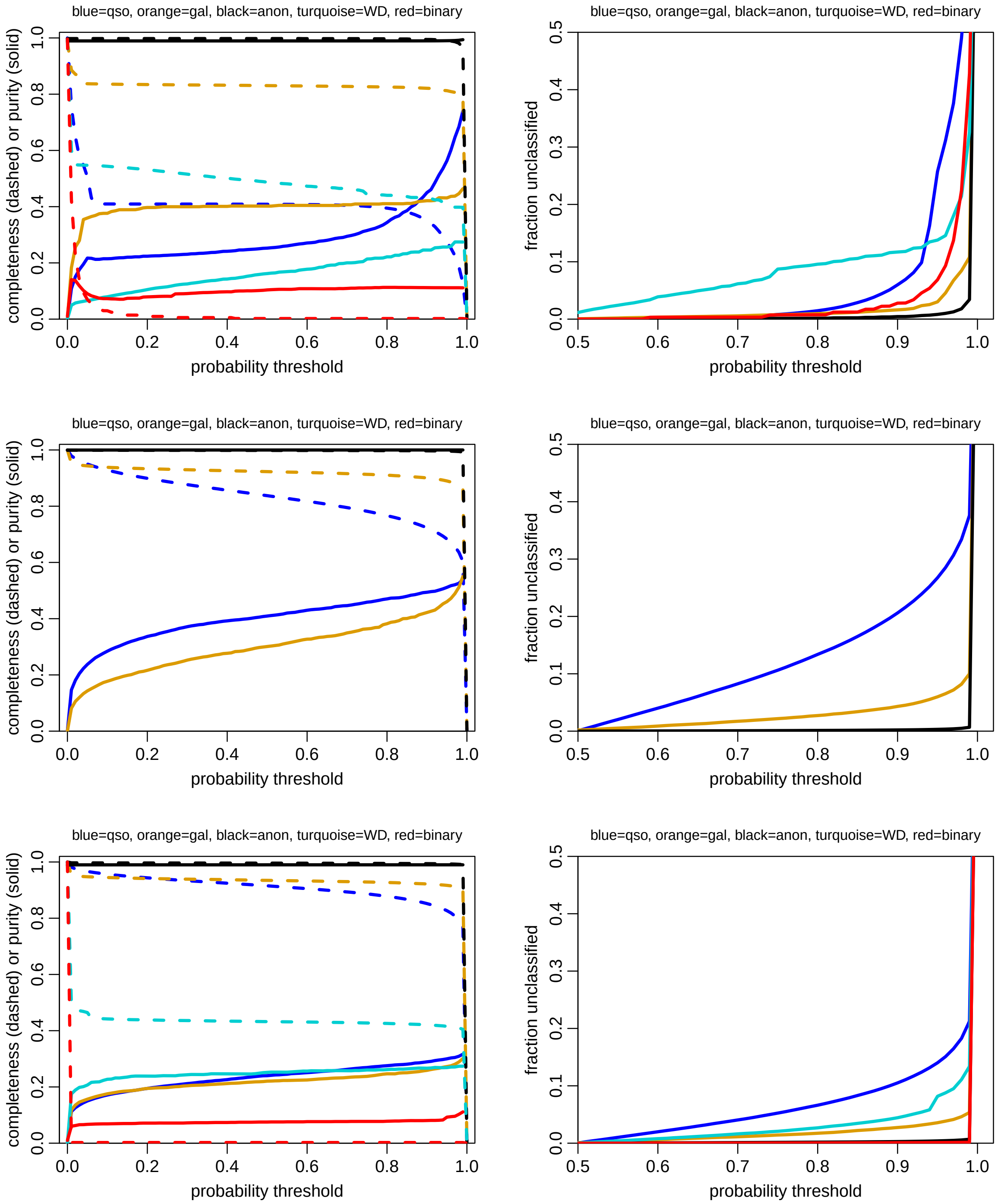


Figure 11: Dataset: defset. Variation of classification completeness and purity (left) and unclassified fraction (right) in as a function of probability threshold for each class (different colours) for Specmod (top), Allosmod (centre), Combmod (bottom).

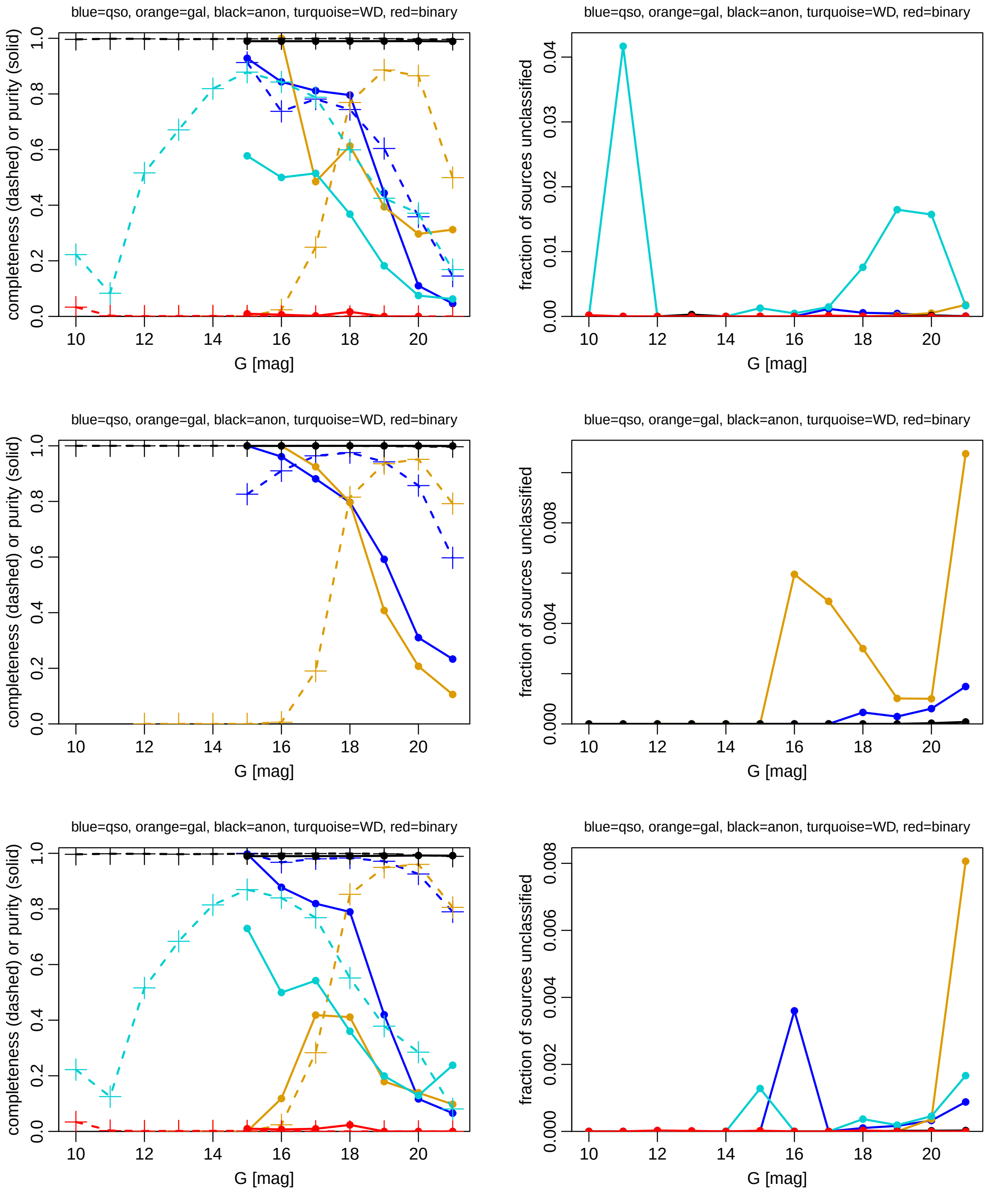
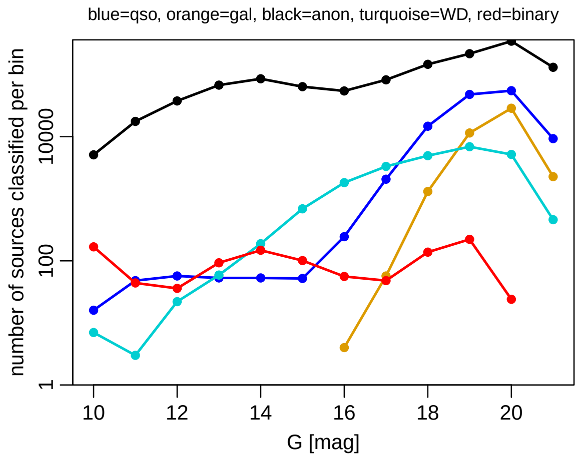
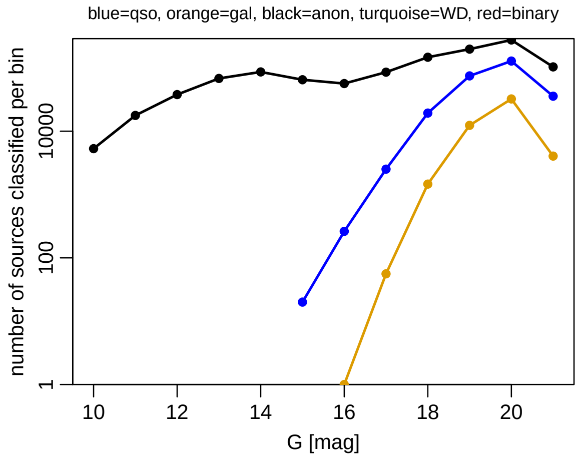


Figure 12: Dataset: defset. The left column shows the variation of classification completeness (dashed lines) and purity (solid lines) as a function of G magnitude for each class (different colours) for Specmod (top), Allosmod (centre), Combmod (bottom). The value shown are for all source in a 1 mag brightness range centered on the point shown; the lines are just for convenience to help identify trends. Classes are assigned by maximum probability above a threshold of 0.5, so can result in some unclassified objects: the fraction left unclassified (in each magnitude bin) is shown in the right column.





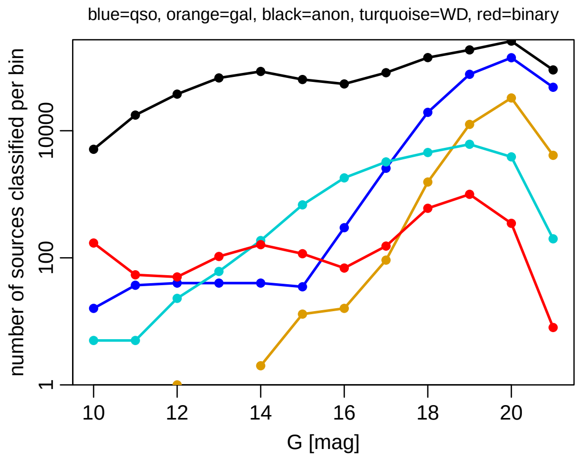


Figure 13: As in Figure 12, but showing the number of sources classified in each magnitude bin on a log scale.

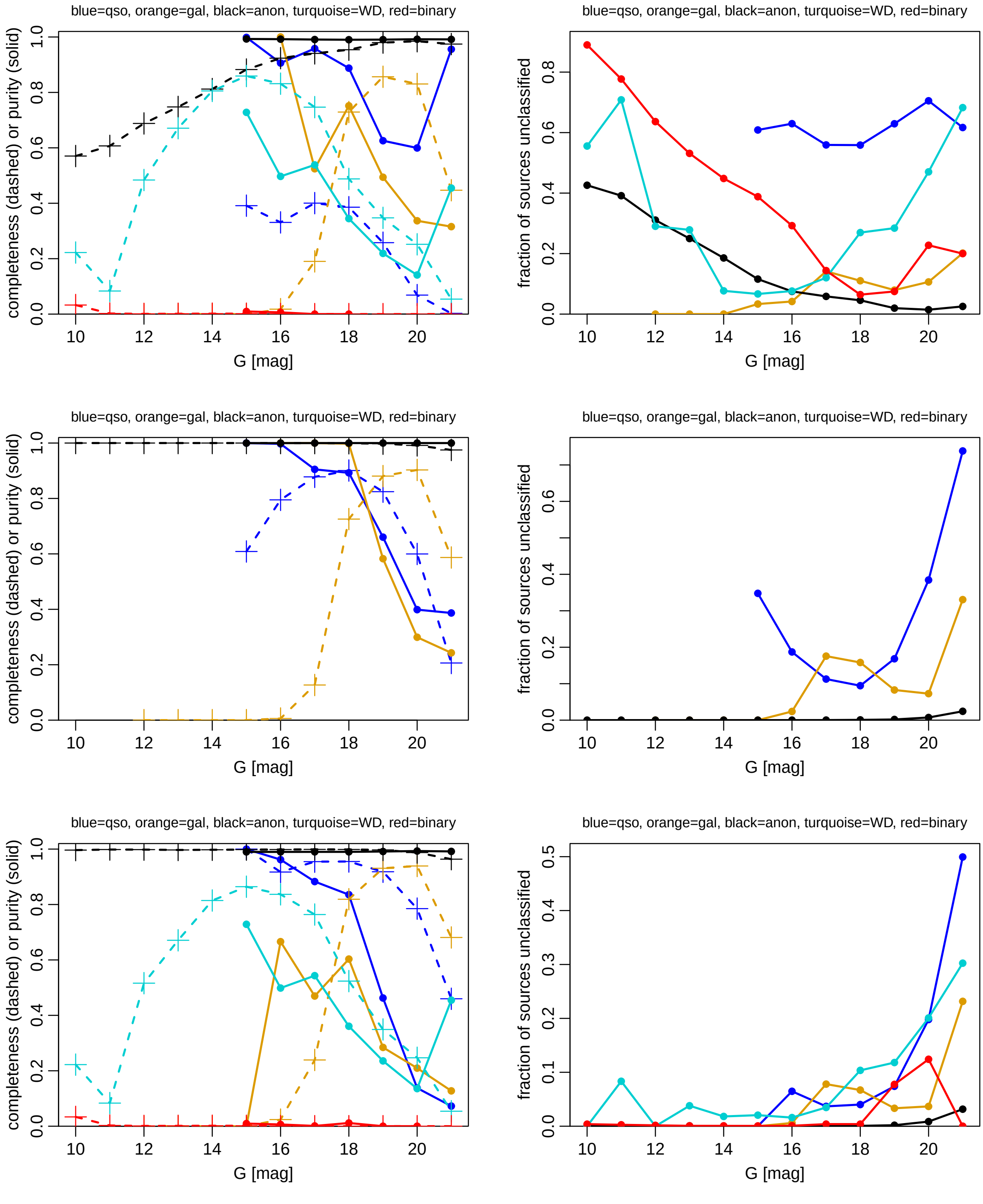


Figure 14: As Figure 12, but for a probability threshold of 0.99.

