Astronomy & Astrophysics

The miniJPAS survey quasar selection

IV. Classification and redshift estimation with SQUEzE

Ignasi Pérez-Ràfols[1,2,3] ![ORCID], Luis Raul Abramo[4], Ginés Martínez-Solaeche[5], Matthew M. Pieri[6], Carolina Queiroz[4],
Natália V. N. Rodrigues[4], Silvia Bonoli[7,8], Jonás Chaves-Montero[2,7], Sean S. Morrison[6,9], Jailson Alcaniz[10],
Narciso Benitez[11], Saulo Carneiro[12], Javier Cenarro[13], David Cristóbal-Hornillos[14], Renato Dupke[10],
Alessandro Ederoclite[13], Rosa M. González Delgado[11], Antonio Hernán-Caballero[13], Carlos López-Sanjuan[13],
Antonio Marín-Franch[13], Valerio Marra[15,16,17], Claudia Mendes de Oliveira[18], Mariano Moles[14], Laerte Sodré Jr[18],
Keith Taylor[19], Jesús Varela[14], and Héctor Vázquez Ramió[13]

(Affiliations can be found after the references)

ABSTRACT

Aims. Quasar catalogues from photometric data are used in a variety of applications including those targeting spectroscopic follow-up, measurements of supermassive black hole masses, Baryon Acoustic Oscillations, or non-Gaussianities. Here, we present a list of quasar candidates including photometric redshift estimates from the miniJPAS Data Release constructed using SQUEzE. miniJPAS is a small proof-of-concept survey covering 1 deg^2 with the full J-PAS filter system, consisting of 54 narrow filters and 2 broader filters covering the entire optical wavelength range.

Methods. This work is based on the machine-learning classification of photometric data of quasar candidates using SQUEzE. It has the advantage that its classification procedure can be explained to some extent, making it less of a 'black box' when compared with other classifiers. Another key advantage is that the use of user-defined metrics means the user has more control over the classification. While SQUEzE was designed for spectroscopic data, we have adapted it for multi-band photometric data; that is we treat multiple narrow-band filters as very low-resolution spectra. We trained our models using specialised mocks. We estimated our redshift precision using the normalised median absolute deviation, σ_{NMAD}, applied to our test sample.

Results. Our test sample returns an f_1 score (effectively the purity and completeness) of 0.49 for high-z quasars (with $z \geq 2.1$) down a to magnitude of $r = 24.3$ and 0.24 for low-z quasars (with $z < 2.1$), also down to a magnitude of $r = 24.3$. For high-z quasars, this goes up to 0.9 for magnitudes of $r < 21.0$. We present two catalogues of quasar candidates including redshift estimates: 301 from point-like sources and 1049 when also including extended sources. We discuss the impact of including extended sources in our predictions (they are not included in the mocks), as well as the impact of changing the noise model of the mocks. We also give an explanation of SQUEzE reasoning. Our estimates for the redshift precision using the test sample indicate a $\sigma_{NMAD} = 0.92\%$ for the entire sample, reduced to 0.81% for $r < 22.5$ and 0.74% for $r < 21.3$. Spectroscopic follow-up of the candidates is required in order to confirm the validity of our findings.

Key words. quasars: general – methods: data analysis – techniques: photometric – cosmology: observations

1. Introduction

In recent years we have seen the appearance of several very large spectroscopic surveys, including hundreds of thousands of spectra. Among these are dedicated cosmology programmes such as the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013), which is part of the third extension of the Sloan Digital Sky Survey (SDSS-III; Eisenstein et al. 2011), where the spectra of quasars play a key role in constraining cosmological parameters. Indeed, BOSS quasars were used in the first baryon acoustic oscillations (BAOs) measurement using the Lyman α (Lyα) forest auto-correlation (Busca et al. 2013; Slosar et al. 2013; Kirkby et al. 2013) and its cross-correlation with quasars (Font-Ribera et al. 2013).

These Lyα BAO measurements were refined in the extended BOSS Survey (eBOSS; Dawson et al. 2016), part of SDSS-IV (Blanton et al. 2017), leading to their measurements on the 16 Data Release (DR16) by du Mas des Bourboux et al. (2020). Quasars in eBOSS were also used to perform BAO clustering analysis (Hou et al. 2021; Neveux et al. 2020). Both analyses were included in the final cosmological results from eBOSS (Alam et al. 2021).

These large spectroscopic surveys use multi-object spectroscopy to achieve such a large number of objects in a reasonable amount of time. In practice, this requires the identification of quasars (or any other object of interest) in photometric surveys to know where to place the optical fibres. The same is true for the next generation of surveys aiming to construct a large spectroscopic quasar sample. The next generation of surveys has already started collecting data.

This next generation includes the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration 2016a,b), which uses targeting data from the DESI Legacy Survey programmes (Dey et al. 2019), and the WEAVE-QSO Survey (Pieri et al. 2016), part of the *William Herschel* Telescope Enhanced Area Velocity Explorer Collaboration (WEAVE; Dalton et al. 2016), which will use targeting data from the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS; Benitez et al. 2014). They will observe a sample of quasars unparalleled in size.

Table 1. Summary of the samples used in this work.

Type	Sample	Objects	Star	Galaxy	Quasar
Real data	All	40 805	…	…	…
	Point-like	10 282	…	…	…
	SDSS cross-match	272	115	40	117
	SDSS cross-match & point-like	254	27	40	113
Mocks	Training	297 959	99 931	99 109	98 919
	Validation	29 783	9991	9901	9891
	Test	29 779	9995	9897	9887
	Test 1 deg^2	9036	2187	6347	502

Notes. The first column specifies the type of sample (mocks or real data). The second and third columns give the name of the sample and the number of objects it contains. Where available, in the following columns we also provide the number of stars, galaxies, and quasars separately.

While being a photometric survey, J-PAS has the particularity of using many narrow-band filters spaced approximately every 100 Å producing pseudo-spectra (or j-spectra) of the observed objects. Therefore, in addition to providing a target sample for WEAVE-QSO, J-PAS promises to deliver a sample of sufficient quality to enable various quasar analyses from J-PAS data alone (e.g. Abramo et al. 2012). J-PAS is currently undergoing commissioning, but the authors have released the data taken with their pathfinder camera, the miniJPAS data release (Bonoli et al. 2021).

Up until now, we have discussed the use of photometric quasar catalogues as targeting catalogues in large spectroscopic surveys as this is our main motivation to build this catalogue. However, we note that photometric quasar catalogues are also interesting in their own right. There are plenty of uses for quasars in photometric surveys, including measuring supermassive black hole masses (e.g. Chaves-Montero et al. 2022), BAO (e.g. Abramo et al. 2012), and non-Gaussianity (e.g. Leistedt et al. 2014).

Previous papers in this series (Rodrigues et al. 2023; Martínez-Solaeche et al. 2023) introduced different types of machine-learning algorithms to construct a catalogue from miniJPAS data. Here, we present the classification of SQUEzE (Pérez-Ràfols et al. 2020). SQUEzE is a machine-learning code designed to identify spectra of quasars that can be extended to using the j-spectra from J-PAS (Pérez-Ràfols & Pieri 2020). The particularity of SQUEzE is that it does not only perform the classification of quasars but also provides photometric redshift estimates.

Having redshift estimates is a key feature as they are essential if this catalogue is to be used, for example, to measure BAOs (Abramo et al. 2012). In general, previous efforts to obtain photometric quasar redshift estimates used quasar templates. Examples of this are Wolf et al. (2003) on COMBO-17 data, Salvato et al. (2009, 2011) on COSMOS data, Matute et al. (2012) and Chaves-Montero et al. (2017) on ALHAMBRA data, or Mendes de Oliveira et al. (2019) on S-PLUS data. Other surveys such as J-PLUS (Cenarro et al. 2019; Spinoso et al. 2020) detect the position of the Lyman α emission line to estimate the redshifts of Lyman α emitters. However, by analysing a single emission line they are susceptible to interlopers. Also, even if the detected line is indeed Lyman α emission, they cannot distinguish between different types of Lyman α emitters (i.e. star-forming galaxies and quasars). We used an alternative method that mimics the visual inspection of quasar spectra by quasar experts. Searches are performed using SQUEzE by finding multiple emission lines, their relative wavelengths and strengths. In doing so, we simultaneously provide quasar identification and redshift estimation.

We start by describing the data we used in Sect. 2. Then, we describe SQUEzE behaviour and its particularities when running it on J-PAS j-spectra in Sect. 3 and present the results on synthetic data (mocks) in Sect. 4. We present our catalogues of quasar candidates in Sect. 5 and discuss our findings in Sect. 6. Finally, we summarise our conclusions in Sect. 7.

2. miniJPAS data and mocks

In this Section, we describe the datasets used in this work. The number of objects in each sample is summarised in Table 1. Where known, we also give the number of quasars, galaxies, and stars separately. We now describe each of these samples.

In this work, we used data from the First Data Release of J-PAS, also known as the miniJPAS survey (Bonoli et al. 2021). The miniJPAS survey is a photometric survey using 56 filters, of which 54 are narrow band filters with a Full Width at Half Maximum of ~140 Å and two are broader filters extending to the ultraviolet and the near-infrared. These 56 filters are complemented by the u, g, r and i SDSS broadband filters. The survey covers ~1 deg^2 on the AEGIS field.

For this work, we used the sources identified using the software SEXTRACTOR (Bertin & Arnouts 1996) using its dual mode. This means that the positions and sizes of the apertures used to estimate the photometry are derived from the reference filter (SDSS r-band). We refer the reader to Bertin & Arnouts (1996) and Bonoli et al. (2021) for more detailed explanations of the software and object detection. The observations were carried out with the 2.55 m T250 telescope at the Observatorio Astrofísico de Javalambre, a facility developed and operated by the Centro de Estudios de Física del Cosmos de Aragón (CEFCA) in Teruel (Spain) using the pathfinder instrument. This is a single CCD direct imager ($9.2k \times 9.2k$, 10 μm pixel) located at the centre of the T250 field of view with a pixel scale of 0.23 arcsec pix^{-1}, vignetted on its periphery. It provides an effective FoV of 0.27 deg^2.

In this dual catalogue, there are a total of 64 293 identified objects[1]. A fraction of these objects are flagged as having known issues (see Bonoli et al. 2021 for a description of the flags). We discarded flagged objects to construct a clean sample of 46 440 objects. However, since high-redshift quasars are

---

[1] Available at https://archive.cefca.es/catalogues/minijpas-pdr201912

typically point-like sources, our main sample is limited to point-like sources by using the stellarity index constructed from image morphology using Extremely Randomised Trees (ERT; Baqui et al. 2021). Following Queiroz et al. (2023), Rodrigues et al. (2023), and Martínez-Solaeche et al. (2023), we require objects to be classified as stars (point-like sources) with a probability of at least 0.1, defined in their catalogue as ERT ≥ 0.1. In some cases, the ERT classification failed (identified as ERT = –99.0). In these cases, we used the alternative classification using the stellar-galaxy locus classifier from López-Sanjuan et al. (2019), requiring a minimum probability of SGLC ≥ 0.1. 11 419 objects meet this point-like source criterion and constitute our point-like sample.