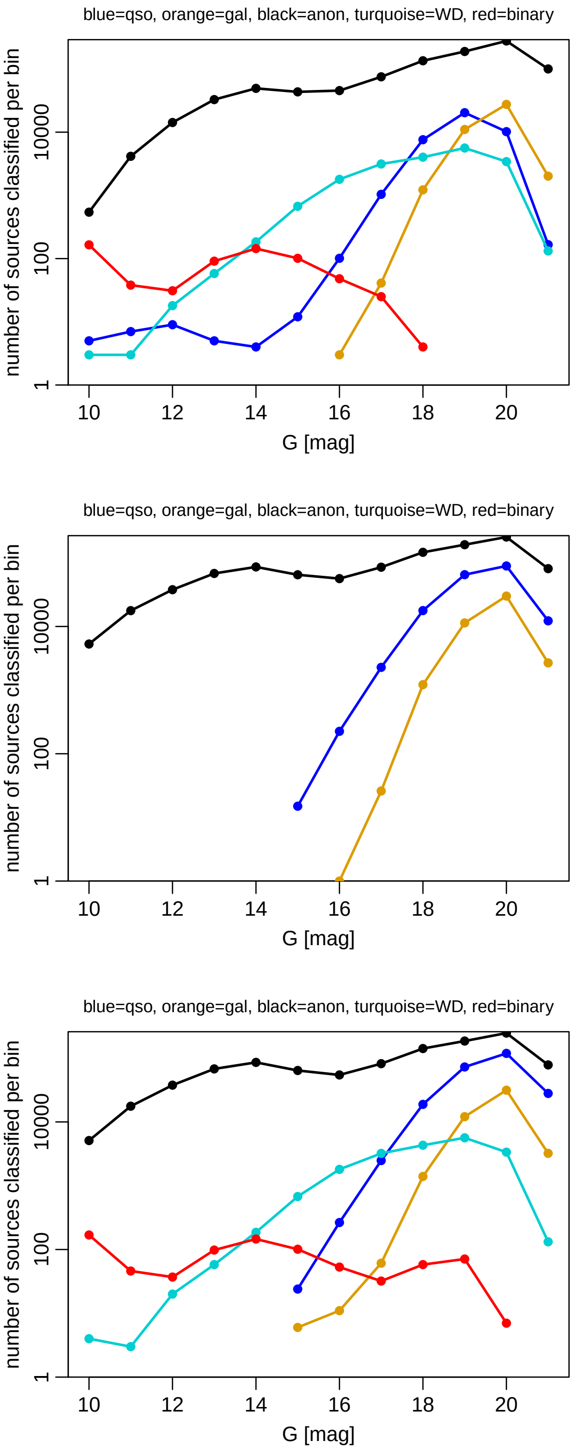


Figure 15: As Figure 13, but for a probability threshold of 0.99.

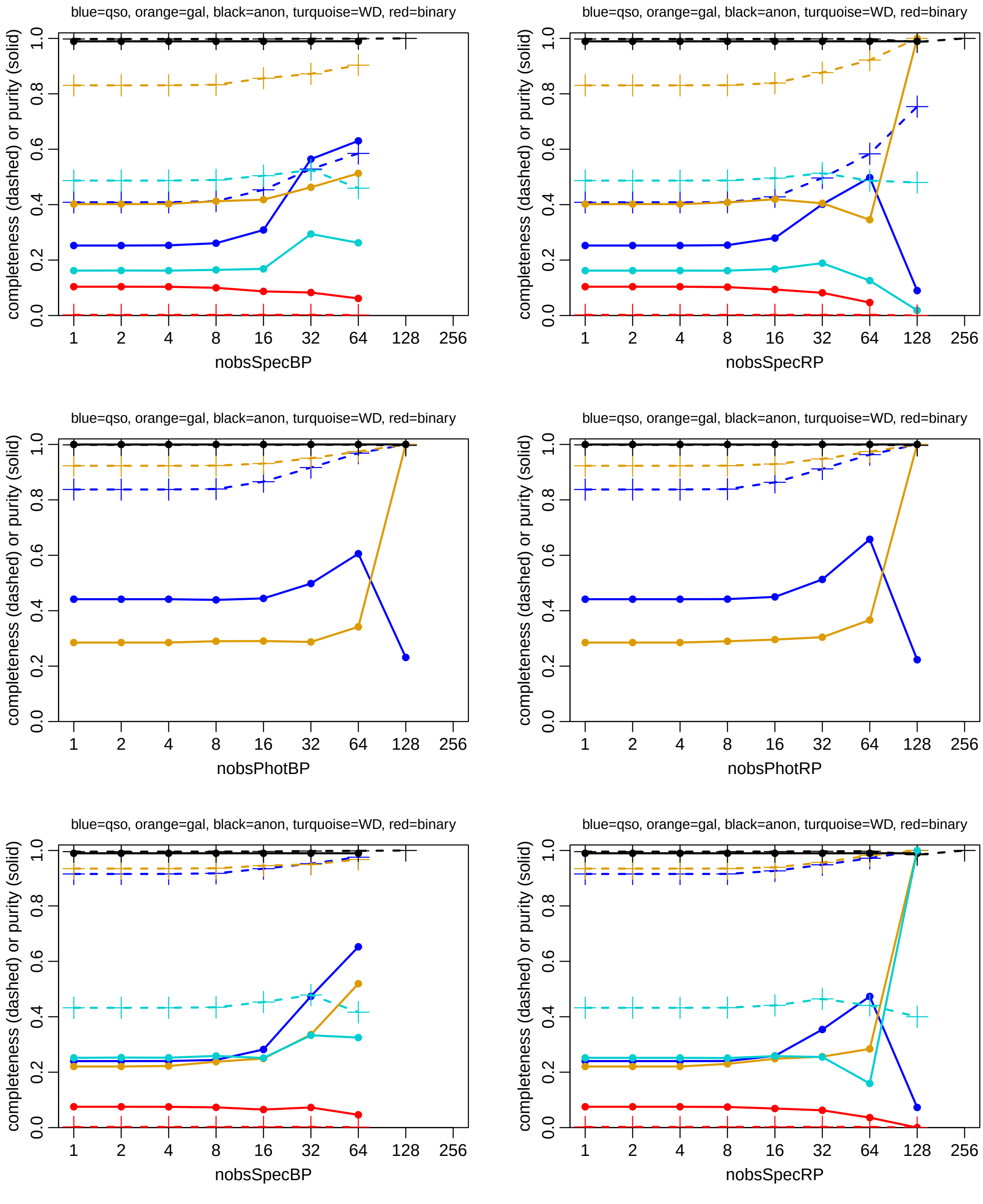


Figure 16: Dataset: defset. The variation of classification completeness (dashed lines) and purity (solid lines) as a function of the minimum number of transits for each classes (different colours) for Specmod (top), Allosmod (centre), Combmod (bottom), for BP (left) and RP (right). The number of transits means the number of focal plane crossings that were used to form the mean data that is the input to DSC. For Specmod and Combmod the horizontal axis is the number of spectroscopic transits, whereas for Allosmod it is the number of photometric transits. Note the log scale. Classes are assigned by maximum probability above a threshold of 0.5.

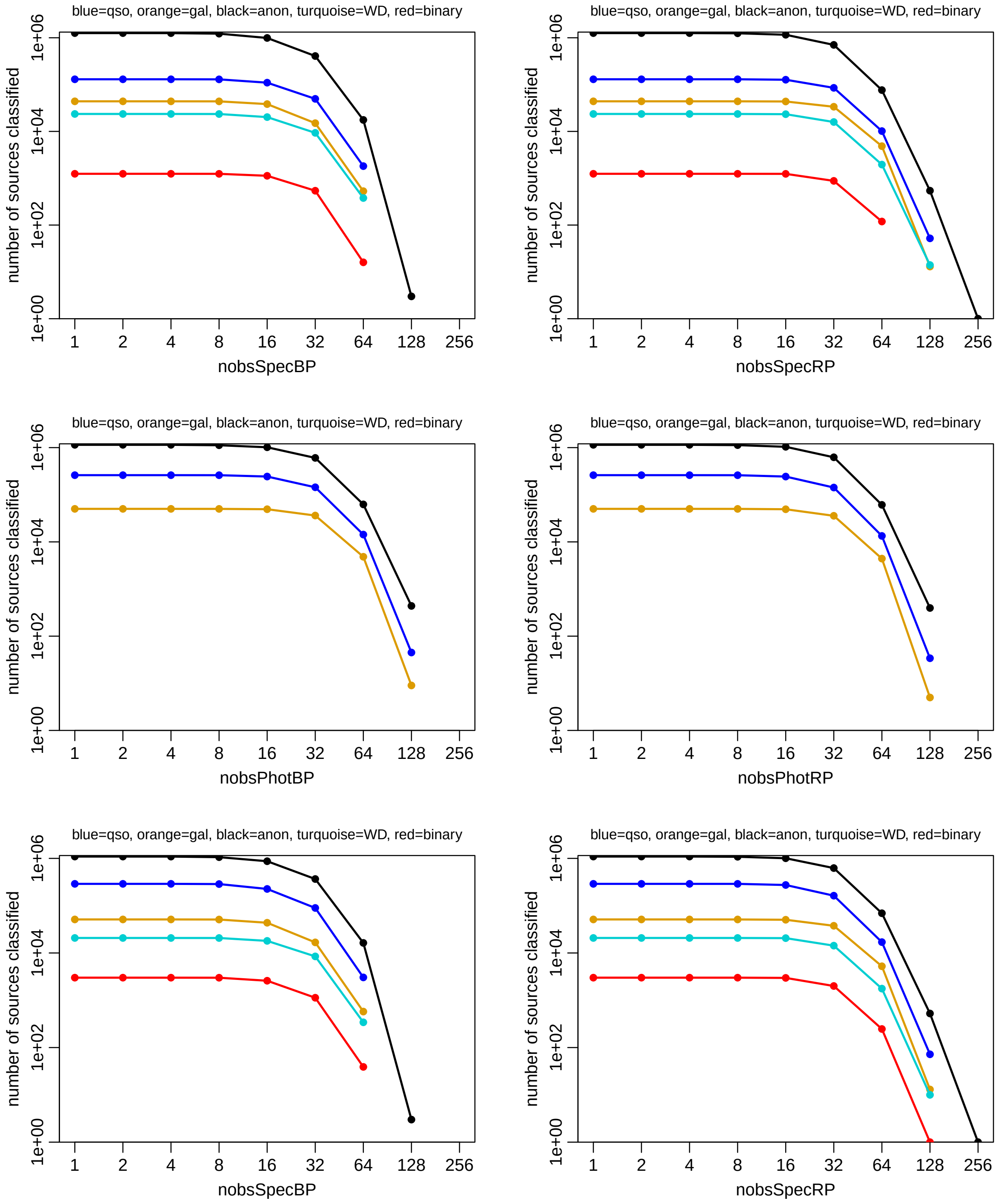


Figure 17: As Figure 16, but now showing the number of sources in the defset with the minimum number of transits.

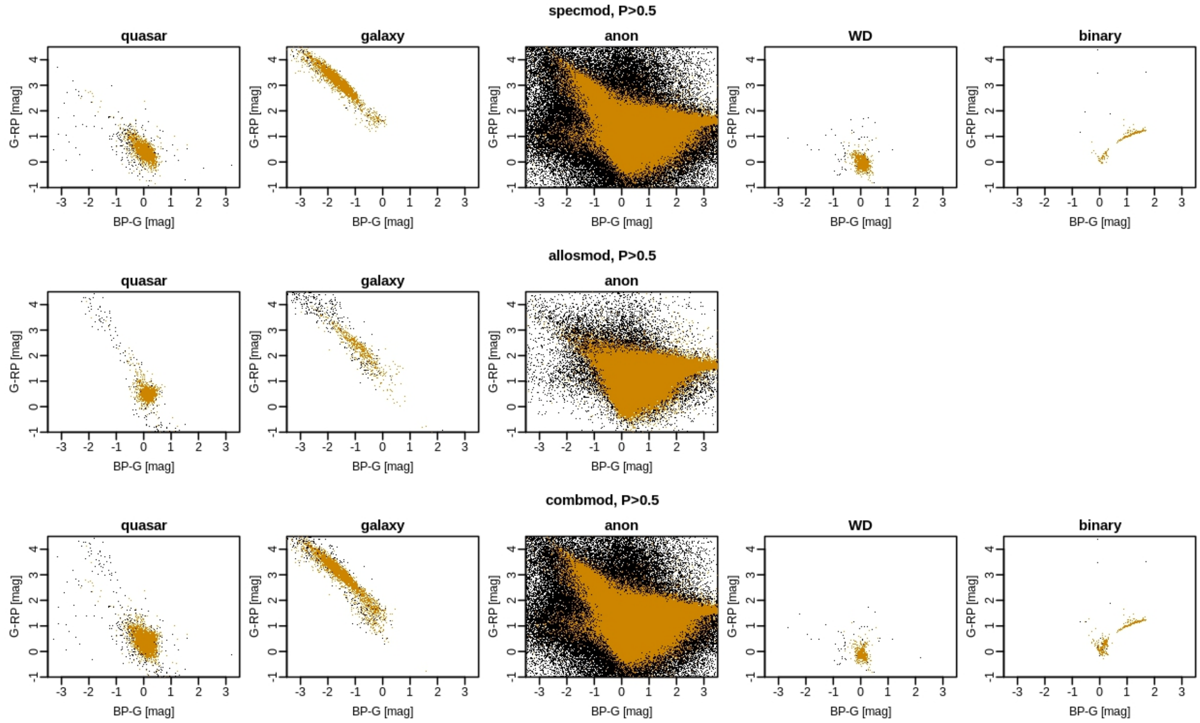


Figure 18: Random subset of full data. CCD for each class (one per panel) in the defset for Specmod (top), Allosmod (centre), and Combmod (bottom). Classes are assigned by the largest posterior probability above 0.5. Black shows the full set, orange the subset of that with more than five BP and RP spectral transits.