A small number of the objects observed in miniJPAS have spectroscopic observations from other surveys. This allows us to have a spectroscopically confirmed classification of these objects. In particular, 272 objects were also observed by SDSS, 117 of which are quasars, 40 are galaxies, and 115 are stars. Of these, 18 were not classified as point-like sources following the aforementioned criteria, four are quasars, one is a star, and 13 are galaxies.

In this work, apart from using miniJPAS data, we also used synthetic data (mocks). This is necessary because larger data volumes with associated truth tables are needed than are currently available. The mocks we used are based on SDSS spectra convolved with the J-PAS filters and with added noise to match miniJPAS expected signal-to-noise ratios. More details on the mocks can be found in Queiroz et al. (2023). There are a total of 360 000 objects distributed between the training (300 000), validation (30 000), and test (30 000) sets. They are evenly split among quasars, galaxies, and stars. Additionally, we have a special 1 deg^2 test set that has the expected relative fraction for each type of object. In general, we used the mocks generated using noise model 11, since that is the noise model believed to be closest to the actual noise distribution from miniJPAS data, but we checked the impact of choosing a different noise model in Appendix A.

Both in data and in mocks, we restricted ourselves to r-band magnitudes $17.0 < r \leq 24.3$. In the process of creating the mocks, the original spectra are rescaled to match the expected magnitude distribution. Noise is added after this rescale is done, modifying the reported values of the flux (and thus the magnitudes). This means that a few of the mocks end up with a magnitude that is fainter than 24.3. We discarded these spectra. Overall, we analysed 40 805 miniJPAS sources, 10 282 of which meet the criterion of being point-like sources. 272 sources have spectroscopic observations from SDSS, 254 of which meet the criterion of being point-like sources. SQUEzE is trained using 99 931 stars, 99 109 galaxies, and 98 919 quasars and validated using 9991 stars, 9901 galaxies, and 9891 quasars. The test sample contains 9995 stars, 9897 galaxies, and 9887 quasars. The special 1 deg^2 test sample contains 2187 stars, 6347 galaxies, and 502 quasars.

3. SQUEzE description and setup

In this section, we provide a brief description of SQUEzE and explain the particularities of applying it to photometric data from J-PAS. A full, detailed description of SQUEzE is given in Pérez-Ràfols et al. (2020). SQUEzE is a quasar classifier that works in three steps. In the first step, we identify peaks in the spectra. We then assign trial redshifts to these peaks in the second step, and we end by classifying these trial redshifts to

discriminate between the correct and incorrect identifications. We now detail each of the steps.

3.1. Peak identification

The first step is peak identification. In this step, emission lines are identified in the spectra. In SQUEzE, this step is performed using a very simple peak finder: each spectrum is first smoothed, and then peaks are located by finding those pixels with higher flux than the two nearby pixels.

miniJPAS data contain j-spectra with data in 56 filters, and the broad emission lines are typically expected to cover three filters (though some very broad quasar emission lines can span more than three J-PAS filters; see e.g. Fig. 2 of Chaves-Montero et al. 2022). This means that any smoothing we apply will typically decrease the signal-to-noise ratio of the peak detection. However, using this simple peak finder, we have many peaks that are purely arising from noise: cases where a filter has more flux than their neighbours.

To solve this, we developed a new, more refined peak finder[2] (see Appendix B for details of its performance compared to the original peak finder). The new peak finder works as follows. First, a power-law fit is applied to reproduce the continuum emission. Outliers to this fit, defined as the data points that are off the fit by more than N sigma, are discarded, and the process is repeated until convergence is reached that is when no data points are discarded in an iteration. Here, N is the minimum significance to detect outliers, and we choose $N = 2$ as our fiducial choice (see Appendix C.2). Upon fit convergence, the outliers below the model are discarded and the outliers above the model are kept as emission peaks. Contiguous peaks (i.e. with pixel number i, $i + 1$, $i + 2$, ...) are compressed into a single peak by performing a weighted average of their wavelengths. The weights are defined by the significance of the outliers. Sometimes, too many pixels are discarded as outliers and the power-law fit fails. In those cases, no emission peaks are reported and the spectra are discarded. The overall performance of SQUEzE is improved when the new peak finder is used (see Appendix B).

3.2. Trial redshifts

Once the peaks have been identified, a list of trial redshifts is generated. For each peak of each spectrum, a trial redshift, z_{try}, is computed assuming that the peak corresponds to the Lyα, C IV, C III], Mg II, Hα, and Hβ emission lines. Negative trial redshifts are immediately discarded. Line metrics are computed for each of the remaining trial redshifts as described in Eqs. (1)–(3) of Pérez-Ràfols et al. (2020). These metrics describe the amplitude of the line, its significance, and the slope at the base of the line. For each trial redshift, we compute the metrics for 17 bands (see Table C.1) corresponding to the predicted position of quasar emission lines and other relevant features (see Appendix C.3 for details). These bands are defined at the potential quasar rest frame, and therefore the spectral coverage of these bands will change as a function of redshift. Figure 1 shows this evolution as the number of filters used in line metrics as a function of redshift.

The wavelength bands used to compute line metrics were designed and optimised for spectra from BOSS, with a resolution of ∼1 Å, while here the resolution is 140 Å and the separation between filter centres is 100 Å. We therefore explored whether the size of the bands impacts the classification performance. The boundaries of these bands have been tuned to a more reliable

---

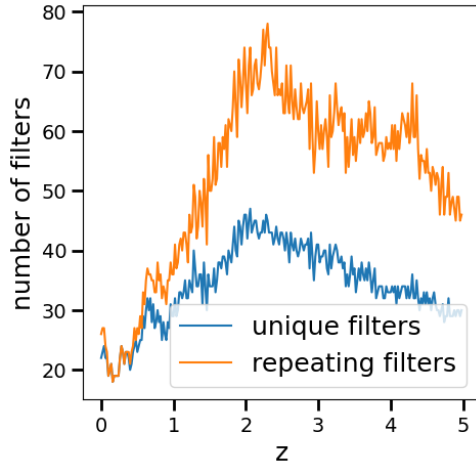[2] The new peak finder is now included in the SQUEzE package.

Fig. 1. Number of J-PAS filters used to compute the metrics as a function of redshift. The blue line shows the number of unique filters used in the computation of the metrics. Given that some bands are overlapping, the total number of filters used is generally larger, as shown by the orange line.

Fig. 2. Magnitude distribution of trial redshifts in the training sample. This distribution does not match the distribution of objects as there are typically a few trial redshifts per object.

separate peak from the side band in the data of this resolution. Furthermore, given our limited use of the filters available, we added bands to allow the machine learning algorithm (see Sect. 3.3) access to information on the absence of emission lines as well as the presence of them. These 'flat-emission-line' metrics are placed at wavelengths where redshift confusion leads to an emission line arising where one should not occur. The desired lack of emission line signal can hence be part of the random forest classification (see Appendix C.3 for more details on the tests that led to our selection of this 'wide+extra' set of wavelength bands). We note that even with our more inclusive approach, only a relatively small number of the filters (out of a total of 60 available) are used to compute the metrics and, thus, to identify the quasars (see the left panel in Fig. C.3).

3.3. Classification

Once we have a list of trial redshifts and associated metrics, they are fed to the random forest classifiers. In training mode, trial redshifts are flagged as correct peaks if the spectra are that of a quasar with trial redshift at most $\Delta z = 0.10$ away from the true redshift. Pérez-Ràfols et al. (2020) used a larger value of 0.15 for this criterion, but we obtained better redshift errors (without a decrease in performance) by using a tighter constraint (see Appendix C.5).

Pérez-Ràfols et al. (2020) used two different classifiers, one for high-redshift quasars and another one for low-redshift quasars. The split in redshift was performed at $z = 2.1$ since this is where the Lyα emission line enters the spectra. They argued that a single classifier could be used but that they observed better performance when splitting by redshift. The reason for this is that high-redshift quasars have more emission lines compared to low-redshift quasars. We checked that this statement is also valid for our dataset (see Appendix C.4) and decided to also use two random forests. In SQUEzE default choices, only the metrics are passed to the random forest classifiers. However, we note that by also passing the trial redshift and the r-band magnitude we obtain slightly better results (see Appendix C.6). Thus, we adopted this change of the default settings.

The final stage in the classification is to select, for each spectrum, the trial redshift with the highest probability. At this point,
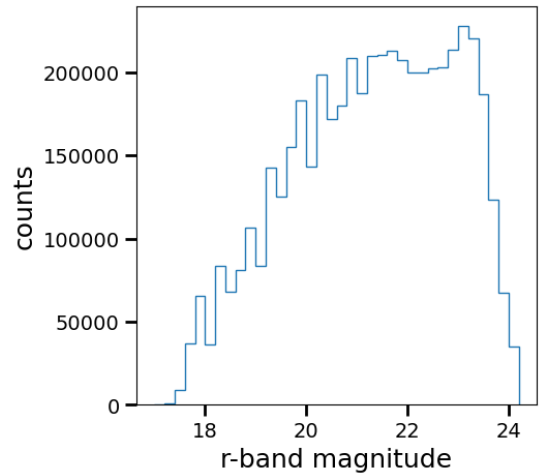
it is worth noting that it is more convenient to separate quasars by their observed r-band magnitude as the faint quasars dominate the training set. This can be seen in Fig. 2, where we show the magnitude distribution of the list of trial redshifts. Here, we note that this distribution is different from the distribution of objects, as each object will typically have a few trial redshifts. As explained in detail in Appendix C.1, we ran SQUEzE in four magnitude bins: $r \in (17.0, 20.0]$, $r \in (20.0, 22.5]$, $r \in (22.5, 23.6]$, and $r \in (23.6, 24.3]$.

4. Performance

We assessed the performance of SQUEzE based on the test sample results. In any sample (either from mocks or from real data), there are quasars, galaxies, and stars. However, we need to keep in mind that here we were interested in creating a quasar catalogue. Thus, in terms of performance, we do not need to penalise the cases where stars are classified as galaxies and vice versa. We measured our performance level based on the correctly classified quasars. However, for a correct classification, we required not only that the object be a quasar, but also that its redshift be correct. Formally, we require $\Delta z = |z_{\text{true}} - z_{\text{try}}| < 0.10$ (see Appendix C.5), as this is enough to ensure that we are not suffering from line confusion (i.e. finding a true emission line but failing to label it correctly). We note, however, that the actual redshift precision is typically better (see below).

We define purity p as the number of true quasars (at the correct redshift) in the catalogue over the total number of sources classified as quasars, and completeness c as the number of true quasars in the catalogue (again, at the correct redshift) over the total number of true quasars in the sample analysed. For each of the classifications, we also have the confidence of the classification, given by the fraction of decision trees that agree with that classification. To some extent, we can tune the purity and completeness of the sample by applying some cuts on this confidence of classification.