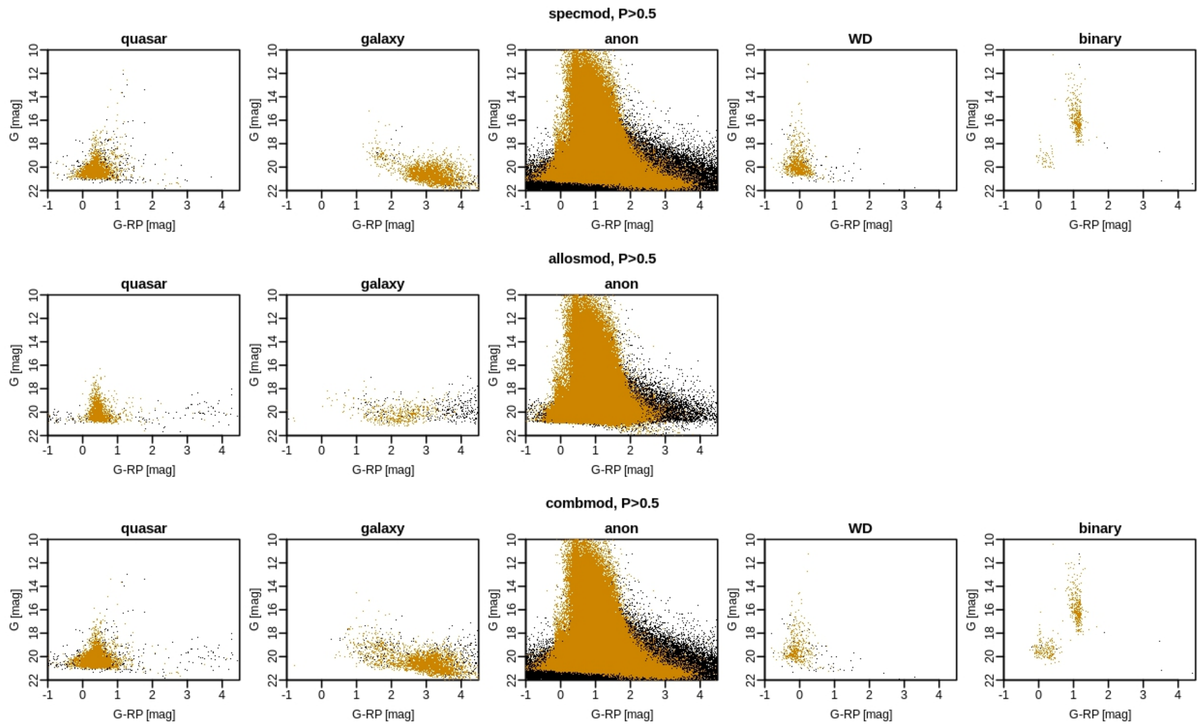


Figure 19: As Figure 18, but showing the CMD.

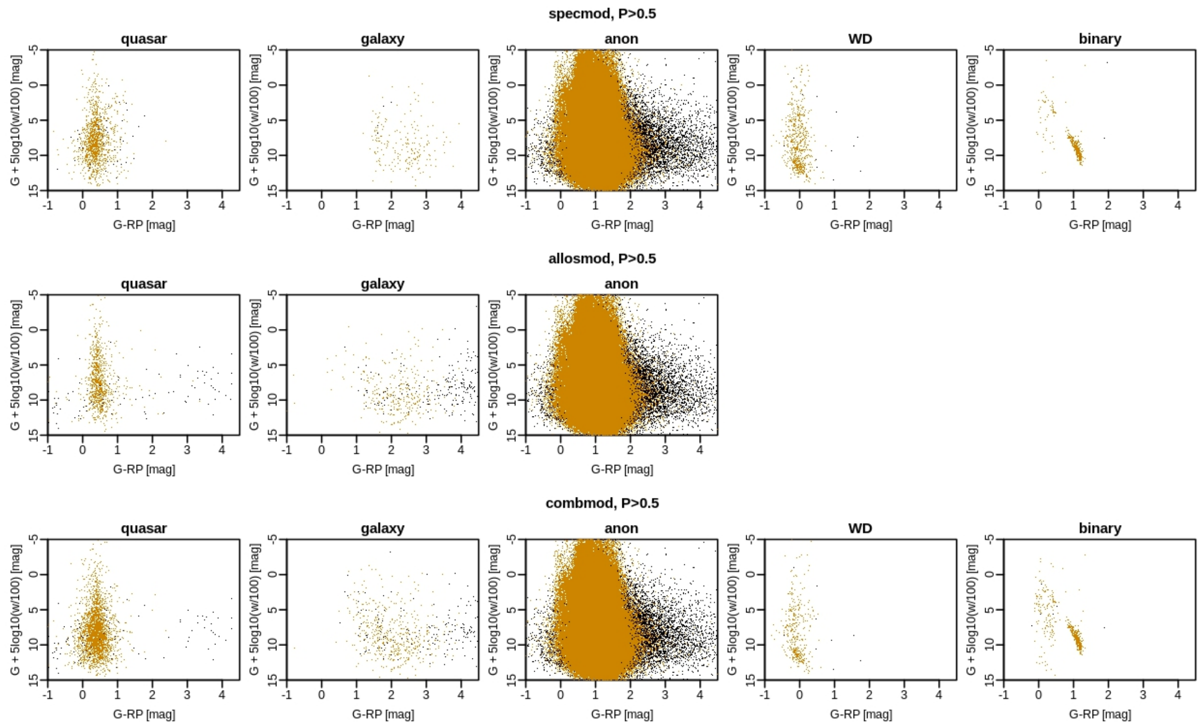


Figure 20: As Figure 18, but showing the CQD. The vertical axis would be equal to the absolute magnitude of a single star in the case of no interstellar extinction and no noise. This is clearly not meaningful for quasars, galaxies, or binary stars, but such a plot is potentially useful for sources classified as those classes to see whether there is an obvious stellar contaminant.

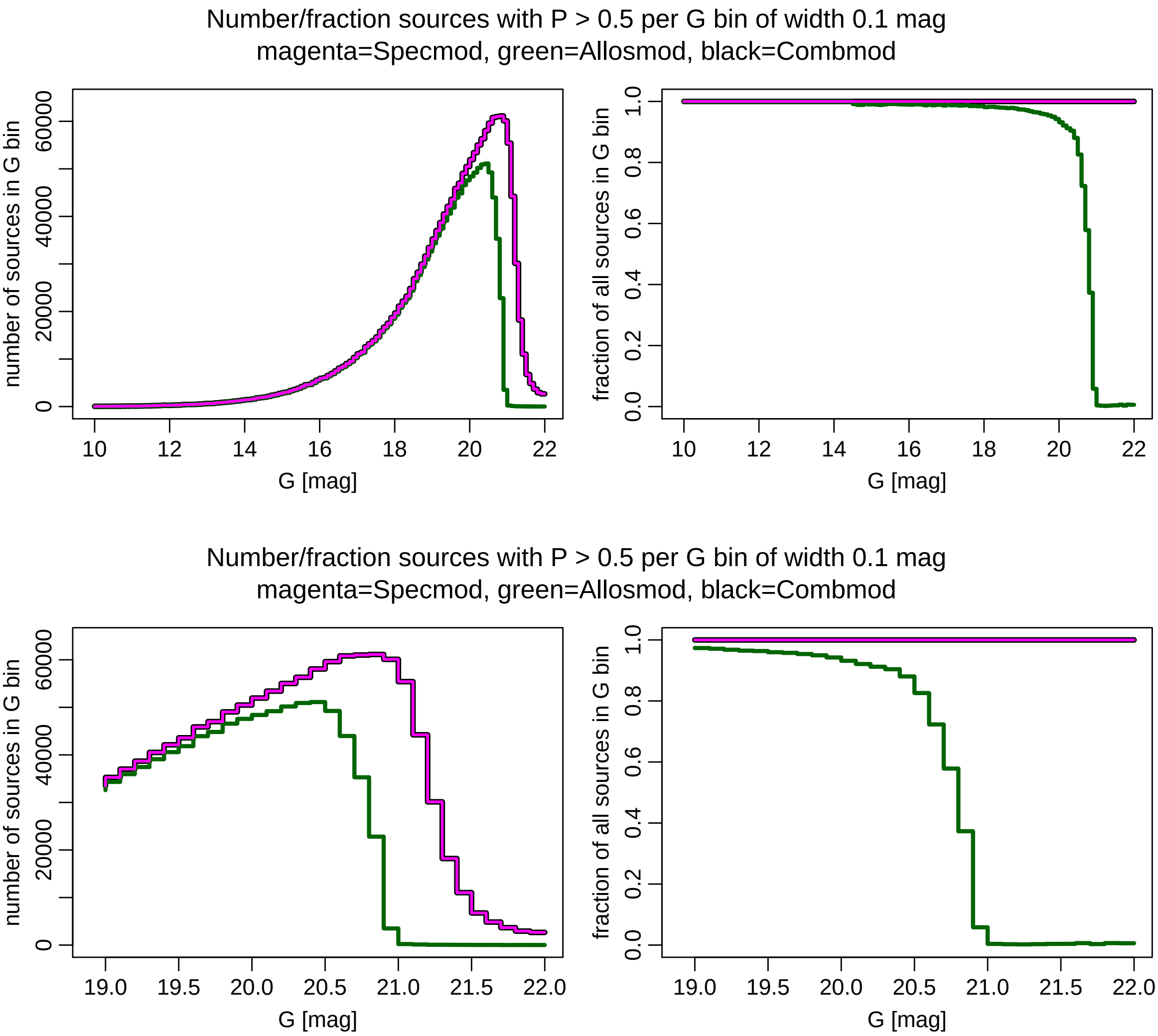


Figure 21: Random subset of full data. Top: The number (left) and fraction (right) of sources with a probability above 0.5 (for any class) in narrow magnitude bins for the three classifiers. The actual numbers for the full data set will be around 1000 times larger than those shown on the left. Bottom: A zoom of the top row for the fainter magnitudes.

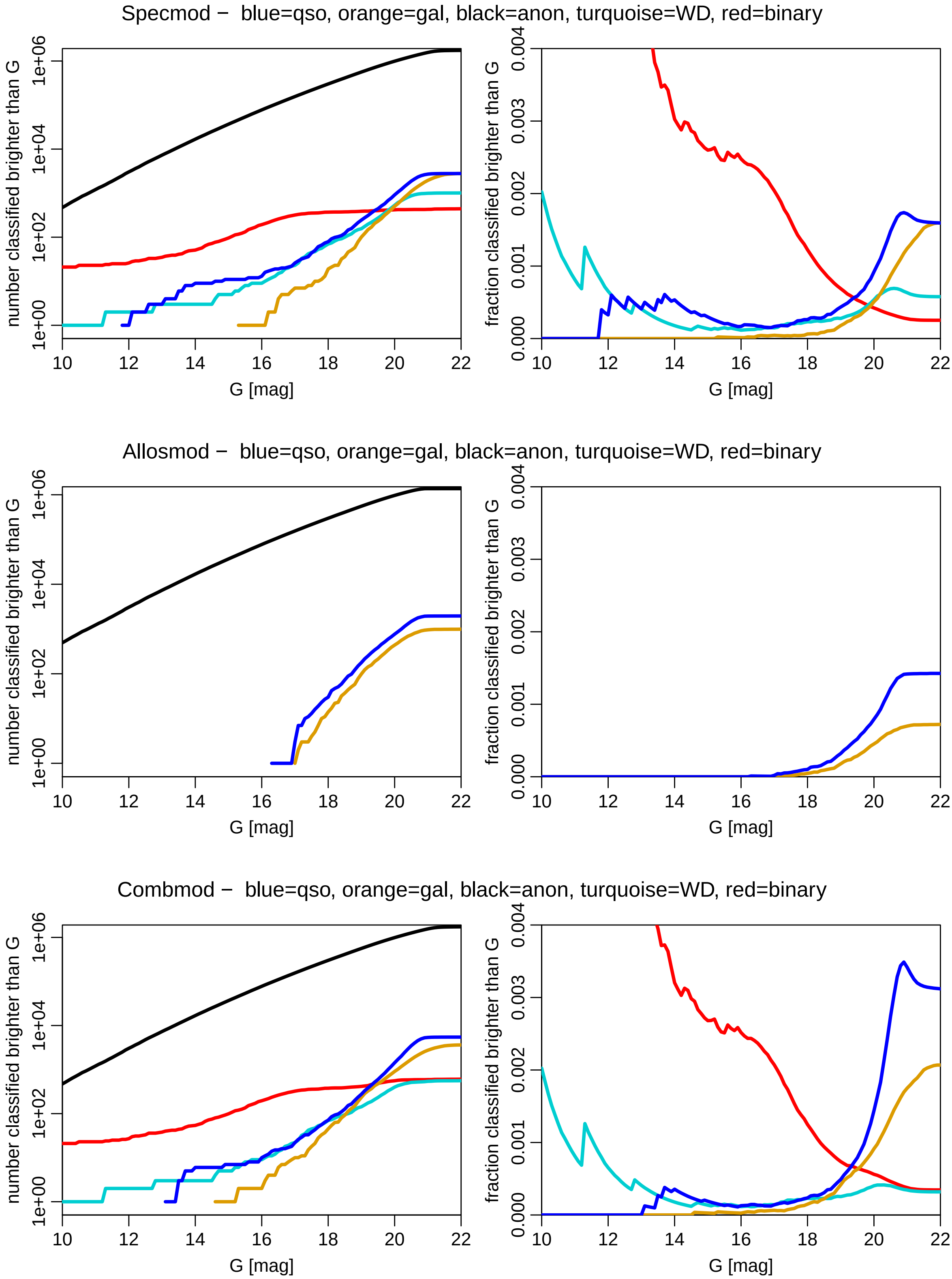


Figure 22: Random subset of full data. The number (left) and fraction (right) of sources with a probability above 0.5 for each class (colours) that are brighter than a given magnitude, for Specmod (top), Allosmod (middle), and Combmod (bottom). At any magnitude the fractions in each right panel sum to one; in all cases the black line (stars) is off the top of the scale.

P_Combmod>0.5

Selection	Number	Fraction
1 Quasars :	5 454 722	
2 Quasars with colours:	5 393 659	0.9888
3 Quasars with astrometry:	4 968 877	0.9109
4 Quasars with both:	4 968 875	0.9109
1 Galaxies :	3 663 012	
2 Galaxies with colours:	3 662 509	0.9999
3 Galaxies with astrometry:	1 214 296	0.3315
4 Galaxies with both:	1 214 296	0.3315

		astrometry ✓	astrometry ✗	
colour	✓	4		2
	✗			
		3		1

Most galaxies (classified by Combmod) do not have astrometry and so have no Allosmod results

P_Combmod>0.5 & classlabelDscJoint: all sources of course have colours and astrometry

Figure 23: The number of quasars (`P_Combmod(quasar)>0.5`) and galaxies `P_Combmod(galaxy)>0.5` in the extragalactic sample which have defined (non-missing) features. "Colours" means they have both BP and RP photometry. "Astrometry" means they have parallaxes and proper motions (always all or none are present). The table in the top-right show the contingency table for the four lines 1–4. From the numbers given one can compute any other elements of this table, e.g. the number with colours but not astrometry.

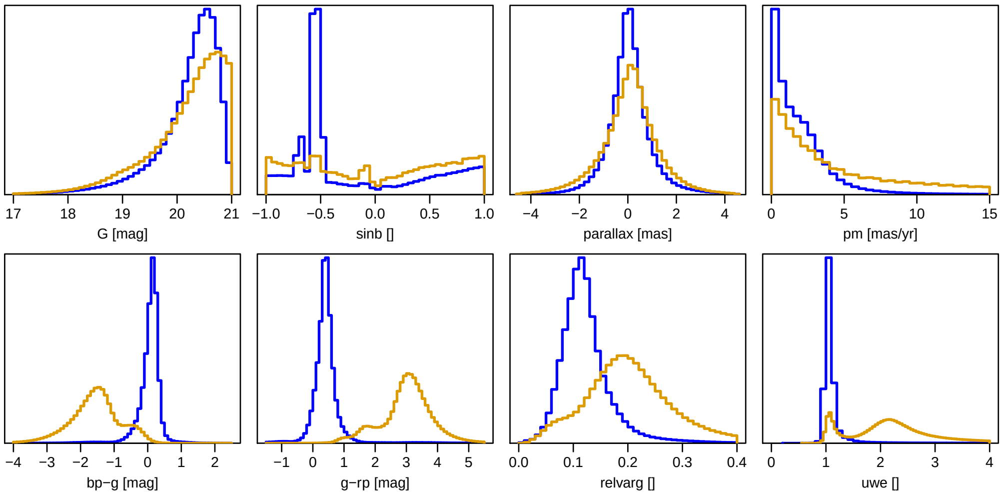


Figure 24: Extragalactic sample. Distribution (linear scale) of various features (those used for Allosmod training) for those objects with `P_Combmod(quasar)>0.5` (blue) and `P_Combmod(galaxy)>0.5` (orange). Units are as in EDR3.

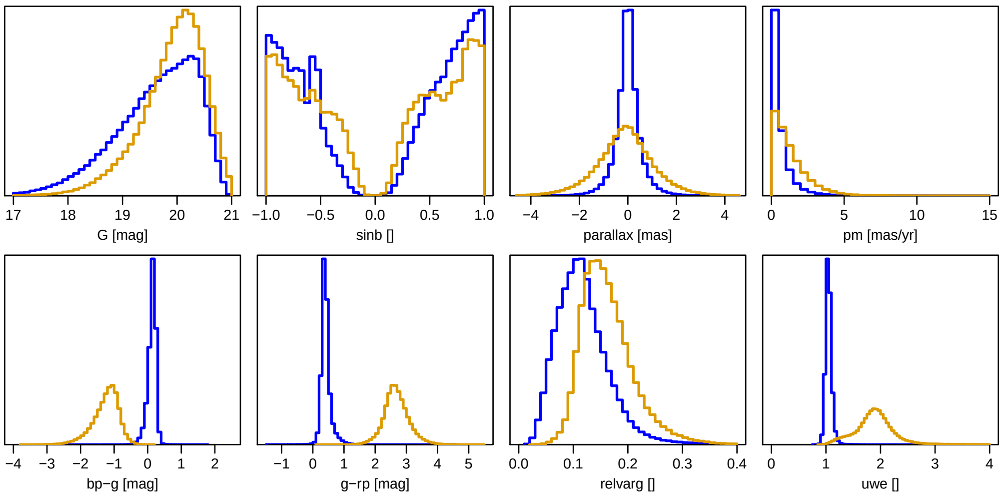


Figure 25: As Figure 24, but for the subset with `classlabelDscJoint` set, i.e. Specmod and Allosmod probabilities are both above 0.5 for that class.

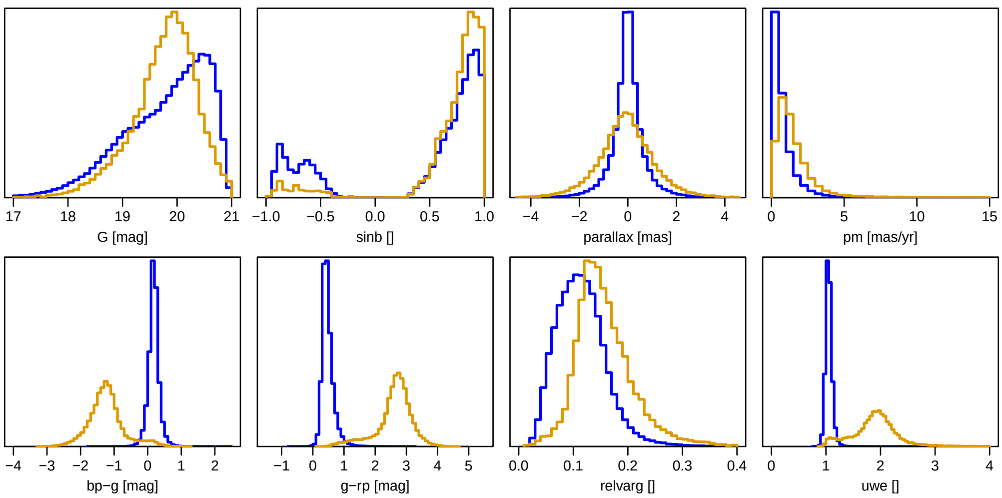


Figure 26: As Figure 24, but for the training set.

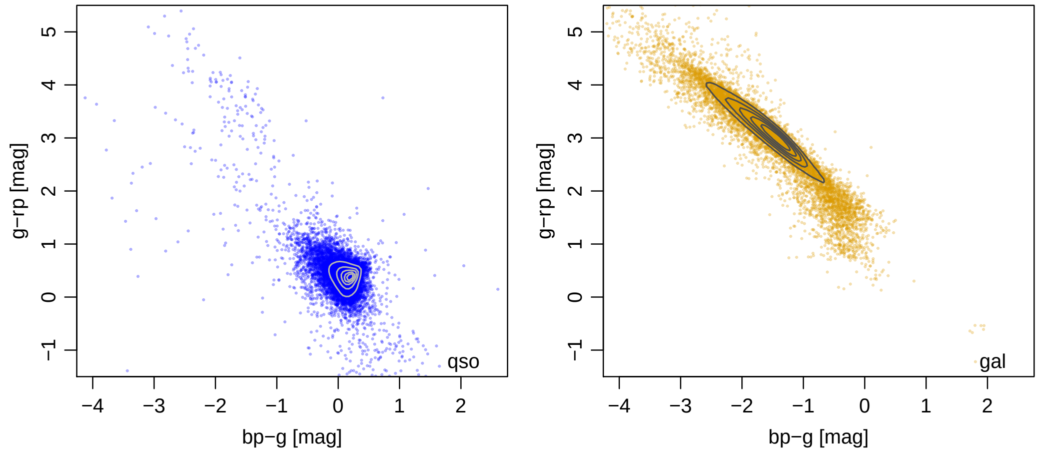


Figure 27: Extragalactic sample. Colour–colour diagram for those objects with `P_Combmod(quasar)>0.5` (blue) and `P_Combmod(galaxy)>0.5` (orange). Units are as in EDR3.

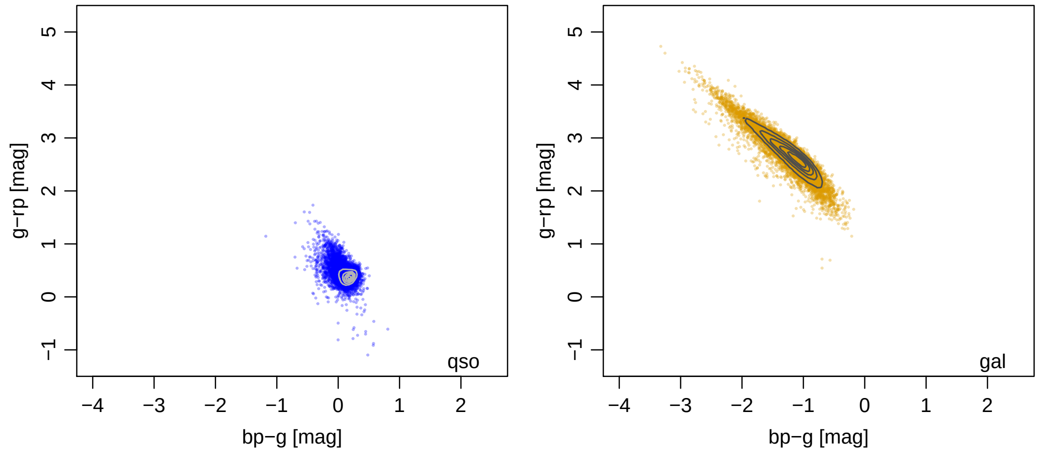


Figure 28: As Figure 27, but for the subset with `classlabelDscJoint` set, i.e. Specmod and Allosmod probabilities are both above 0.5 for that class.

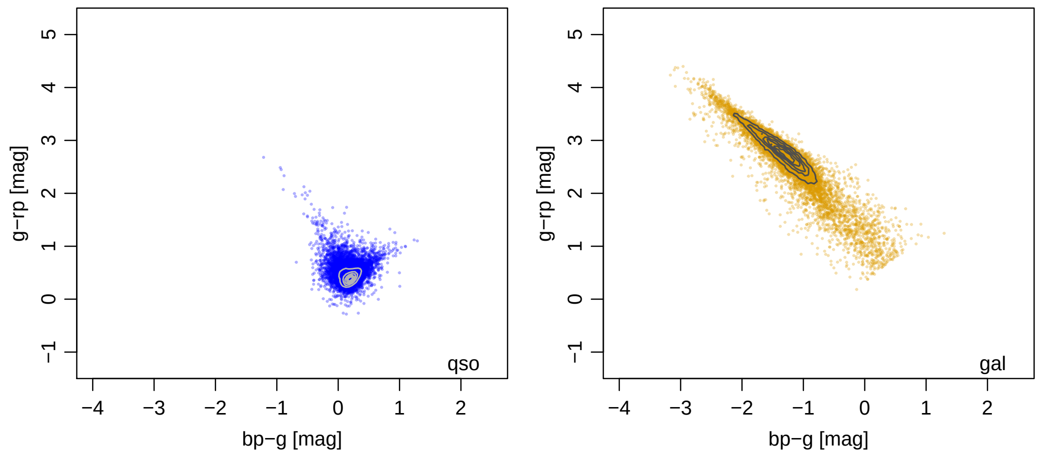


Figure 29: As Figure 27, but for the training set.

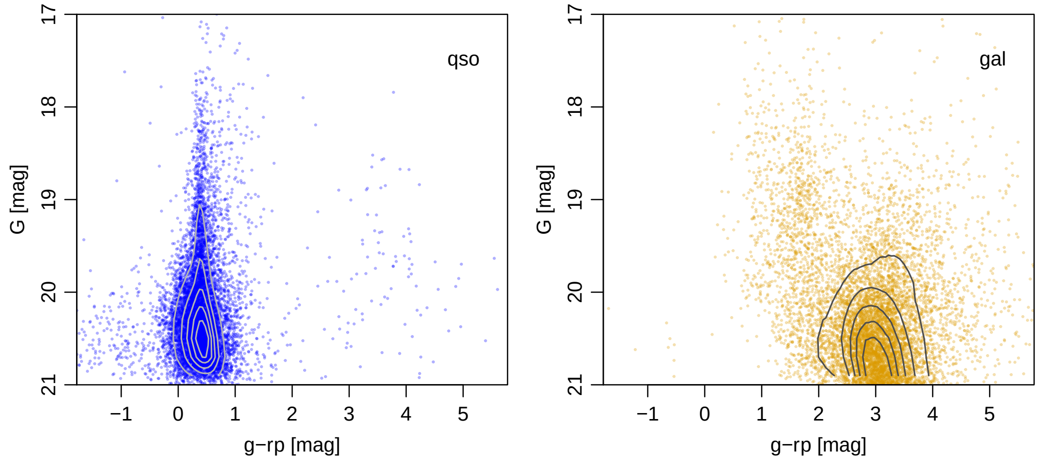


Figure 30: Extragalactic sample. Colour–magnitude diagram for those objects with `P_Combmod(quasar)>0.5` (blue) and `P_Combmod(galaxy)>0.5` (orange). Units are as in EDR3.

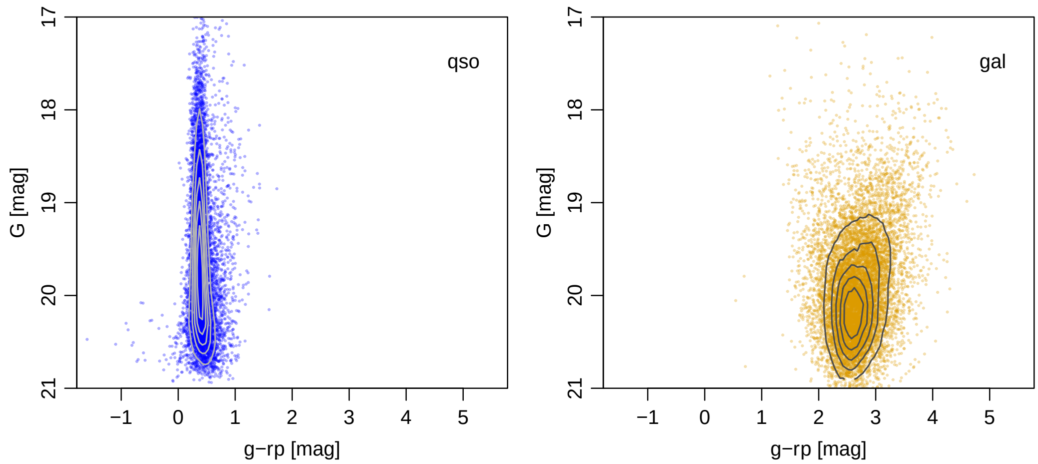


Figure 31: As Figure 30, but for the subset with `classlabelDscJoint` set, i.e. Specmod and Allosmod probabilities are both above 0.5 for that class.