A higher confidence requirement will result in a purer but less complete sample. Similarly, a lower confidence requirement will result in a more complete, but less pure, sample. Even though the choice of a confidence threshold can be tuned for specific analysis, a common general-purpose choice is to balance purity and completeness. An optimised balance can be found by

maximising the f_1 score, which is defined as

$$f_1 = \frac{2pc}{p+c}. \tag{1}$$

We note that with this definition performance is expected to be worse than the other classifiers presented in the companion papers (Rodrigues et al. 2023; Martínez-Solaeche et al. 2023). Part of the reason for this is that they have a more relaxed criterion to determine good classifications. Since they are not measuring redshifts, they require the quasars to be correctly classified as high-redshift quasars ($z \geq 2.1$) or low-redshift quasars ($z < 2.1$). We also adopt this criterion to make a more direct comparison. We denote this criterion as f_1^*.

4.1. Test sample

The top panels in Fig. 3 show the performance as a function of limiting magnitude. Blue solid lines correspond to the f_1 score, whereas the orange dashed lines show the more relaxed criterion f_1^*. For each limiting magnitude, we perform a cut in the confidence threshold of the classification such that the f_1 score is maximised (green dotted lines). As expected, the performance drops as fainter objects are added to the sample. This is because fainter objects are more difficult to classify as they are noisier and have a larger number of filters with non-detections. The f_1 score including all objects down to $r = 24.3$ is 0.49 (with a confidence threshold of 0.55) for high-z quasars and 0.24 for low-z quasars (with a confidence threshold of 0.39). The values of f_1^* are higher than those of f_1, as expected. Including all objects down to $r = 24.3$ its values are 0.56 for high-z quasars (with a confidence threshold of 0.58) and 0.41 for low-z quasars (with a confidence threshold of 0.32).

The comparison of these results with those obtained by the algorithms from Rodrigues et al. (2023) and Martínez-Solaeche et al. (2023) is not straightforward. They report the averaged f_1 score including the f_1 for high-z quasars, low-z quasars, galaxies and stars. Here, by using f_1^*, we can only compute equivalent quantities for high-z quasars and low-z quasars. As such, only a qualitative comparison is possible. Even so, we provide our measurement of f_1^* using the same magnitude bins in Fig. 4. This figure should be compared with the top panel of Fig. 4 of Rodrigues et al. (2023) and with Fig. 1 of Martínez-Solaeche et al. (2023).

We can see that SQUEzE performance is qualitatively higher than the RF from Rodrigues et al. (2023). It has a qualitatively similar performance level compared to LGMB and CNN1 (without errors), also from Rodrigues et al. (2023), and is qualitatively lower than CNN1, CNN2 from Rodrigues et al. (2023), and the classifiers from Martínez-Solaeche et al. (2023). However, this is expected as our method tackles the harder problem of solving both the redshift estimation and the classification problems.

We now move on to analysing the contaminants of our sample. To do so, we split it into four different magnitude bins: $17 < r \leq 20$ (bin 1), $20 < r \leq 22.5$ (bin 2), $22.5 < r \leq 23.6$ (bin 3), and $23.6 < r \leq 24.3$ (bin 4). For each of the bins, we plot the predicted redshift, z_{try}, against the true redshift, z_{true}, to study the contaminants. We include only quasars with confidence levels greater than 0.39 when $z_{try} < 2.1$ and 0.55 when $z_{try} \geq 2.1$, corresponding to the confidence thresholds mentioned above. Figure 5 shows galaxies, stars, and quasar contaminants as orange up-pointing triangles, green squares, and blue down-pointing triangles, respectively. Black dots show correct classifications, i.e., those that fulfil the criteria of

$\Delta z = \left| z_{true} - z_{try} \right| < 0.10$ (red band). Grey bands also signal the areas where quasar contaminants (blue down-pointing triangles) are correctly classified in the relaxed classification criterion, i.e., without a redshift requirement.

Bin 1 behaves as expected. Correct classifications are found very close to the red line, showing that the redshift precision is significantly better than the required value of 0.10 (indicated in the plot by the red stripe). Quasar contaminants are rare and follow straight lines showing that there is some degree of line confusion, i.e. we correctly find a quasar emission line but we fail to identify the line responsible for the emission (and thus the redshift error). Galactic contaminants also follow the same straight lines as quasars, suggesting that these galaxies might contain active galactic nuclei (AGNs) with broad emission lines or – more likely – star-formation emission lines (Hα, Hβ, O [III], O [II]) that are misidentified as QSO emission lines (Chaves-Montero et al. 2017). Stellar contaminants are distributed at trial redshifts lower than 2.1. This indicates that only the low-z classifier is adding stellar contaminants.

This simple picture starts to break as we go to bin 2. We see two effects. First, while we still see clear line confusion, we also see some quasar and galactic contaminants that are no longer distributed along these lines, indicating that we are no longer able to always distinguish real emission line peaks from noise peaks. Apart from this, we see by eye that the redshift precision of the correct classifications is significantly worse (see below for a more quantitative statement). This suggests that the chosen redshift tolerance was too large. We discuss this further in Appendix C.5, where we conclude that this is not the case.

This issue is aggravated as we go to bin 3. Now we are not able to see the confusion lines as clearly as before (though some can still be seen). This suggests that our ability to distinguish real peaks from noise peaks starts to break somewhere around magnitude $r \sim 22.5$ (see also Appendix B). We note that in this bin we see an apparent cut of the contaminants in the redshift at $z = 1.5$. Below this redshift, the C IV line is not observable in our spectral coverage. Together with the Lyα line, they are the stronger lines. Thus, the confidence in classification is generally lower whenever they are not present. In practice, this means that many of the trial redshifts below this redshift either do not meet the minimum required classification confidence; otherwise other trial redshifts for the same quasar are preferred.

Finally, for bin 4, there is a strong decrease in the number of contaminants. There are two reasons behind this. First, we are at the faint end of our sample and therefore the number of objects decreases compared to bin 3. Second, the spectra are so noisy that we obtain very few confident classifications.

Overall, a significant fraction of the redshift confusion is causing high-z quasars (with $z \geq 2.1$) to be classified as low-z quasars. This can be seen in the upper left quadrants in Fig. 5. In lower numbers, the same occurs in the opposite direction (lower right quadrants). This can explain the drop in performance compared to the results from Rodrigues et al. (2023) and Martínez-Solaeche et al. (2023) even when we consider the same relaxed criteria f_1^*. However, the fact that there is more than one trial redshift per quasar indicates that if the high- or low-redshift classification could be fixed by these other algorithms, SQUEzE can still be used to provide a redshift estimate (see Pérez-Ràfols et al., in prep.).

We now turn our attention to the redshift precision. As explained in Sect. 3, we formally require a precision of 0.10, but the performance is expected to be much better. Figure 5 shows that this is not the case for the fainter bins. We now quantify
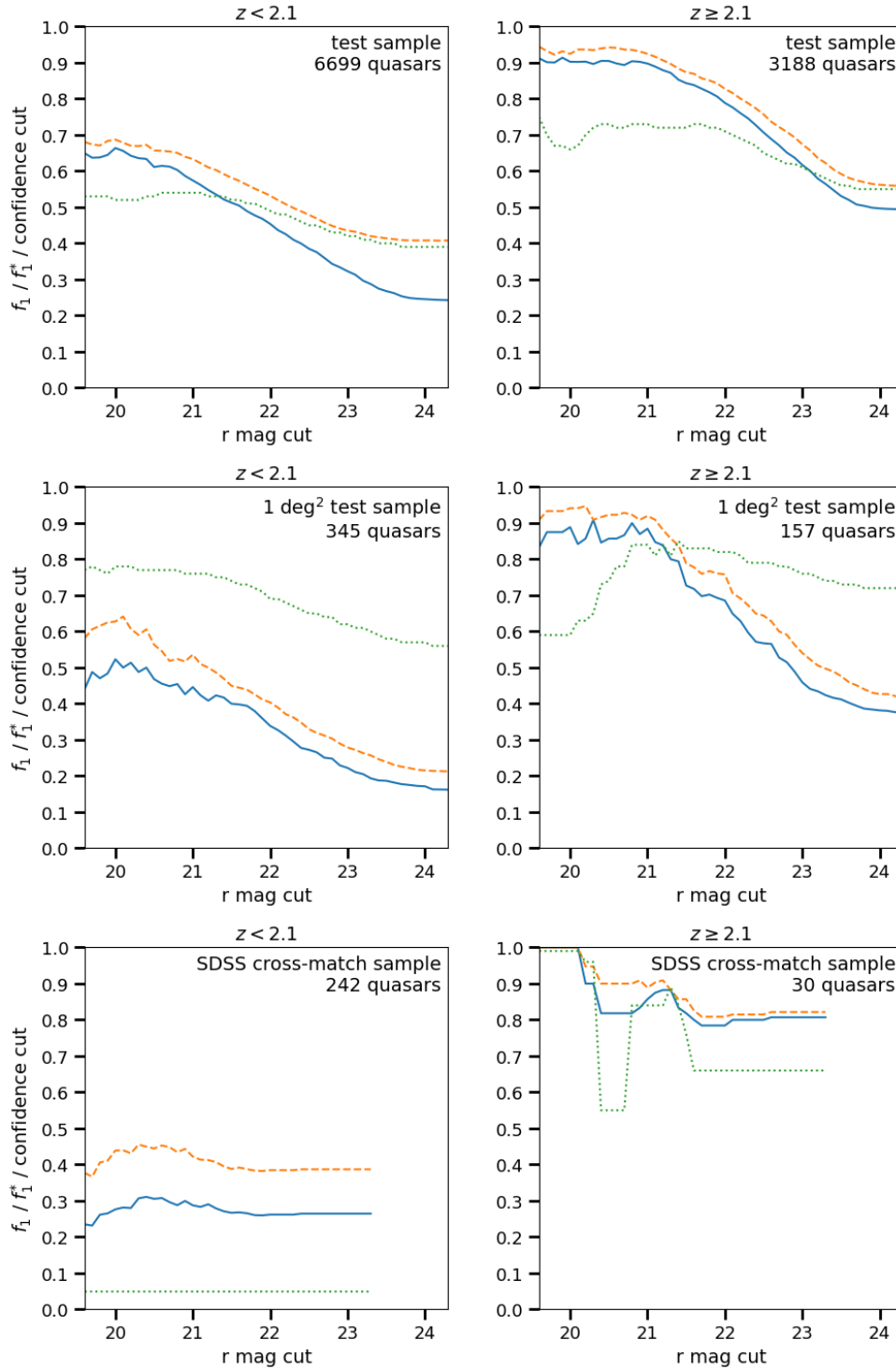
Fig. 3. Performance as function of limiting magnitude. All objects brighter than the magnitude cut in the r band are considered to compute the f_1 score. Blue solid lines show the f_1 score as defined in Eq. (1), and the orange dashed lines show the more relaxed statistic f_1^* (see text for details). Green dotted lines show the confidence threshold used to compute the f_1 score. From top to bottom, we show results for the test sample (Sect. 4.1), the 1 deg^2 test sample (Sect. 4.2), and the SDSS cross-match sample (Sect. 4.3). We note that in the bottom panels, the lines stop at magnitude 23.4; the sample does not have fainter objects.

this statement. We take the correct classifications and measure the distribution of Δz for the different magnitude bins. Results of this exercise are shown in Fig. 6 and in the first block of Table 2. Indeed, the redshift error increases as we go to fainter magnitudes. In fact, our bright bin (bin 1, with $17.0 < r \leq 20.0$) has a typical redshift error of $\sim$2800 km s^{-1}, which is less than two-thirds of the typical error in our faint bin (bin 4, $\sim$4700 km s^{-1}). We also see that there is no significant bias in our measurement of the redshift (the mean offset is an order of magnitude smaller than the typical error).