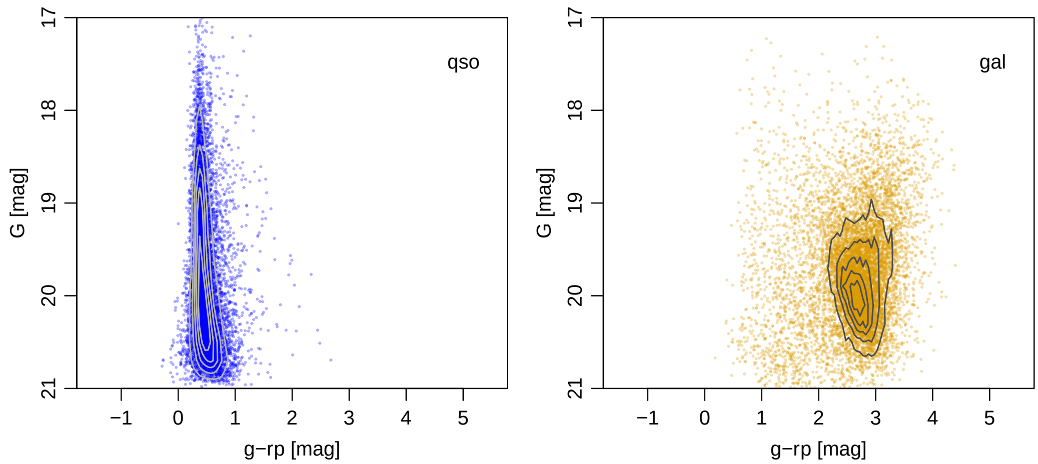


Figure 32: As Figure 30, but for the training set.

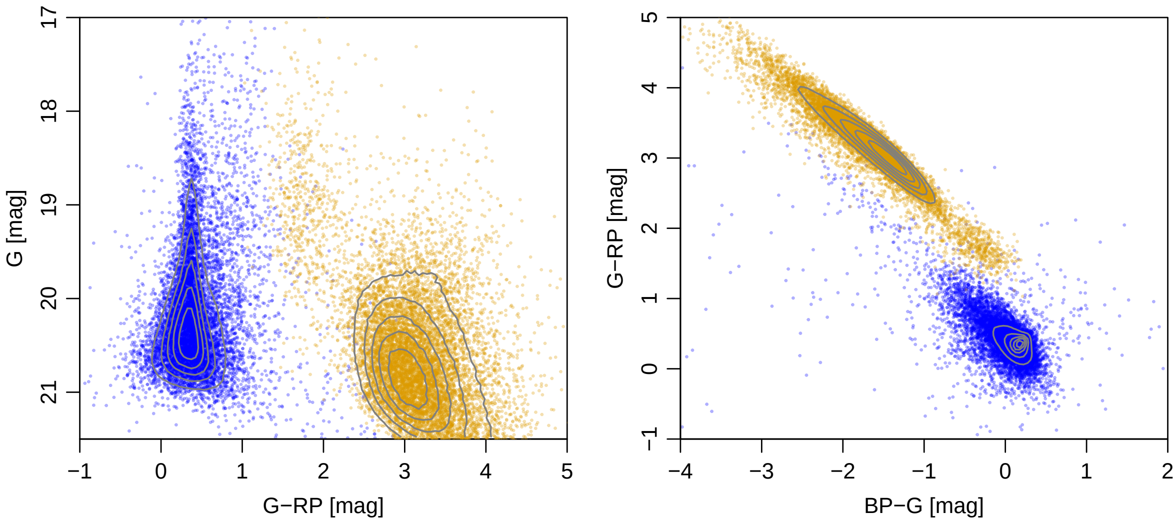


Figure 33: As Figures 27 and 30, but now plotting quasars and galaxies in a single panel to ease comparison.

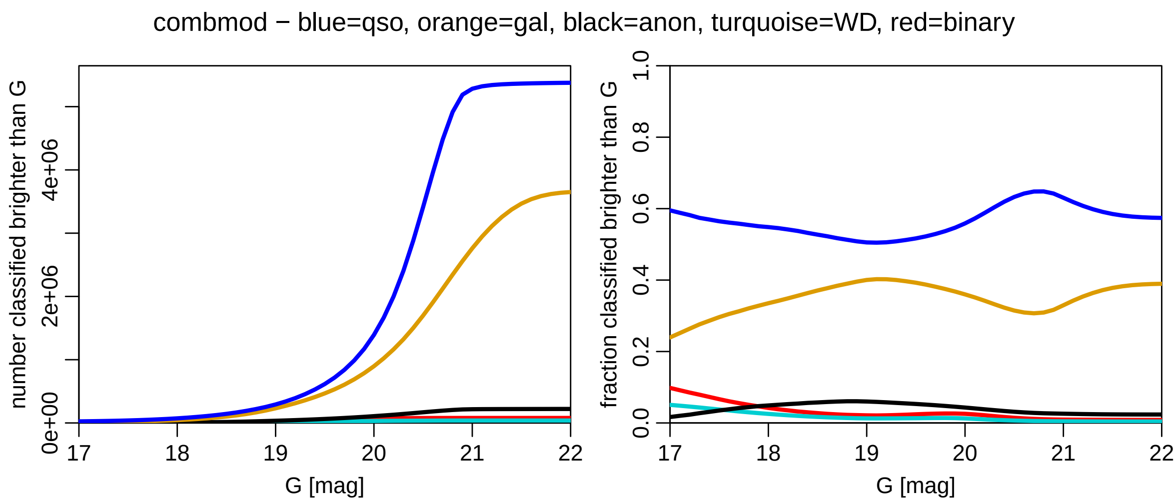


Figure 34: Extragalactic sample. Number (left) and fraction (right) of sources classified according to `P_Combmod(quasar)>0.5` (blue) and `P_Combmod(galaxy)>0.5` (orange) as a function of the limiting G magnitude used to build the sample. At any magnitude the fractions in each right panel sum to one.

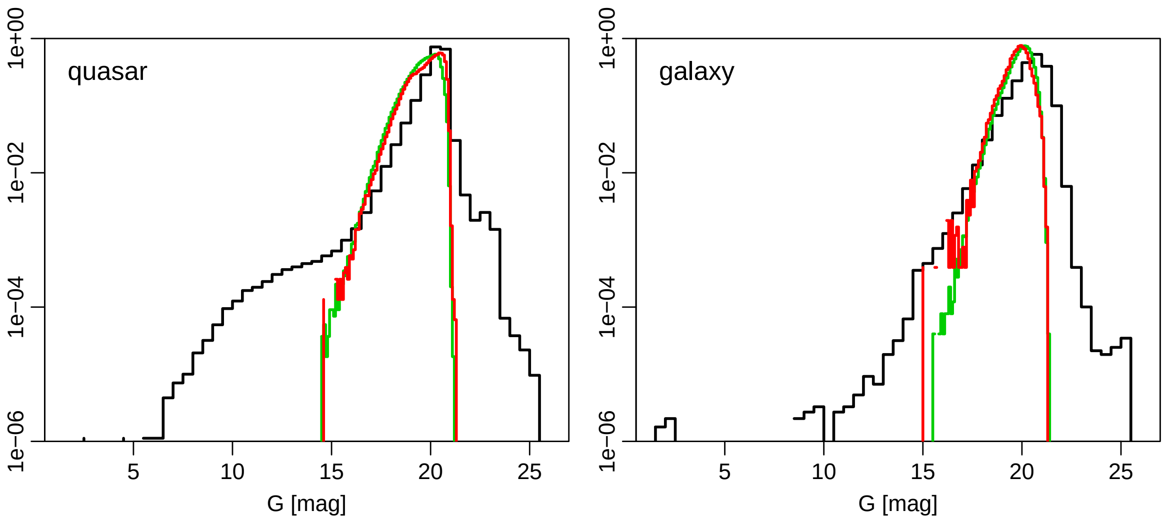


Figure 35: Extragalactic sample. Number of sources per G magnitude bin that are classified as quasars (left) or galaxies (right) in various ways: the full sample (i.e. Combmod probability > 0.5) in black; those that additionally have Specmod and Allosmod probabilities > 0.5 in green; the training data in red. All distributions are normalized to unit area. Note the log scale on the vertical axis. The black lines in the left and right panels are the same data as the blue and orange lines (respectively) in the left panel of Figure 34.

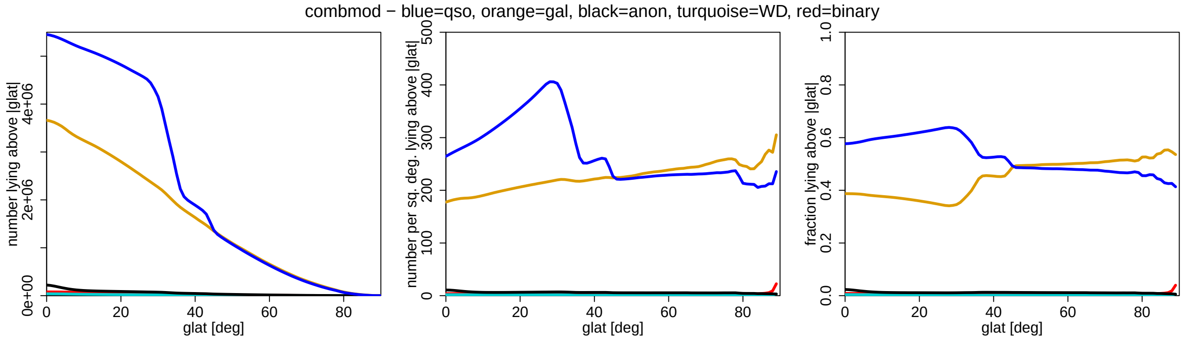


Figure 36: Extragalactic sample. Number (left), number per unit area (middle), and fraction (right) of sources classified according to `P_Combmod(quasar)>0.5` (blue) and `P_Combmod(galaxy)>0.5` (orange) that lie above some absolute Galactic latitude. At any latitude the fractions in each right panel sum to one.

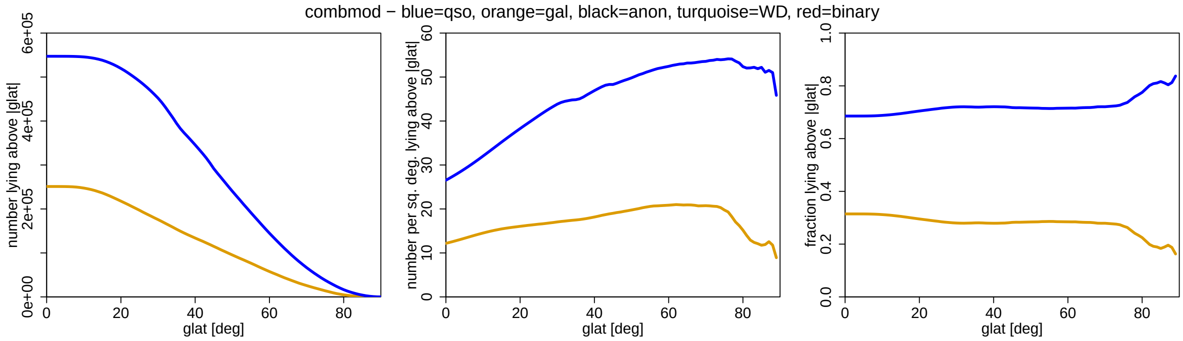


Figure 37: As Figure 36, but for the subset with `classlabelDscJoint` set, i.e. Specmod and Allosmod probabilities are both above 0.5 for that class.

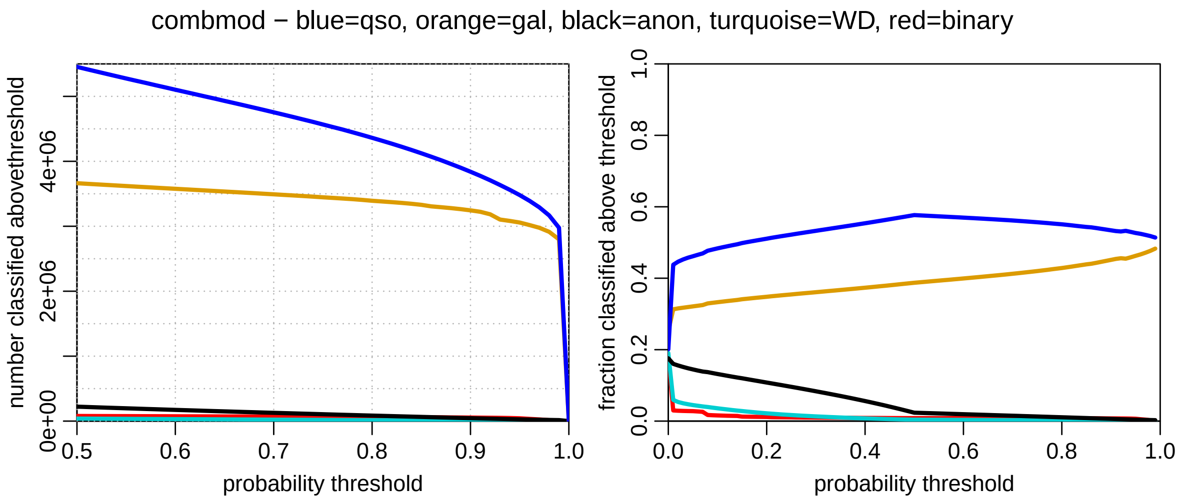


Figure 38: Extragalactic sample. Number (left) and fraction (right) of sources classified according to `P_Combmod(quasar)` $> p$ (blue) and `P_Combmod(galaxy)` $> p$ (orange) as a function of p. At any p the fractions in each right panel sum to one.