Table 2. Statistics of the redshift precision.

Sample	Mag bin	$\overline{\Delta v}$ (km s^{-1})	$\sigma_{\Delta v}$ (km s^{-1})	σ_{NMAD} (per cent)	N
Test	$17.0 < r \leq 20.0$	704.69	2833.20	0.71	504
	$20.0 < r \leq 22.5$	809.09	3083.64	0.94	2265
	$22.5 < r \leq 23.6$	859.76	4333.32	1.51	983
	$23.6 < r \leq 24.3$	372.16	4702.21	1.56	183
1 deg^2 test	$17.0 < r \leq 20.0$	1558.23	2753.08	0.75	35
	$20.0 < r \leq 22.5$	787.62	3212.02	0.86	104
	$22.5 < r \leq 23.6$	2121.36	4577.93	1.80	58
	$23.6 < r \leq 24.3$	2954.34	5542.67	1.41	9
SDSS cross-match sample	$17.0 < r \leq 20.0$	1667.23	2361.88	0.73	22
	$20.0 < r \leq 22.5$	1263.28	2625.00	0.88	61
	$22.5 < r \leq 23.6$	1084.38	NaN	0.53	1

Notes. Mean redshift offset ($\overline{\Delta v}$), dispersion measured ($\sigma_{\Delta v}$), normalised median absolute deviation (σ_{NMAD}), and number of correctly classified quasars (N) in the samples test, 1 deg^2 test, and SDSS cross-match. The mean offset indicates potential biases of our redshift estimate and the dispersion indicates our typical redshift error. The number of objects indicates how reliable the measured statistics are.
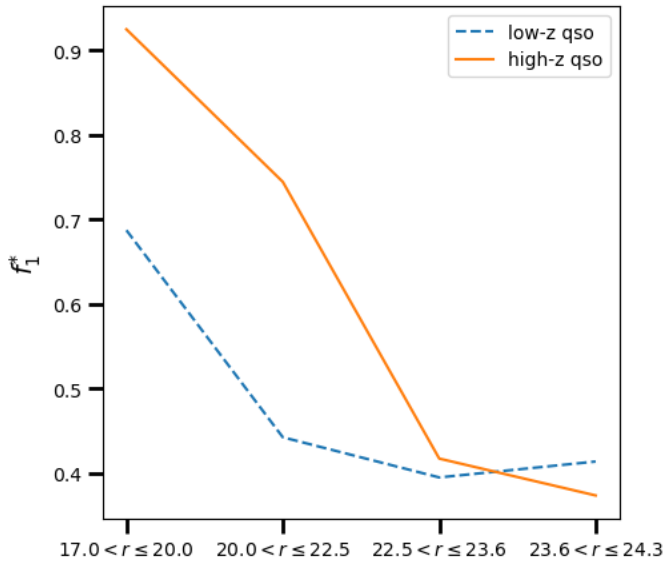
Fig. 4. f_1^* measured using the same bins as Rodrigues et al. (2023) and Martínez-Solaeche et al. (2023). The solid orange line shows the f_1^* score for high-z quasars and the dashed blue line for low-z quasars.

We also provide, in Table 2, the normalised median absolute deviation, σ_{NMAD}, defined by Hoaglin et al. (1983) as

$$\sigma_{\mathrm{NMAD}} = 1.48 \times \mathrm{median}\left(\frac{\left|z_{\mathrm{try}} - z_{\mathrm{true}}\right|}{1 + z_{\mathrm{true}}}\right). \qquad (2)$$

This quantity is less sensitive to redshift outliers than the standard deviation. Nevertheless, we observe the same trend here as we do for the standard deviation.

4.2. 1 deg^2 test sample

We now focus on the special 1 deg^2 test sample, which differs from the normal test sample in the relative number of objects. This sample has the number of quasars, stars, and galaxies expected in a square degree on the sky. Thus, in proportion, the number of quasars is significantly smaller. The middle

panel of Fig. 3 shows the performance as a function of limiting magnitude. We see a similar trend as for the test sample. The f_1 score including all objects down to $r = 24.3$ is 0.38 (with a confidence threshold of 0.72) for high-z quasars and 0.16 for low-z quasars (with a confidence threshold of 0.56). The values of f_1^* including all objects down to $r = 24.3$ are 0.42 (with a confidence threshold of 0.83) for high-z quasars and 0.21 for low-z quasars (with a confidence threshold of 0.68). This decrease compared to the test sample is expected as we now have a larger fraction of contaminants.

We now analyse the redshift precision in this sample. As for the test sample, we compute the distribution of Δz for different magnitude bins. The results are shown in Fig. 7 and tabulated in the second block of Table 2. The distribution of redshift errors in bins 1, 2, and 3 are similar to those of the test sample. For the faint bin, we clearly do not have enough statistics to say anything meaningful.

4.3. SDSS cross-match sample

More interesting than the performance with regard to mock data is that on real data. However, we are limited in this assessment by the lack of a large sample with an available truth table. The only samples with reliable spectroscopic confirmation of the object classes are the SDSS cross-match samples (including and excluding extended objects). We note that both samples are very small (see Table 1) and that they are biased as they only contain the brightest objects. Even though they are not included in the mocks, we include the 18 objects not classified as point-like sources in the performance assessment to have a sample as large as possible. We note that the results stay the same including or not including these 18 objects.

The performance as a function of magnitude is given in the bottom panel of Fig. 3. The distribution of Δz for the classifications in Fig. 8 and summarised in the third block of Table 2. Due to the small size of the sample, the measured f_1 score distribution is much noisier than in the mock test sample. However, the results suggest a similar performance to the 1 deg^2 test sample. The redshift distribution is also noisier.

This sample does not include any object for the faintest bin (bin 4) and only one object for bin 3. This is important as the algorithm has more difficulties when classifying objects in these
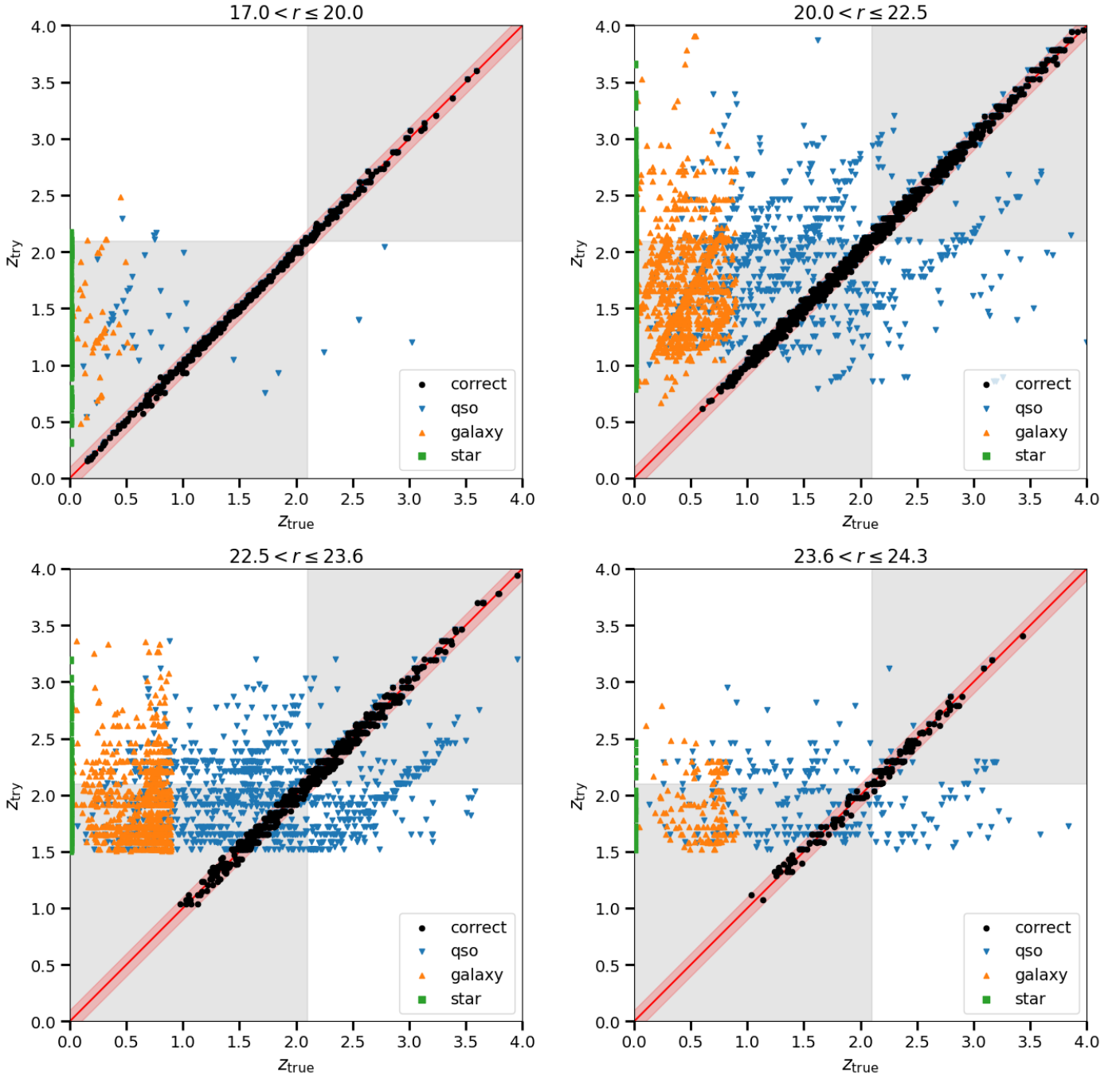
Fig. 5. SQUEzE trial redshift, z_{try}, versus true redshift, z_{true}, for the test sample. Black dots indicate correct classifications. Blue down-pointing triangles, orange up-pointing triangles, and green squares indicate the quasar, galactic, and stellar contaminants respectively. The red solid line shows the perfect classification line and the red stripe show the redshift offset tolerance ($\Delta z = 0.10$). Grey squares indicate the area where quasar contaminants are deemed as correct in the relaxed classification scheme (see text for details). From top to bottom and left to right, panels show the data split in four different magnitude bins: $17 < r \le 20$, $20 < r \le 22.5$, $22.5 < r \le 23.6$, and $23.6 < r \le 24.3$. We note that stellar contaminants (green dots) are always found at $z_{true} = 0$.

fainter bins. In order to properly assess the performance on data, a larger sample of spectroscopic observations would be needed, particularly including the objects at the faint end.

5. miniJPAS quasar catalogue

After assessing the performance of SQUEzE, we shifted our attention to the actual catalogue. We created two different catalogues:

one including only point-like sources and one including extended sources as well (sample *all*). These samples are described in Sect. 2 and in Table 1.

We ran SQUEzE on these two samples and added objects with a classification confidence higher than the cut to the catalogue (green dotted lines in the top panel of Fig. 3). Here, we used the thresholds from the test sample. Even though the number of contaminants in the data should be closer to the 1 deg^2 test sample, this sample is too small for its cuts in classification confidence to
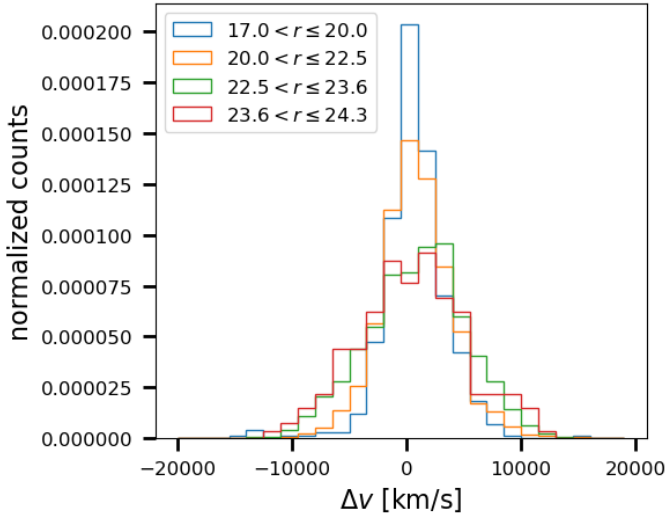
Fig. 6. Distribution of $\Delta z = \left| z_{\text{true}} - z_{\text{try}} \right| < 0.15$ for the test sample for four magnitude bins: $17 < r \leq 20$, $20 < r \leq 22.5$, $22.5 < r \leq 23.6$, and $23.6 < r \leq 24.3$.
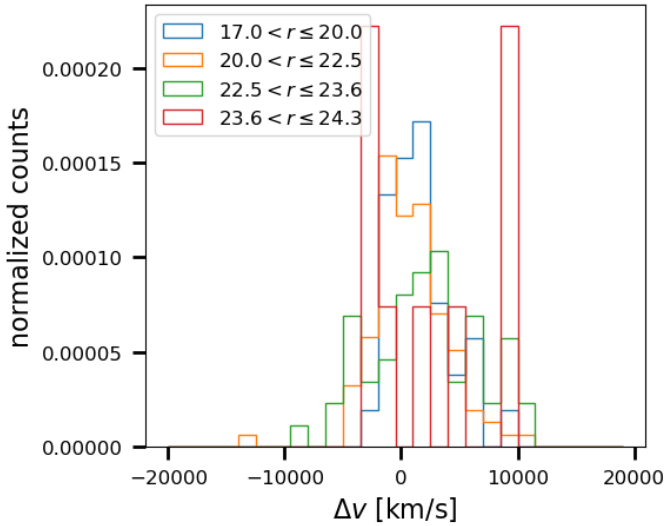
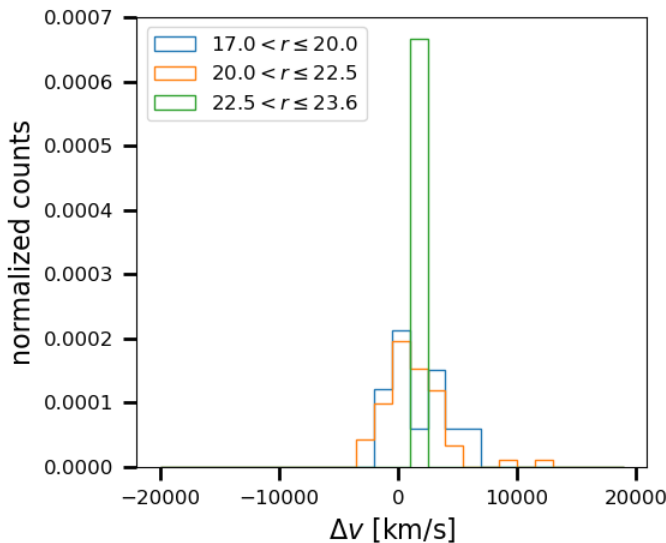**Fig. 7.** Same as Fig. 6, but for the 1 deg^2 test sample.

Fig. 8. Same as Fig. 6, but for the SDSS cross-match sample.

be statistically robust. Thus, we chose those of the test sample. For the final catalogue, we also dropped entries flagged as duplicated to keep only one entry per object. For some objects, no peaks were found, and thus they did not enter the random forest classifiers. This occurred for 906 objects in the point-like sample and 3665 objects in the entire sample. These are dropped from our final catalogue.

The final catalogue contains 301 quasar candidates for the point-like sample and 1049 when also including extended sources. Applying the same criteria for the 1 deg^2 test sample, we obtain a catalogue of 412 quasar candidates. These numbers should be compared to those of the point-like sample, as the mocks were built to match that sample. The similarity between the number of candidates could be suggestive of similar behaviour of SQUEzE on data and mocks.

To go further, we compared the magnitude and redshift distributions of the candidates in the point-like sample to the distributions of the candidates in the 1 deg^2 test sample (Fig. 9). We started with the magnitude distribution (left panel of Fig. 9). There is a deviation of the point-like sample towards fainter magnitudes. While small deviations are expected given the relatively small sample sizes, this could also indicate SQUEzE performs differently in data and mocks at the faint end. Mocks were created from brighter SDSS data, so it would not be surprising if a different type and/or distribution of contaminants appears at the faint end. A different population of contaminants could easily induce a different behaviour in the classifier. A spectroscopic follow-up of these sources is needed to confirm or deny this apparent discrepancy.

The redshift distribution of the samples point-like and all (right panel of Fig. 9) are similar, except for a peak at $z \sim 3$, present only when extended sources are included. This could hint that SQUEzE performance on extended sources is similar to that of point-like sources, even if it was only trained on point-like sources. There are two possible explanations for this behaviour. First, the algorithm to separate point-like sources from extended sources has a certain degree of confusion, leading to some extended sources entering the point-like sample. This does seem to be happening to some degree. For instance, in the SDSS cross-matched sample we have 115 stars, of which only 27 are classified as point-like sources. This would mean that their properties are indeed included in the training sample, thus explaining the similarity between the two distributions. This is expected to happen to some degree, particularly at the faint end, where it is not always trivial to separate the actual source from the sky contribution.

Another possible explanation is that the properties of extended and point-like quasars, as seen by SQUEzE, are similar. This also makes sense as SQUEzE focuses on the emission lines in specific spectral regions. Observing the galactic emission (and making them extended objects) would not change how SQUEzE sees the quasars. However, we note that this could change the way contaminants are seen. Most likely, the truth lies somewhere between the two explanations, but we require a larger sample, with spectroscopic confirmation of the classifications, to ascertain this.

6. Discussion

6.1. Comparison with previous performance estimates

Pérez-Ràfols & Pieri (2020) studied the potential performance of SQUEzE in different surveys including tests for a generic narrow-band survey. In particular, they mentioned that realistic JPAS
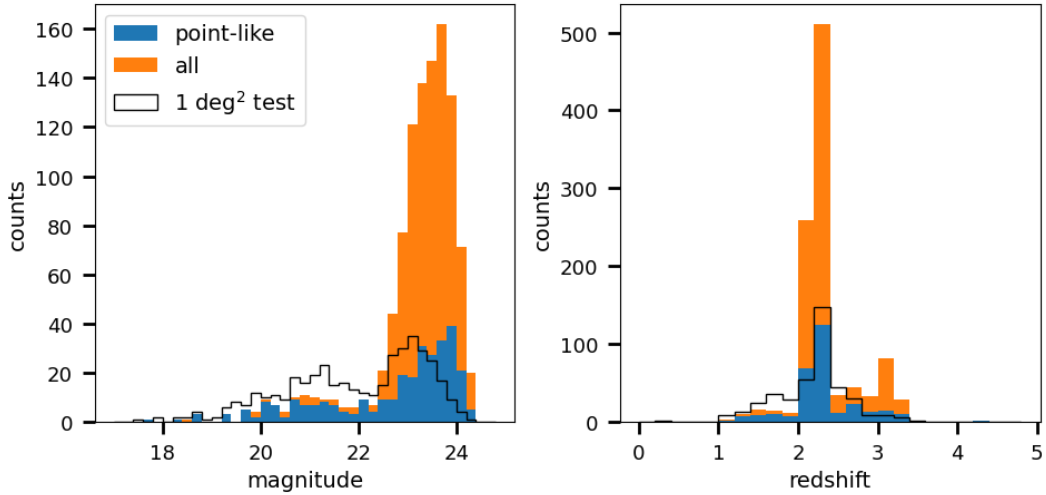
Fig. 9. Solid histograms show distribution of magnitude (left) and redshift (right) for the quasar candidates for the 'point-like' and 'all' samples (see Table 1). Both distributions look similar, except for a peak at $z \sim 3$, present only when extended sources are included, suggesting the code might also work for extended sources. For comparison, empty histograms show the same distributions for the 1 deg^2 test sample. We note that the redshift used here is the best trial redshift for each candidate.

mocks, such as those we use here, would be required to asses the performance of SQUEzE on JPAS data. Nevertheless, they suggested that their rebin100+noise4 could be used as an initial test of this performance. Based on this, they predicted the purity and completeness to be greater than 0.9. This is clearly in conflict with the results obtained here, where this statement only holds to a magnitude of $r < 21.1$. The simplest explanation for this discrepancy lies in the spectra used to asses this performance. To construct their rebin100+noise4, Pérez-Ràfols & Pieri (2020) rebinned SDSS spectra and added noise in a crude simulation of miniJPAS-like data. In this work, we assessed the performance using a set of refined mocks from Queiroz et al. (2023) that are tailored to match the observations. This discrepancy in the performance would be expected if the initial estimates from Pérez-Ràfols & Pieri (2020) were optimistic in the expected signal-to-noise ratio.

In order to test this, we rebuilt the train and test samples but instead of using the miniJPAS mocks outlined here, we follow the Pérez-Ràfols et al. (2020) prescription for building the rebin100+noise4 mocks. We computed the mean signal-to-noise ratio for our regular test sample and for the test sample rebuilt here. Figure 10 shows the histogram of these signal-to-noise ratios. We clearly see that the estimates from Pérez-Ràfols & Pieri (2020) have higher signal-to-noise ratios, confirming our hypothesis.

To further test that changes to SQUEzE are responsible for an apparent decline in performance, we reran the classification using the rebuilt samples to train and classify the code. For this run, the f_1 score including all objects down to magnitude $r = 24.3$ is 0.91 for high-z quasars (with a confidence threshold of 0.30) and 0.47 for low-z quasars (with a confidence threshold of 0.29). This is significantly higher than our regular estimates and it is in agreement with the previous results from Pérez-Ràfols & Pieri (2020). Thus, the discrepancy found here can be accounted for by the better signal-to-noise ratio in the mocks used in the previous work (though we stress that the mocks we use here are more realistic).