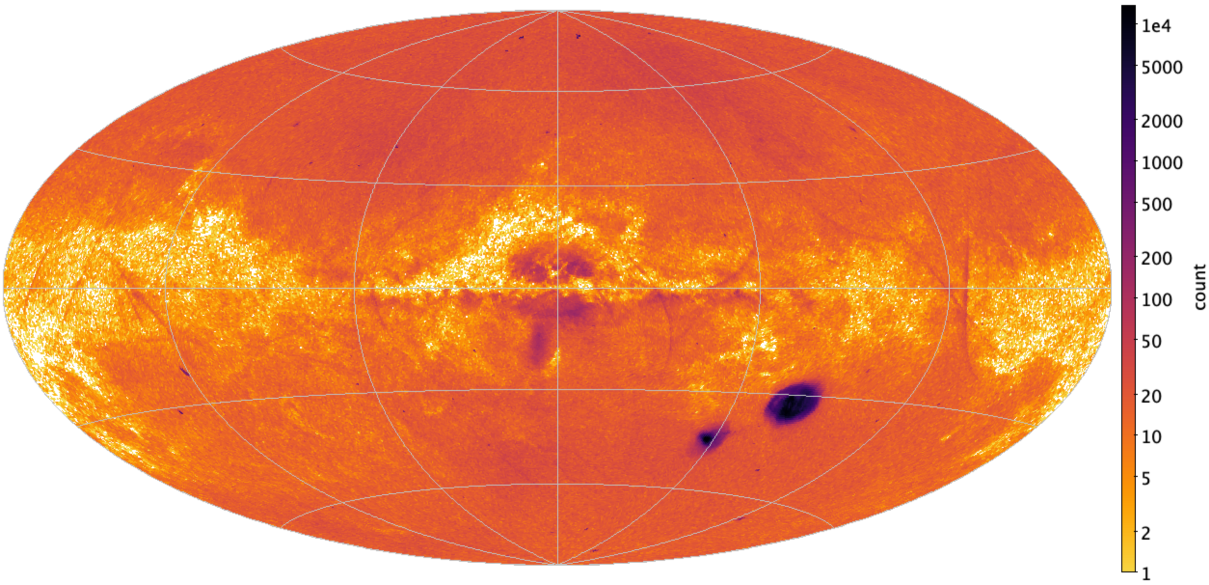


Figure 39: Extragalactic sample. Galactic sky distribution for sources with `P_Combmod(quasar)` $>$ 0.5. Each plot shows the number of objects per healpix at level 7 (0.21 sq. deg.) on a logarithmic scale in an Aitoff (equal-area) projection. The Galactic centre is in the middle and longitude increases to the left.

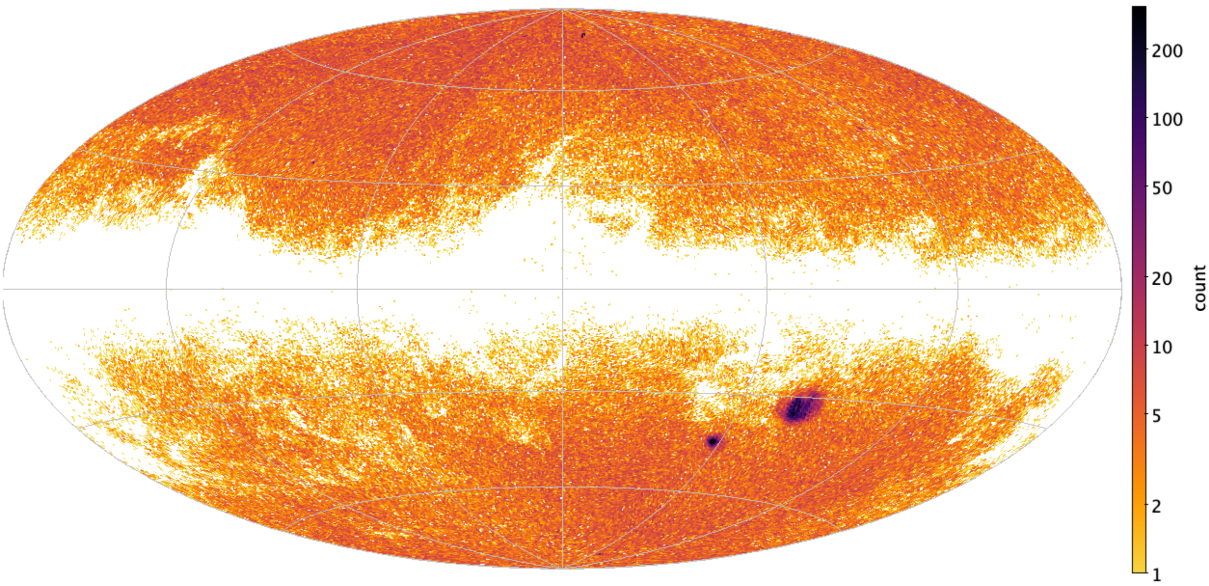


Figure 40: As Figure 39, but for the subset with `classlabelDscJoint = quasar`, i.e. Specmod and Allosmod probabilities are both above 0.5.

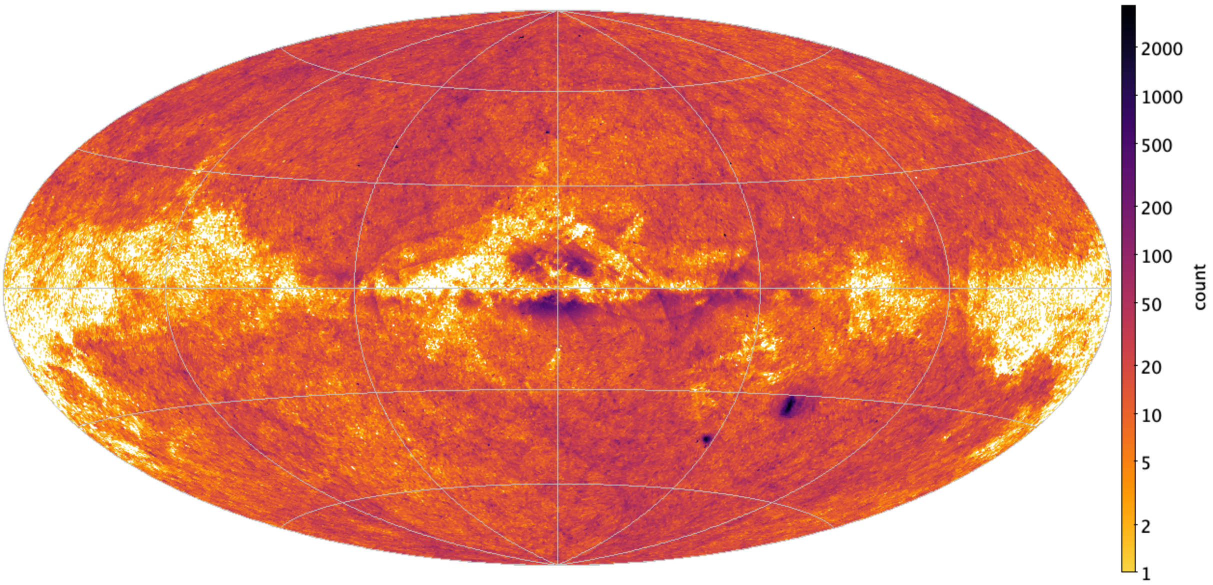


Figure 41: Extragalactic sample. Galactic sky distribution for sources with `P_Combmod(galaxy)>` 0.5. Each plot shows the number of objects per healpix at level 7 (0.21 sq. deg.) on a logarithmic scale in an Aitoff (equal-area) projection. The Galactic centre is in the middle and longitude increases to the left.

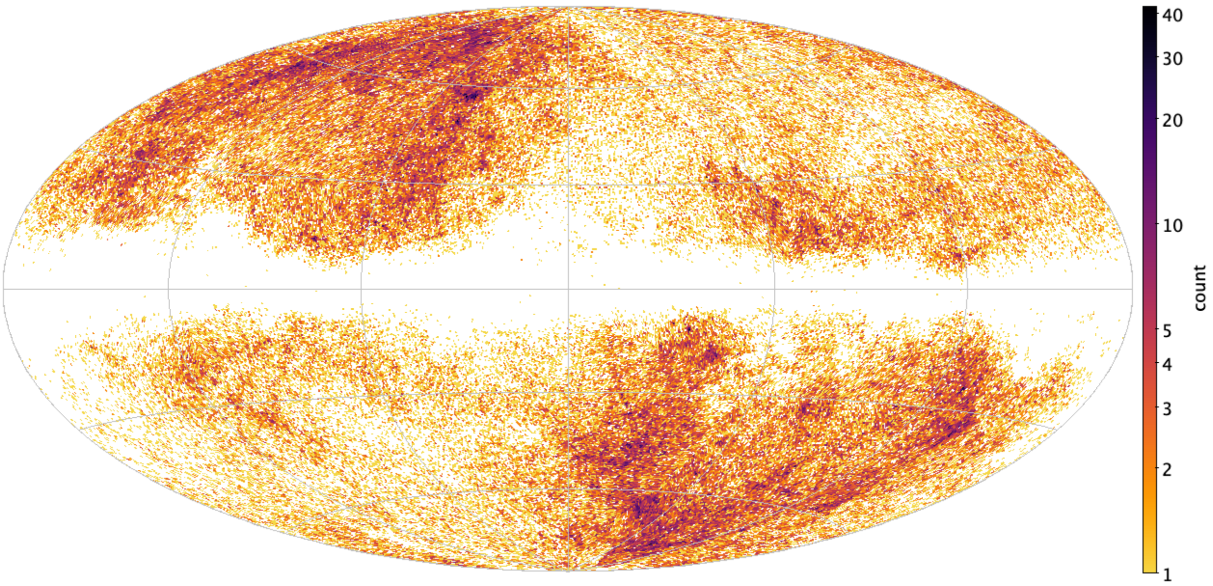


Figure 42: As Figure 41, but for the subset with `classlabelDscJoint = galaxy`, i.e. Specmod and Allosmod probabilities are both above 0.5.

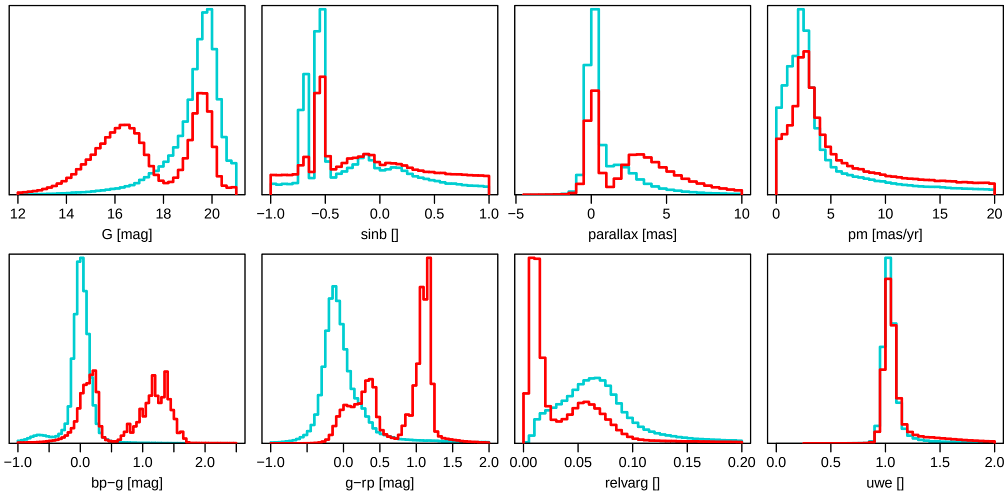


Figure 43: Distribution (linear scale) of various features (those used for Allosmod training) for those objects with `P_Combmod(whitedwarf)>0.5` (cyan) and `P_Combmod(binary)>0.5` (red), Units are as in EDR3.

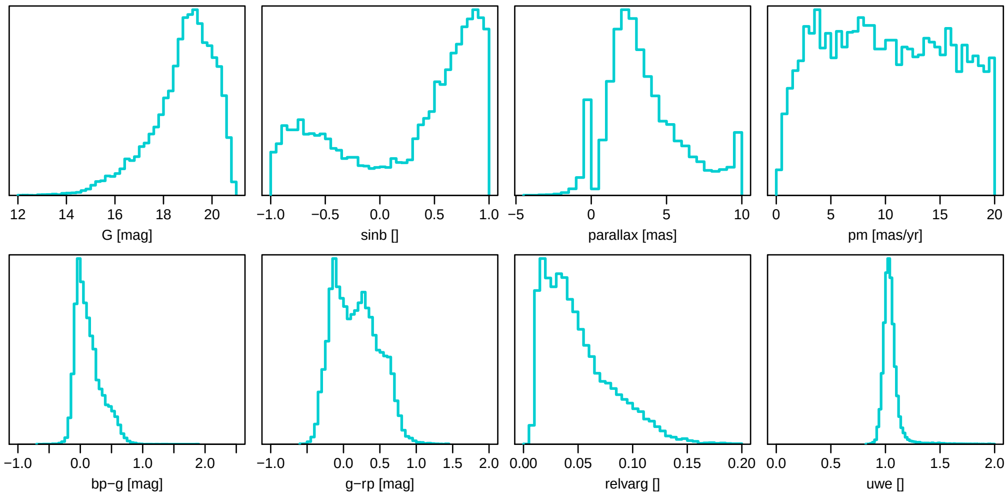


Figure 44: As Figure 43, but for the training set. Binaries are omitted because there are no Gaia features for the training set.

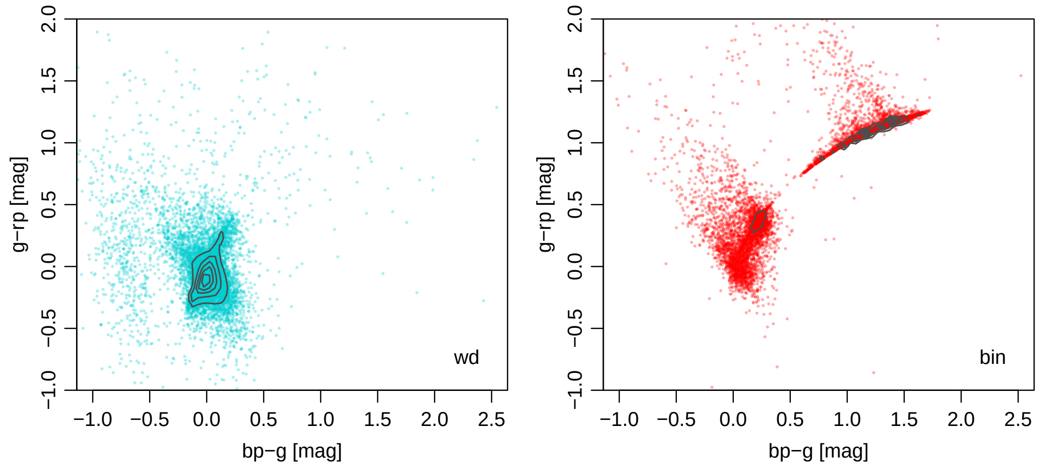


Figure 45: Colour–colour diagram for those objects with `P_Combmod(whitedwarf)>0.5` (cyan) and `P_Combmod(binary)>0.5` (red). Units are as in EDR3.

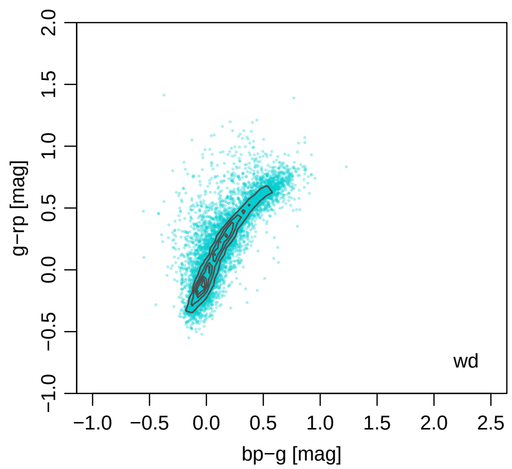


Figure 46: As Figure 45, but for the training data. Binaries are omitted because there are no Gaia features for the training set.

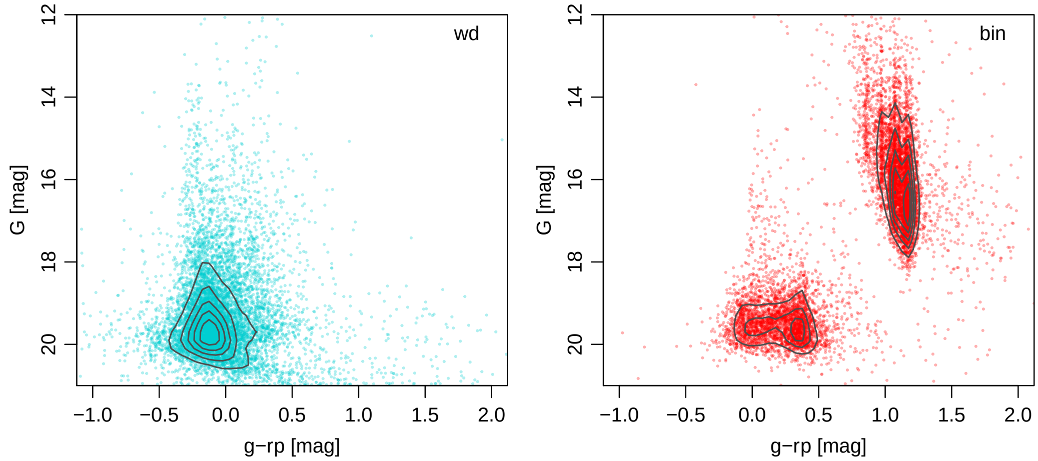


Figure 47: Colour–magnitude diagram for those objects with `P_Combmod(whitedwarf)>0.5` (cyan) and `P_Combmod(binary)>0.5` (red). Units are as in EDR3.

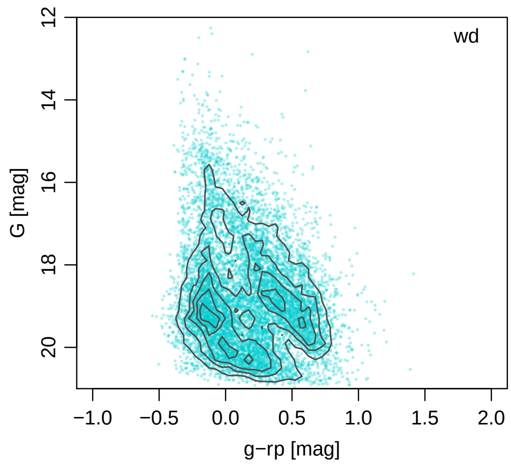


Figure 48: As Figure 47, but for the training data. Binaries are omitted because there are no Gaia features for the training set.

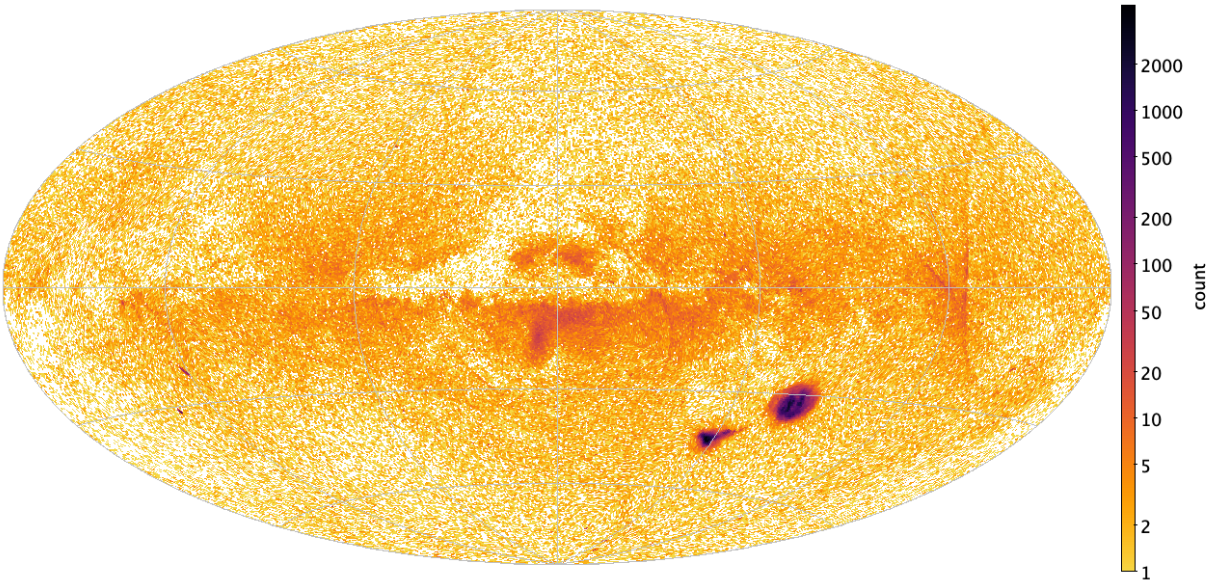


Figure 49: Galactic sky distribution for sources with `P_Combmod(white dwarf)`> 0.5. Each plot shows the number of objects per healpix at level 7 (0.21 sq. deg.) on a logarithmic scale in an Aitoff (equal-area) projection. The Galactic centre is in the middle and longitude increases to the left.

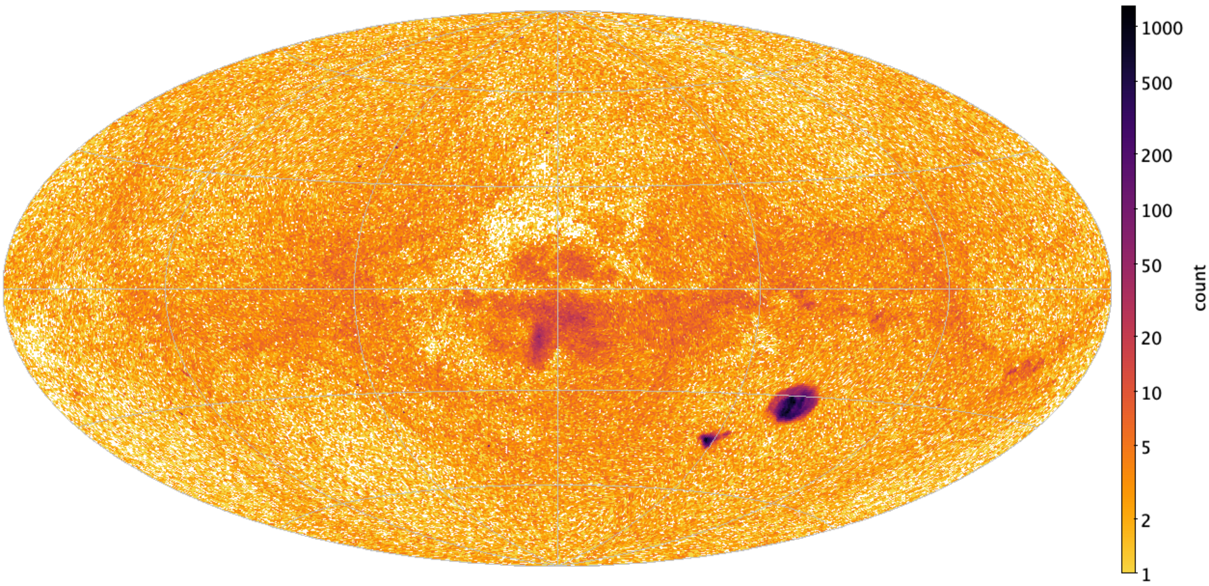


Figure 50: As Figure 41, but for `P_Combmod(binary)`> 0.5

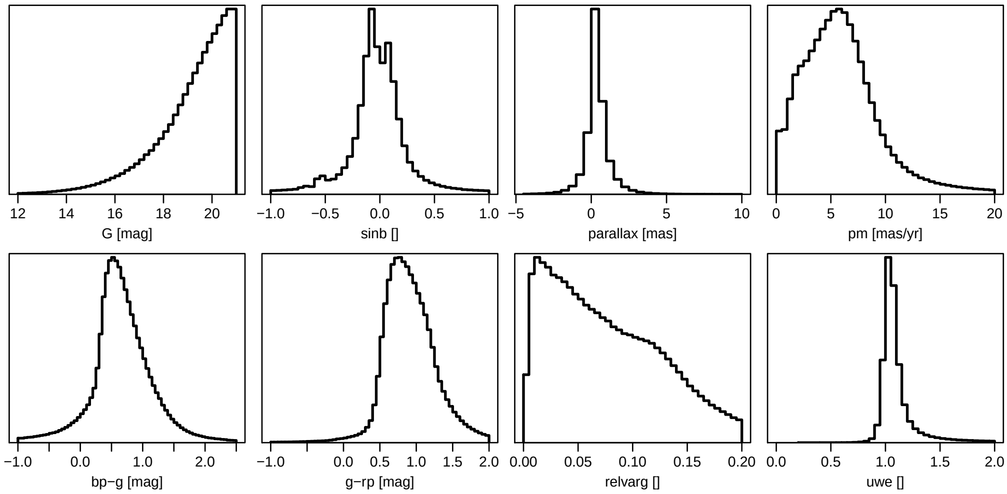


Figure 51: Random subset of full data. Distribution (linear scale) of various features (those used for Allosmod training) for those objects with `P_Combmod(anonymous)>0.5`. Units are as in EDR3.

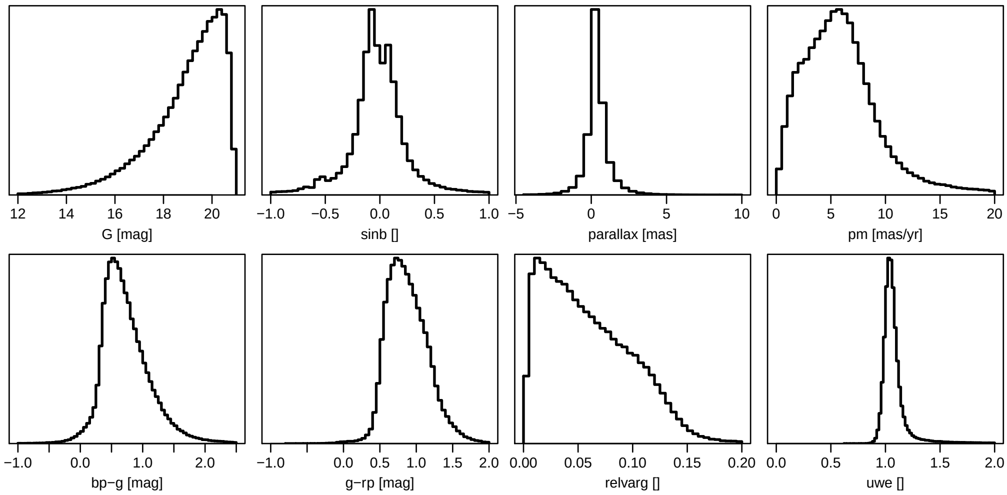


Figure 52: As Figure 51, but for the training set.

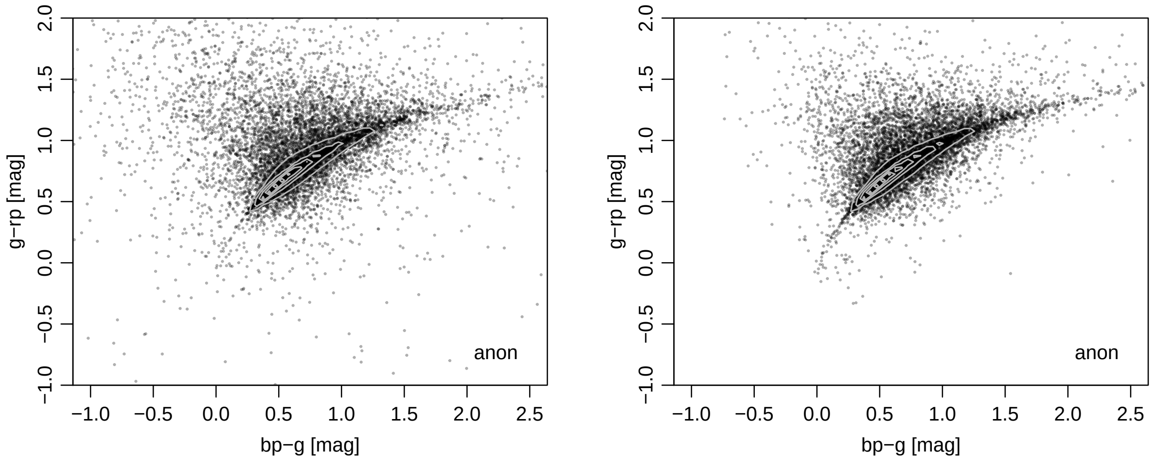


Figure 53: Colour–colour diagram of anonymous sources. Left: Random subset of full data, sources with P_Combmod(anonymous)>0.5. Right: the training set. Units are as in EDR3.