6.2. Explainability of the classifiers

One of the main problems of using machine learning algorithms for classification problems is that often they are used as black
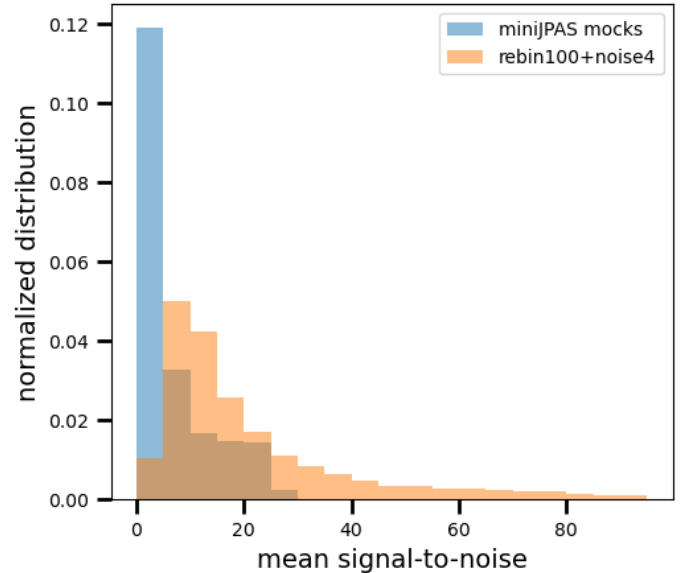

Fig. 10. Normalised distributions of the mean signal-to-noise ratio for the spectra of the test sample (blue) and the rebuilt test sample (orange). We rebuilt the test sample following the Pérez-Ràfols et al. (2020) prescription for building the rebin100+noise4 mocks.

boxes offering no explainability as to how the classification is being done. Because of the way SQUEzE is built, it provides some degree of explainability. Moreover, the results of this paper also offer some degree of explainability regarding the classifiers presented in the previous papers in the series. We analysed SQUEzE training to review this.

To explain SQUEzE behaviour, the most important thing is the coupling between classification and redshift estimation. The classification is done using a random forest classifier on a set of features. However, contrary to standard random forest usage, each spectrum can enter the random forest classifier multiple times. The key element here is that what we are classifying are not spectra, but trial redshifts. These are derived from the position of emission peaks found. This means that one of the key elements for SQUEzE to correctly identify the quasar spectra is its

ability to detect real emission lines. Indeed, as shown in Fig. 3, our ability to detect the emission lines declines with increasing magnitude leading to worse performance.

Having confirmed that only emission peaks drive the classification (which is not necessarily the case in Rodrigues et al. 2023 and Martínez-Solaeche et al. 2023), we now explore a feature importance analysis for each of the two random forest classifiers in SQUEzE in the four magnitude bins. This is performed by computing the mean (across the different trees in the forest) decrease in impurities when a particular feature is included or not. Higher values for the mean decrease indicate higher importance of the feature.

The results of this exercise are shown in Figs. 11 and 12. Section 3 describes three metrics for each of the lines (but see, in more detail, Eqs. (1) to (3) in Pérez-Ràfols et al. 2020). In terms of SQUEzE outputs, the amplitude of line X is labelled as X_LINE_RATIO, its significance as X_LINE_RATIO_SN, and the slope at the base of the line as X_LINE_RATIO2.

Generally speaking, the most important lines for high-z objects are Lyα, C IV, and C III], in this order. Their amplitudes are the most important characteristics, followed by the signal-to-noise ratio. This makes sense as these are the most prominent emission lines, and these are the lines a human visual inspector usually looks for. This is suggestive that SQUEzE is indeed in agreement with the visual inspection analysis by a human expert. We see that the red and blue halves of the C IV line are also relatively important. This line is most affected by the presence of broad absorption line (BAL) features, and therefore these lines are important for including BAL quasars in our sample.

At fainter magnitudes, the importance of these emission lines decreases in favour of the trial redshift and the magnitude. Because we are no longer able to distinguish real emission line peaks from noise peaks, it seems that SQUEzE is relying more on the magnitude and redshift distribution of the objects. This explains why adding these two columns helps to improve the results (see Appendix C.6). The use of these two columns is equivalent to having priors on their expected distributions. We warn the reader that this could potentially bias us towards the expected distributions and that we might be misled into overestimating the use of these parameters.

Similarly, for low-z objects, the most important lines are Mg II and C III], in this order. Interestingly, the extra bands that we added to avoid line confusion, in particular, LC3, LC4, and LC5 (see Table C.2) have a non-negligible importance. For instance, the mean decrease for these lines is similar to real lines such as the Hα or Hβ+O [III]. They are also more important than the Lyβ, Lyα, Si IV, and C IV lines, but this is expected as these are high-z lines.

This explains the improvement seen when changing the default set of lines (see Appendix C.3). As expected, when going fainter in magnitude we see a similar behaviour to that for high-z candidates.

6.3. Redshift precision

In this section, we discuss the estimated redshift precision of our catalogue and compare it with previous results from the literature. The typical redshift errors of our correctly classified quasars are discussed in Sect. 4, but there we analyse the entire sample, and not only the objects that would enter the quasar catalogue. Later, in Sect. 5, we explain how we build our quasar catalogue. We follow the same procedure to build catalogues for

Table 3. Normalised median absolute deviation, σ_{NMAD}, of the correct classifications for the test and 1 deg^2 test samples.

Sample	Name	σ_{NMAD} (per cent)	N
	All	0.92	2694
	High-z	0.99	1451
Test	Low-z	0.80	1243
	$r < 22.5$	0.81	2075
	$r < 21.3$	0.74	1289
	All	0.88	139
	High-z	0.88	71
1 deg^2 Test	Low-z	0.90	68
	$r < 22.5$	0.79	102
	$r < 21.3$	0.75	69

Notes. For each sample, we give the values for the entire sample, for high-z quasars, and low-z quasars only, and for two bright samples with $r < 22.5$ and $r < 21.3$, respectively (see text for details).

the test and 1 deg^2 test samples, where the redshift and the classification are known, to estimate the precision of our redshift estimates. We quantify this precision in terms of σ_{NMAD} (see Eq. (2)).

We computed σ_{NMAD} considering only correct classifications for the entire sample, and we did this for high-z quasars and low-z quasars separately. We also computed σ_{NMAD} for two bright subsamples. For the first one, we considered $r < 22.5$, i.e. including bins 1 and 2, where the emission lines are clearly detected (see Sect. 4). For the second bright sample, we considered quasars with $r < 21.3$, where we expect a much higher level of purity (in particular >0.90 for high-z quasars).

Table 3 summarises the results of this exercise. As expected, The results are similar in the test and 1 deg^2 test samples. In particular, the dispersion is lower for low-z quasars and also for brighter objects. Overall, our results are below the percentage level. This is an order of magnitude better than the reported values of $\sigma_{\mathrm{NMAD}} = 9\%$ by Matute et al. (2012), who analysed a similar sample of quasars in the ALHAMBRA survey down to magnitude $r = 24$. These results are comparable to the findings by Chaves-Montero et al. (2017), where they get $\sigma_{\mathrm{NMAD}} = 1.15\%$ and $\sigma_{\mathrm{NMAD}} = 0.91\%$ for their AGN-X sample in the 2-line and 3-line detection mode, respectively. For their AGN-S sample, they find $\sigma_{\mathrm{NMAD}} = 1.01\%$ and $\sigma_{\mathrm{NMAD}} = 0.86\%$ for the 2-line and 3-line detection modes. However, we note that their quasars have magnitudes $F814W < 22.5$ for the AGN-S sample and $F814W < 23$ for the AGN-X sample. While there is no direct comparison between r-band magnitudes and $F814W$ magnitudes, they have generally brighter quasars. If we compare their results to our brighter samples, we see that we recover σ_{NMAD} values that are ~5−20% lower (depending on the exact compared samples).