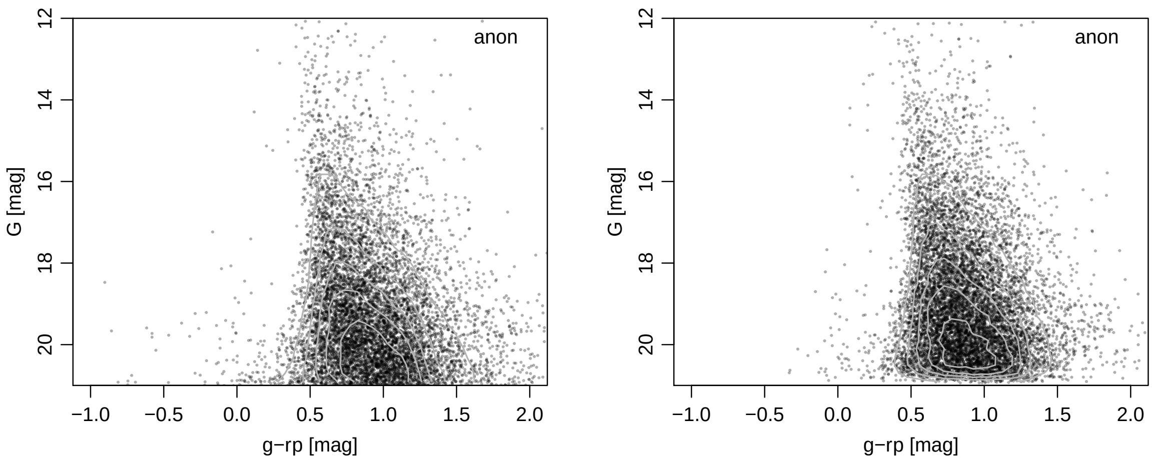


Figure 54: Colour–magnitude diagram of anonymous sources. Left: Random subset of full data, sources with P_Combmod(anonymous)>0.5. Right: the training set. Units are as in EDR3.

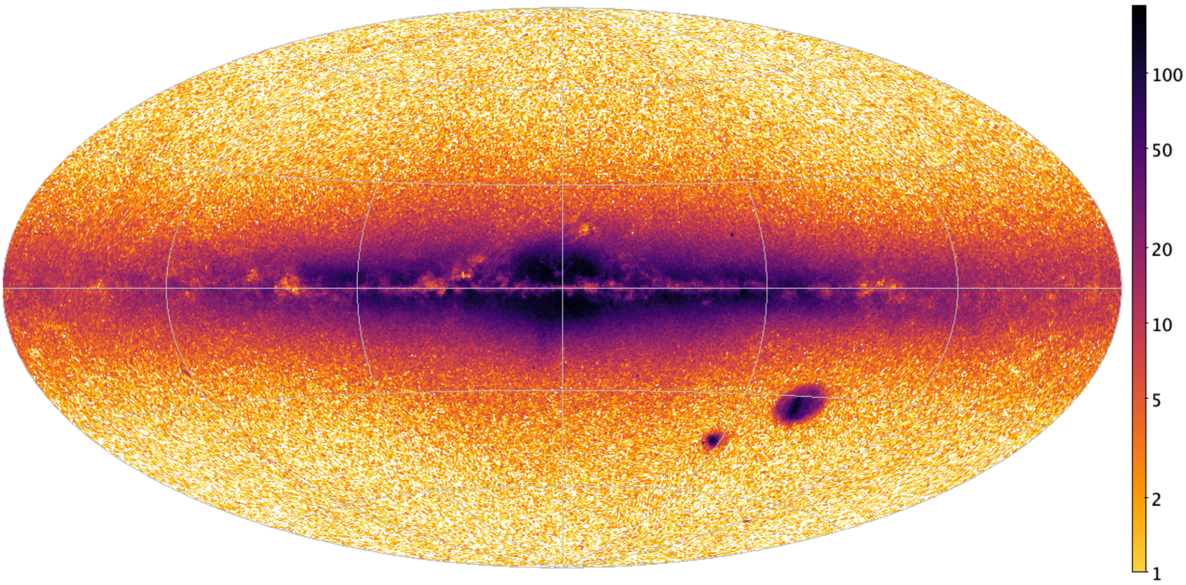


Figure 55: Random subset. Galactic sky distribution for sources with P_Combmod(anonymous) > 0.5. Each plot shows the number of objects per healpix at level 7 (0.21 sq. deg.) on a logarithmic scale in an Aitoff (equal-area) projection. To get the approximate number of objects in the full DSC sample, multiply by 1000. The Galactic centre is in the middle and longitude increases to the left.

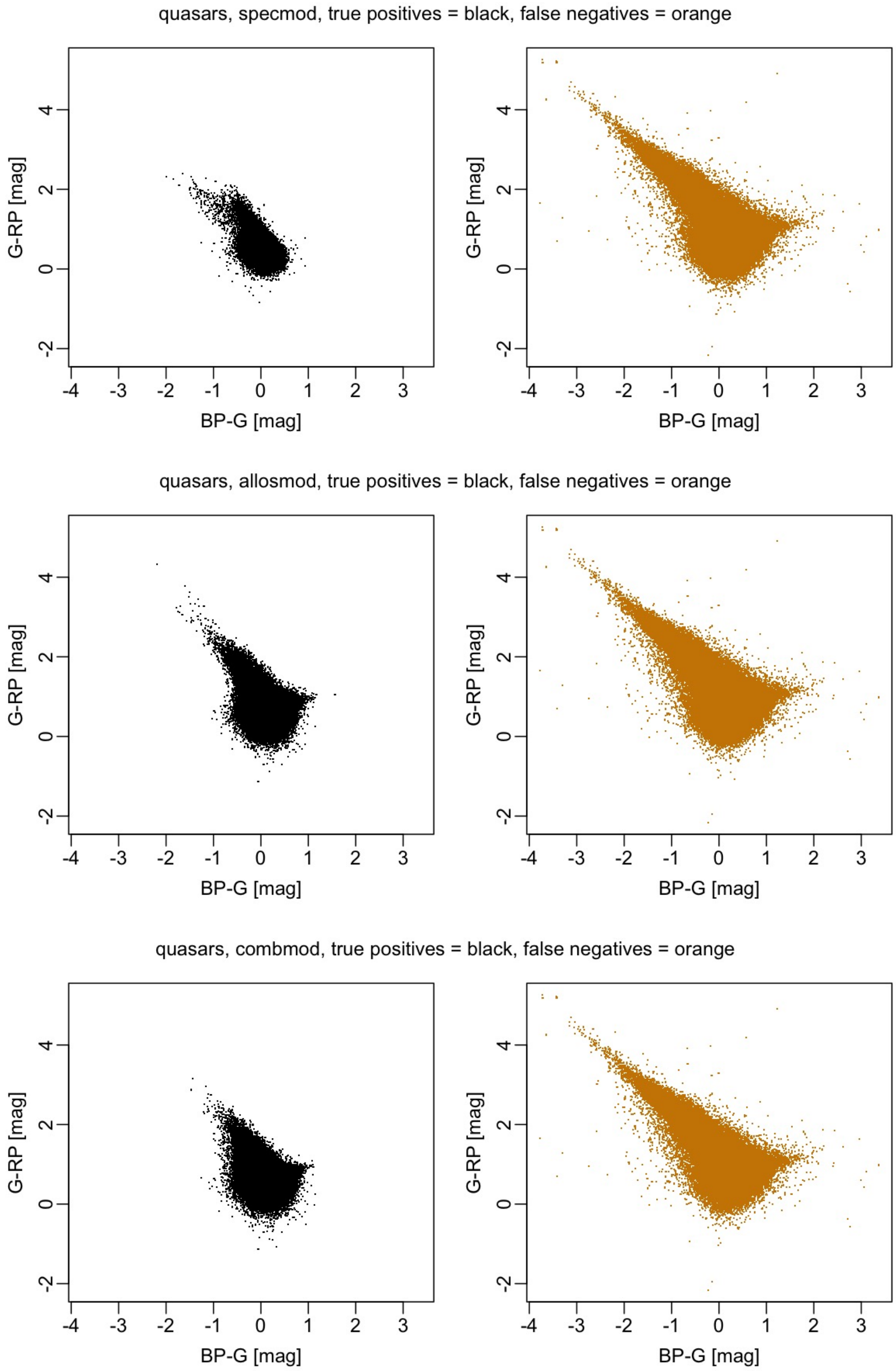


Figure 56: CCD for Klioner AGN for those classified as quasars (left, black) and those classified as something else (right, orange) for Specmod (top), Allosmod (centre), and Combmod (bottom). Classification are assigned by the largest posterior probability.

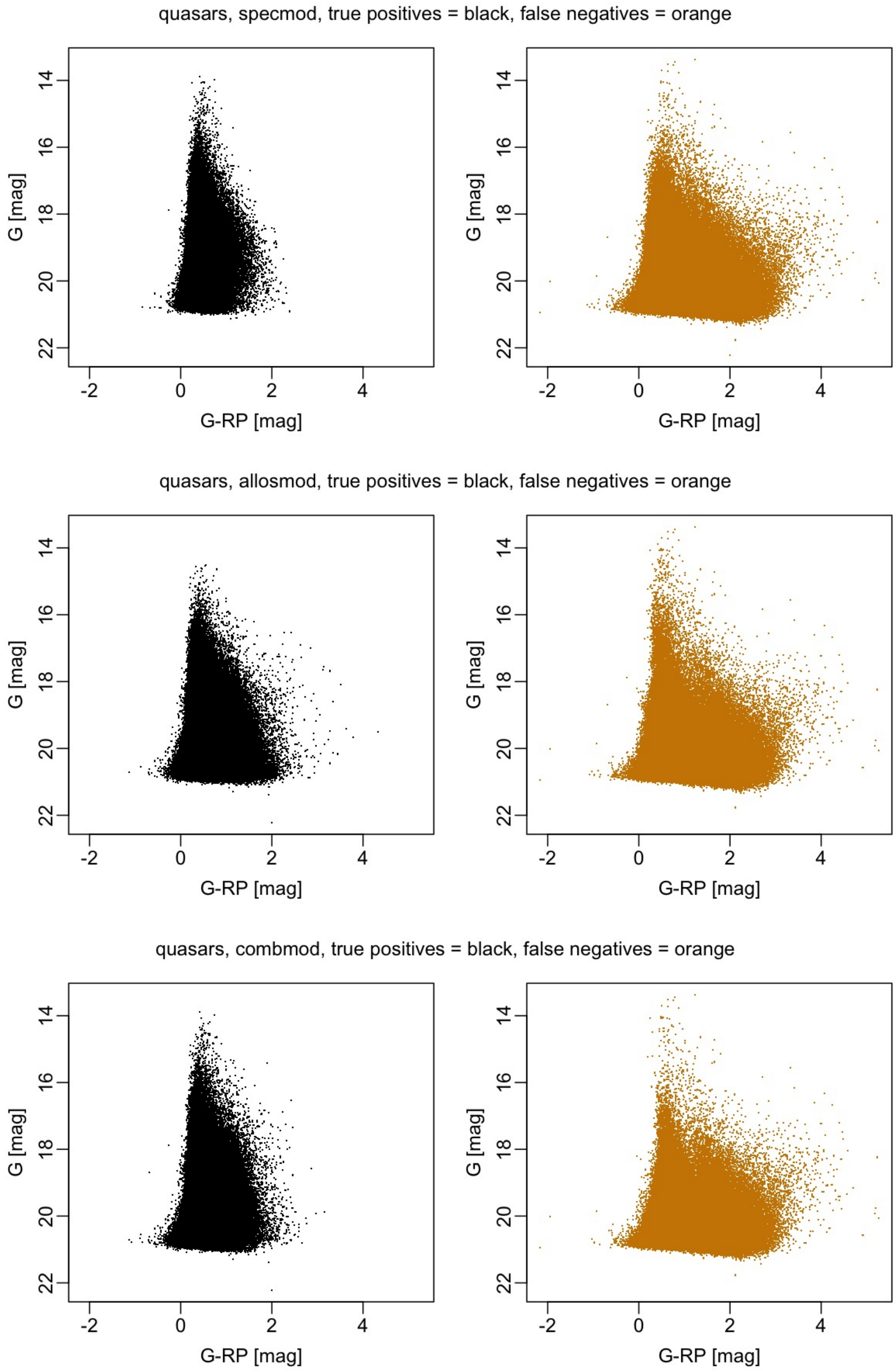


Figure 57: CMD for Klioner AGN for those classified as quasars (left, black) and those classified as something else (right, orange) for Specmod (top), Allosmod (centre), and Combmod (bottom). Classification are assigned by the largest posterior probability.

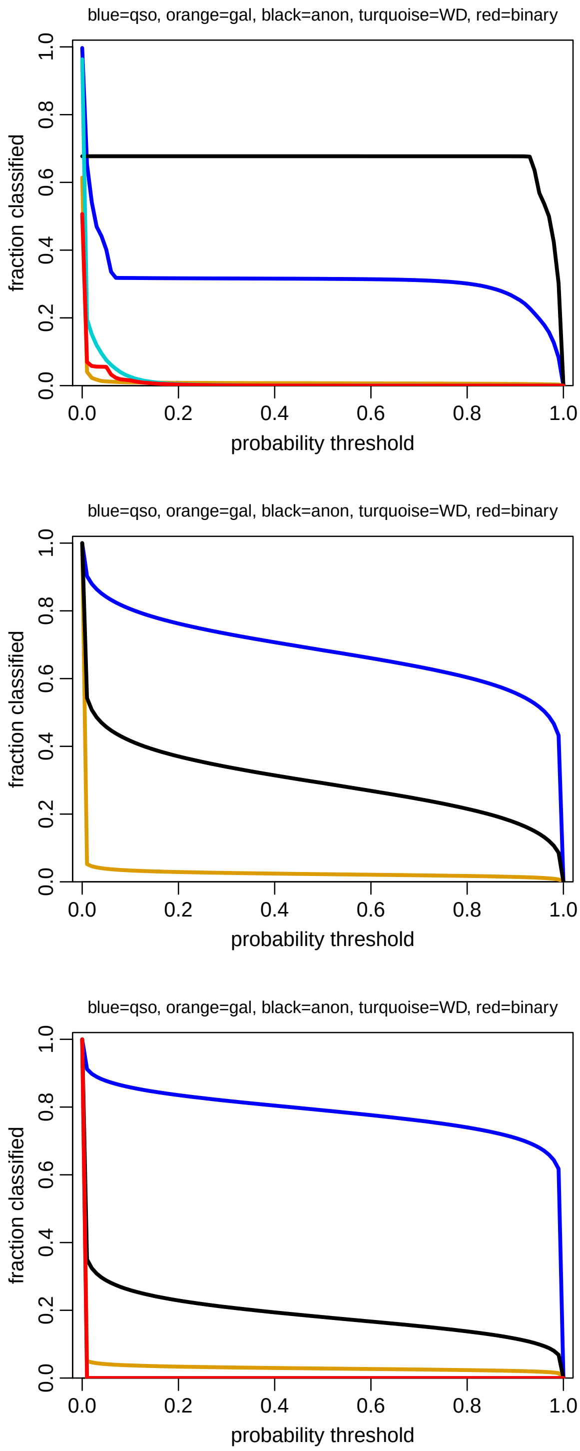


Figure 58: Fraction of the Klioner AGN assigned to each of the DSC classes when varying the probability threshold for classification, for Specmod (top), Allosmod (centre), and Combmod (bottom).