7. Summary and conclusions

In this work, we analysed miniJPAS data using SQUEzE. We present the particularities of applying SQUEzE to this dataset and a catalogue of quasar candidates. Following previous papers in this series (Rodrigues et al. 2023; Martínez-Solaeche et al. 2023), we trained the models on the miniJPAS mocks developed by Queiroz et al. (2023) for this purpose. We tested the performance on three different datasets, two of them synthetic
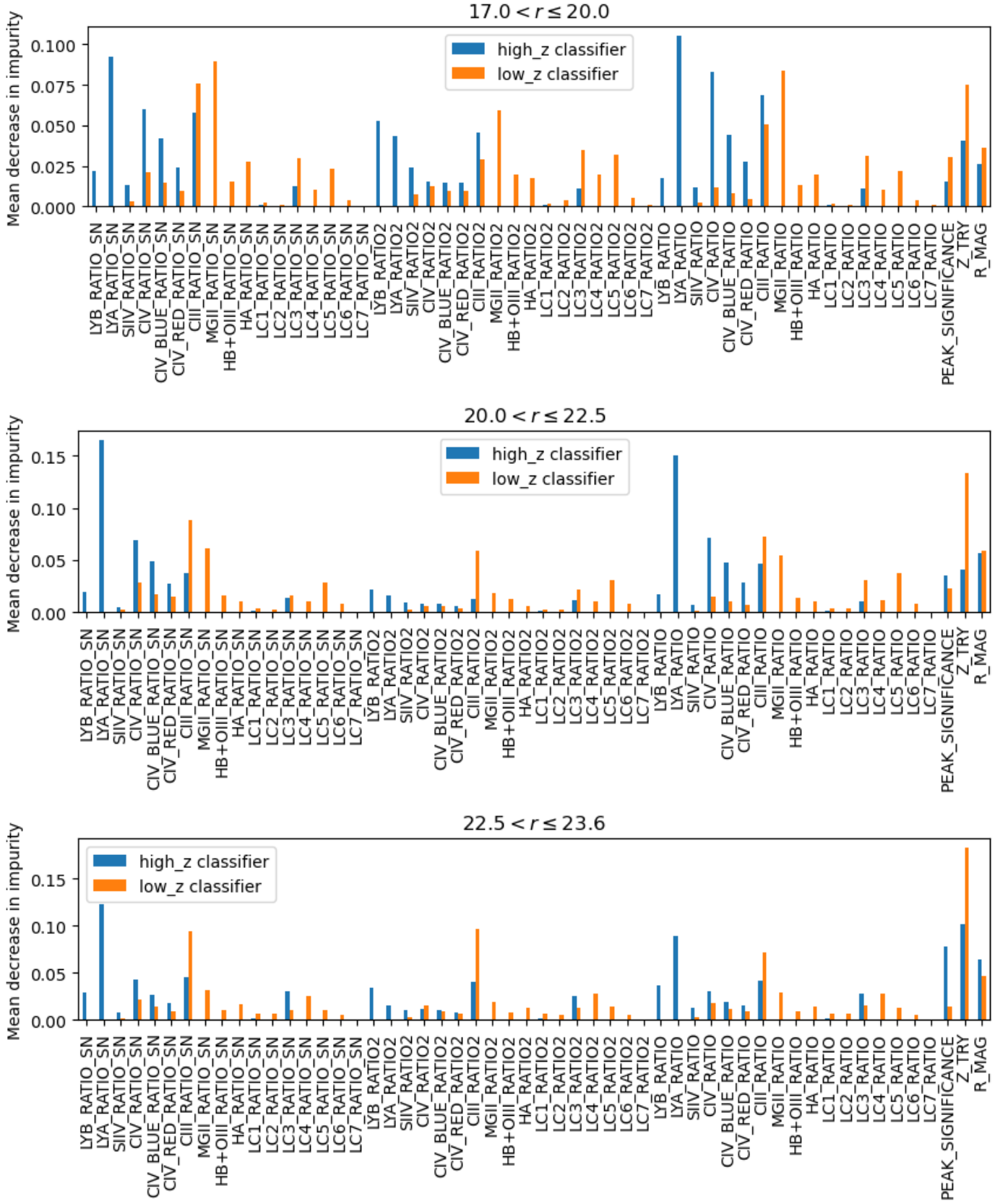
Fig. 11. Feature importance analysis performed on SQUEzE training based on the mean decrease in impurity. Higher values indicate that the feature is more important. From top to bottom, the different panels show the result for the first three magnitude bins, with $17 < r \leq 20$, $20 < r \leq 22.5$, and t$22.5 < r \leq 23.6$. The results for the remaining magnitude bin are shown in Fig. 12.
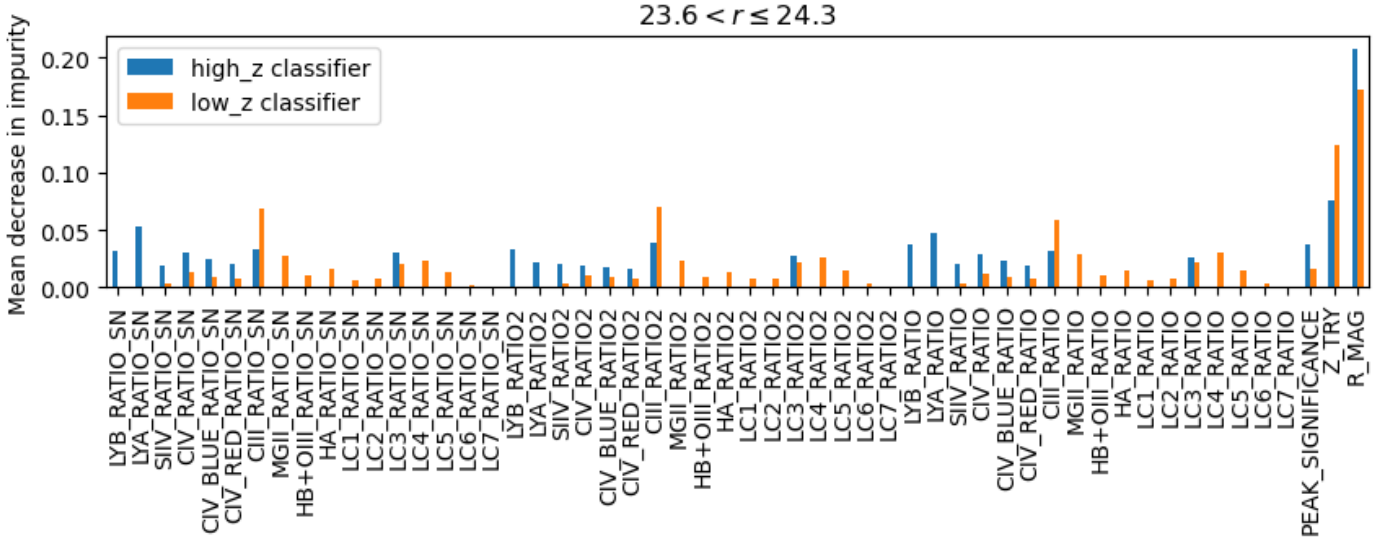
Fig. 12. Same as Fig. 11, but for the fourth magnitude bin, with $23.6 < r \leq 24.3$.

and one of them using the relatively small subset of miniJPAS data with spectroscopic counterparts from SDSS. Finally, we compared our results to previous estimates of SQUEzE performance, attempted to explain the reasoning behind SQUEzE, assessed the impact of using different noise models to build the mocks, and evaluated the redshift precision of our samples. Our main conclusions are as follows.

– Our results in the test samples suggest that the f_1 score including all objects down to $r = 24.3$ is 0.49 for high-z quasars and 0.24 for low-z quasars. For high-z quasars, this is increased to 0.9 for magnitudes of $r < 21.0$.

– While SQUEzE performance is lower than some of the other classifiers of the series, it provides us with redshift estimates.

– We assesed our redshift precision using the normalised median absolute deviation, σ_{NMAD}. For our test sample, we reach a value of 0.92%, an order of magnitude better than similar samples in the literature. For brighter samples, this decreases further to 0.81% ($r < 22.5$) and 0.74% ($r < 21.3$).

– Contrary to other machine-learning classifiers, the SQUEzE decisions can be explained: as we go fainter in magnitude SQUEzE is no longer able to distinguish real emission lines from noise peaks and more weight is given to the magnitude and redshift distributions.

– It is possible that SQUEzE is able to run on extended sources with similar levels of performance even if the training set only characterises point-like sources. This could imply that the photometric properties of extended and point-like quasars are similar or that the criteria used to split between extended and point-like sources occasionally fail.

– Changing the noise model used to create the mocks has an impact mostly at the faint end and sometimes results in lower redshift estimates.

– We computed a catalogue of quasar candidates for both point-like sources, with 301 candidates, and also including extended sources, with 1049 candidates. While extended sources are not included in our mocks, the comparison of the magnitude and redshift distributions of both catalogues suggests that SQUEzE could show a similar performance level on extended objects compared to point-like objects.

– A spectroscopic follow-up of a large number of objects is crucial to verify the results of this work and could lead to improvements in the classifiers.

Summing up, we found that SQUEzE can complement the other classifiers presented in this series. Even if it has slightly lower performance levels than some of the other classifiers, it provides us with redshift estimation. This is crucial for many science cases. We remark that SQUEzE might work when presented with extended sources, but a spectroscopic follow-up is needed to verify our findings.

References

Abramo, L. R., Strauss, M. A., Lima, M., et al. 2012, MNRAS, 423, 3251
Alam, S., Aubert, M., Avila, S., et al. 2021, Phys. Rev. D, 103, 083533
Baqui, P. O., Marra, V., Casarini, L., et al. 2021, A&A, 645, A87
Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv e-prints [arXiv:1403.5237]
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Bonoli, S., Marín-Franch, A., Varela, J., et al. 2021, A&A, 653, A31
Busca, N. G., Delubac, T., Rich, J., et al. 2013, A&A, 552, A96
Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2019, A&A, 622, A176
Chaves-Montero, J., Bonoli, S., Salvato, M., et al. 2017, MNRAS, 472, 2085
Chaves-Montero, J., Bonoli, S., Trakhtenbrot, B., et al. 2022, A&A, 660, A95
Dalton, G., Trager, S., Abrams, D. C., et al. 2016, Proc. SPIE, 9908, 99081G
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44
DESI Collaboration (Aghamousa, A., et al.) 2016a, arXiv e-prints [arXiv:1611.00036]
DESI Collaboration (Aghamousa, A., et al.) 2016b, ArXiv e-prints [arXiv:1611.00037]
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
du Mas des Bourboux, H., Rich, J., Font-Ribera, A., et al. 2020, ApJ, 901, 153
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Font-Ribera, A., Arnau, E., Miralda-Escudé, J., et al. 2013, JCAP, 2013, 018
Hoaglin, D. C., Mosteller, F., & Tukey, J. W. 1983, in Wiley Series in Probability and Mathematical Statistics (New York: Wiley)
Hou, J., Sánchez, A. G., Ross, A. J., et al. 2021, MNRAS, 500, 1201
Kirkby, D., Margala, D., Slosar, A., et al. 2013, JCAP, 2013, 024
Leistedt, B., Peiris, H. V., & Roth, N. 2014, Phys. Rev. Lett., 113, 221301
López-Sanjuan, C., Varela, J., Cristóbal-Hornillos, D., et al. 2019, A&A, 631, A119
Martínez-Solaeche, G., Queiroz, C., González Delgado, R. M., et al. 2023, A&A, 673, A103
Matute, I., Márquez, I., Masegosa, J., et al. 2012, A&A, 542, A20
Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., et al. 2019, MNRAS, 489, 241
Neveux, R., Burtin, E., de Mattia, A., et al. 2020, MNRAS, 499, 210
Pérez-Ràfols, I., & Pieri, M. M. 2020, MNRAS, 496, 4941
Pérez-Ràfols, I., Pieri, M. M., Blomqvist, M., Morrison, S., & Som, D. 2020, MNRAS, 496, 4931
Pieri, M. M., Bonoli, S., Chaves-Montero, J., et al. 2016, in SF2A-2016: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, 259
Queiroz, C., Abramo, L. R., Rodrigues, N. V. N., et al. 2023, MNRAS, 520, 3476
Rodrigues, N. V. N., Raul Abramo, L., Queiroz, C., et al. 2023, MNRAS, 520, 3494
Salvato, M., Hasinger, G., Ilbert, O., et al. 2009, ApJ, 690, 1250
Salvato, M., Ilbert, O., Hasinger, G., et al. 2011, ApJ, 742, 61
Slosar, A., Iršič, V., Kirkby, D., et al. 2013, JCAP, 2013, 026
Spinoso, D., Orsi, A., López-Sanjuan, C., et al. 2020, A&A, 643, A149
Wolf, C., Wisotzki, L., Borch, A., et al. 2003, A&A, 408, 499

[1] Departament de Física Quàntica i Astrofísica. Institut de Ciències del Cosmos, Universitat de Barcelona (UB-IEEC), Martí i Franquès 1, 08028 Barcelona, Catalonia, Spain
e-mail: iprafols@fqa.ub.edu

[2] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Edifici Cn, Campus UAB, 08193 Bellaterra (Barcelona), Spain

[3] Institut d'Estudis Espacials de Catalunya (IEEC), Gran Capità, 2-4 Edifici Nexus, 08034 Barcelona, Spain

[4] Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, Rua do Matão 1371, CEP 05508-090, São Paulo, Brazil

[5] Instituto de Astrofísica de Andalucía (CSIC), P.O. Box 3004, 18080 Granada, Spain

[6] Aix-Marseille Univ., CNRS, CNES, LAM, 38 rue Frédéric Joliot-Curie, 13388 Marseille, France

[7] Donostia International Physics Center, Paseo Manuel de Lardizabal 4, 20018 Donostia-San Sebastian, Spain

[8] Ikerbasque, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Spain

[9] Department of Astronomy, University of Illinois at Urbana Champaign, 1002 W Green St, Urbana, IL 61801, USA

[10] Observatório Nacional, Rua General José Cristino 77, São Cristóvão, 20921-400 Rio de Janeiro, Brazil

[11] Instituto de Astrofísica de Andalucía (CSIC), Gta. de la Astronomía s/n, 18008 Granada, Spain

[12] Instituto de Física, Universidade Federal da Bahia, Campus Universitário de Ondina - Ondina, Salvador 40210-340, Brasil

[13] Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Unidad Asociada al CSIC, Plaza San Juan 1, 44001 Teruel, Spain

[14] Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza San Juan 1, 44001 Teruel, Spain

[15] Núcleo de Astrofísica e Cosmologia & Departamento de Física, Universidade Federal do Espírito Santo, Av. Fernando Ferrari 514, 29075-910 Vitória, Brazil

[16] INAF – Osservatorio Astronomico di Trieste, Via Tiepolo 11, 34131 Trieste, Italy

[17] IFPU – Institute for Fundamental Physics of the Universe, Via Beirut 2, 34151 Trieste, Italy

[18] Departamento de Astronomia, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Rua do Matão 1226, Butantã, São Paulo 05508-090, Brasil

[19] Instruments4, 4121 Pembury Place, La Canada Flintridge, CA 91011, USA

Appendix A: Effect of the noise model in mocks

As mentioned in Section 2, the mock sets used are described in detail in Queiroz et al. (2023). In particular, we used the noise model 11, which is precisely the one closest to the observed data. Here, we explore the impact of using a different noise model. We trained SQUEzE using the second-best noise model (model 1) to assess the performance on the respective test sample (computed also using the alternative noise model).

Figure A.1 shows the change in f_1 score when the different noise models are used. The alternative noise model performs slightly better. Including objects at all magnitudes, there is an increase in the performance of 0.08 for high-z quasars and 0.04 for low-z quasars. However, the noise model 1 is simpler than the 11th, and the test sample is also rerun for each noise model. Because noise model 1 is simpler, it is not unexpected that the performance is slightly better, as the classifiers have to learn a simpler distribution. However, we stress that noise model 11 is closer to the actual measurement of the noise distribution and thus should be more realistic (hence our choosing it as our fiducial model).

Perhaps more interesting is the impact of using these noise models on real data. Table A.1 shows the number of candidates recovered when using the different noise models to train the

Table A.1. Number of quasar candidates in different noise models.

noise model	point-like candidates	all candidates
11	301	1049
1	419	1514

Number of quasar candidates computed using different noise models in the mocks to train our classifier. Noise model 11 is the closest to the observed data (see Queiroz et al. 2023).

classifiers. Using model 11 results in a smaller number of candidates (by ∼ 30%). A similar decrease is observed for both the point-like and the entire samples. We explored this difference further for the point-like sample by analysing the distributions of redshift and magnitude (Figure A.2). The magnitude distributions are essentially the same, but we observe some small differences in the redshift distribution. Model 1 seems to slightly favour larger redshifts. Clearly, one of the models has better redshift precision. One would tend to think that model 11, being a more realistic noise model, would produce better redshifts, but a spectroscopic follow-up of the objects is required to know for certain.
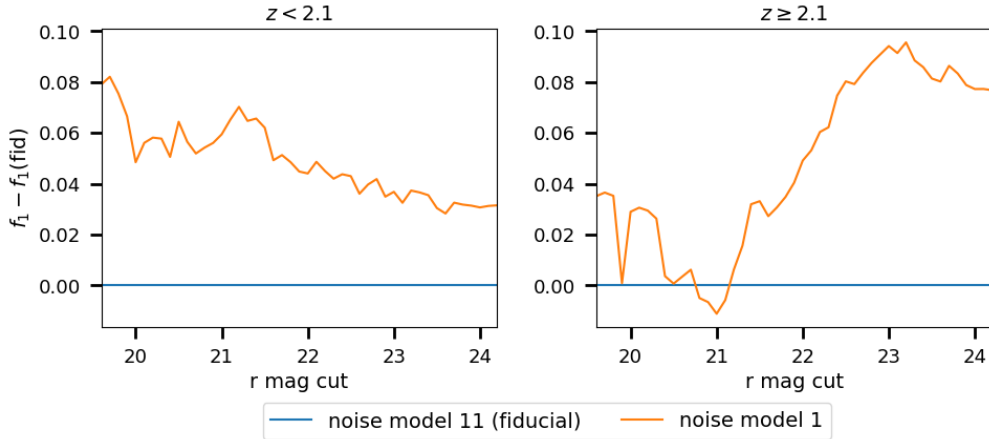

Fig. A.1. Change in the f_1 score when different noise models are used to generate the mocks. Note that the change is applied both to the train and the test samples. The left (right) panel shows the performance for the low (high) redshift quasars.
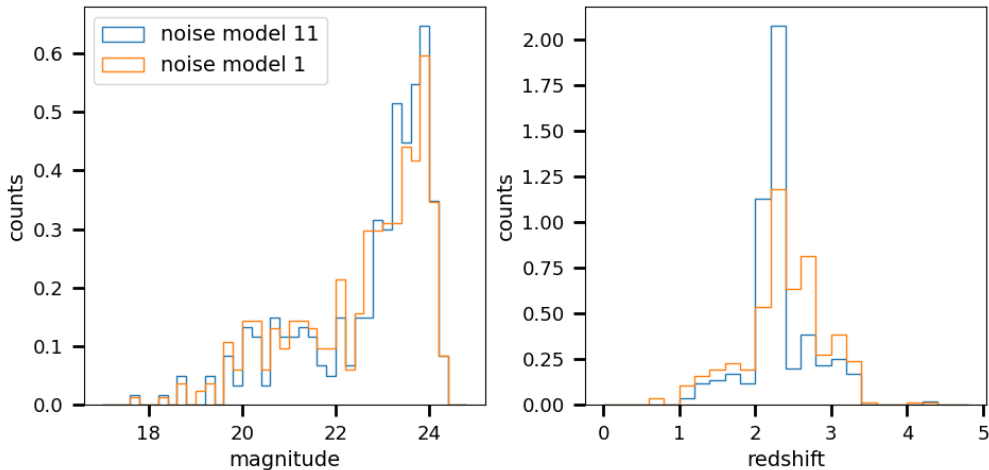

Fig. A.2. Distribution of magnitude (left) and redshift (right) for the quasar candidates computed using different noise models in the mocks to train our classifier. Noise model 11 is the closest to the observed data (see Queiroz et al. 2023).

Appendix B: Performance of the peak finders

We stress that one of the key elements of SQUEzE is the peak finder. In this work, we changed the original peak finder of SQUEzE by a new peak finder (see Section 3.1). Here, we discuss the performance of the peak finders.

We start by performing a qualitative assessment based on the performance of a few bright objects (see Figure B.1). The first example shows the performance of both the new peak finder and the original one on a synthetic spectrum of a quasar. We can see that the new peak finder provides a smaller number of peaks and that they are at the expected positions. The other two examples show the performance of synthetic spectra of a galaxy and of a star, where the assumption of a power-law continuum does not necessarily hold here. Nevertheless, the number of noise peaks is significantly reduced.

These examples correspond to a relatively bright spectrum, and the peak identification is less successful on fainter objects as the noise is louder. Nevertheless, we find that the new peak finder better filters the noise peaks.

To estimate the performance of the peak finders we used three quantities. First, we considered the completeness after the peak finder step. As mentioned above, quasars for which we fail to detect the correct peak here will not be recovered at a later stage. Thus, we want this quantity to be as high as possible. Apart from completeness, it is also important to consider the number of correctly identified peaks and the total number of peaks.

Figure B.2 shows the result of this exercise. The completeness level of the new peak finder is lower than that of the original one. For high-z quasars, the decrease is 0.014 at magnitudes of $r < 22.1$, where the original peak finder has a completeness of one. As we go fainter in magnitude, the difference increases up to 0.14 at magnitudes of $r < 24.3$, where the original peak finder completeness level remains at 0.99. For low-z quasars, we see a similar trend albeit with a larger decrease: 0.043 at magnitudes of $r < 22.1$ and 0.23 at magnitudes of $r < 24.3$.

This decrease in completeness is compensated by a drastic reduction in the number of incorrect peaks when the new peak finder is used. We see a decrease of a factor of ∼ 3 for high-z quasars and a factor of ∼ 2 for low-z quasars. At the same time, the number of correct peaks per spectrum stays roughly constant. This means that the random forest algorithms will have an easier job finding the correct entries. Indeed, we see an increase in performance when using the new peak finder (see Figure B.1). We conclude that this decrease in the number of incorrect peaks more than compensates for the decrease in completeness.
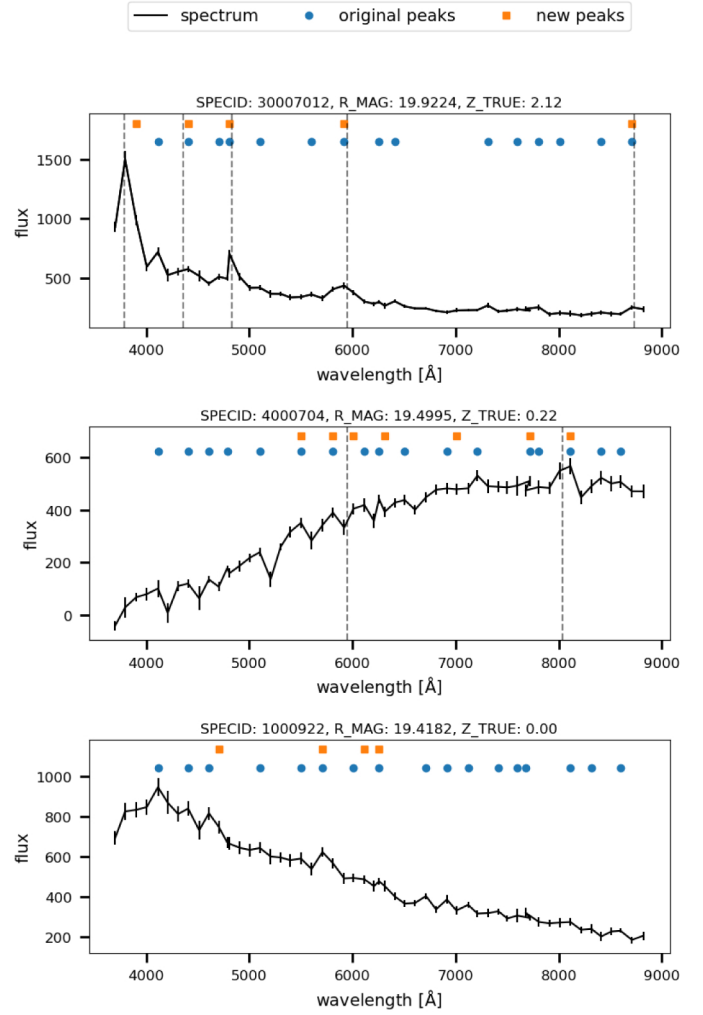

Fig. B.1. Example of performance of new peak finder compared to the original one. To illustrate the difference between peak finders, we show a quasar with magnitude $r = 19.9224$ and redshift of $z = 2.12$ (top panel), a galaxy with magnitude $r = 19.4995$ and redshift of $z = 0.07$ (mid panel), and a star with magnitude $r = 19.4182$. Blue circles show the peaks as detected by the original peak finder. Orange squares are the peaks detected by the new peak finder. Dashed lines indicate the expected position of the main emission lines. For the quasar, from left to right, they indicate the Lyα, Si IV, C IV, C III], and Mg II. For the galaxy, also from left to right, they indicate the O [III] and Hα.
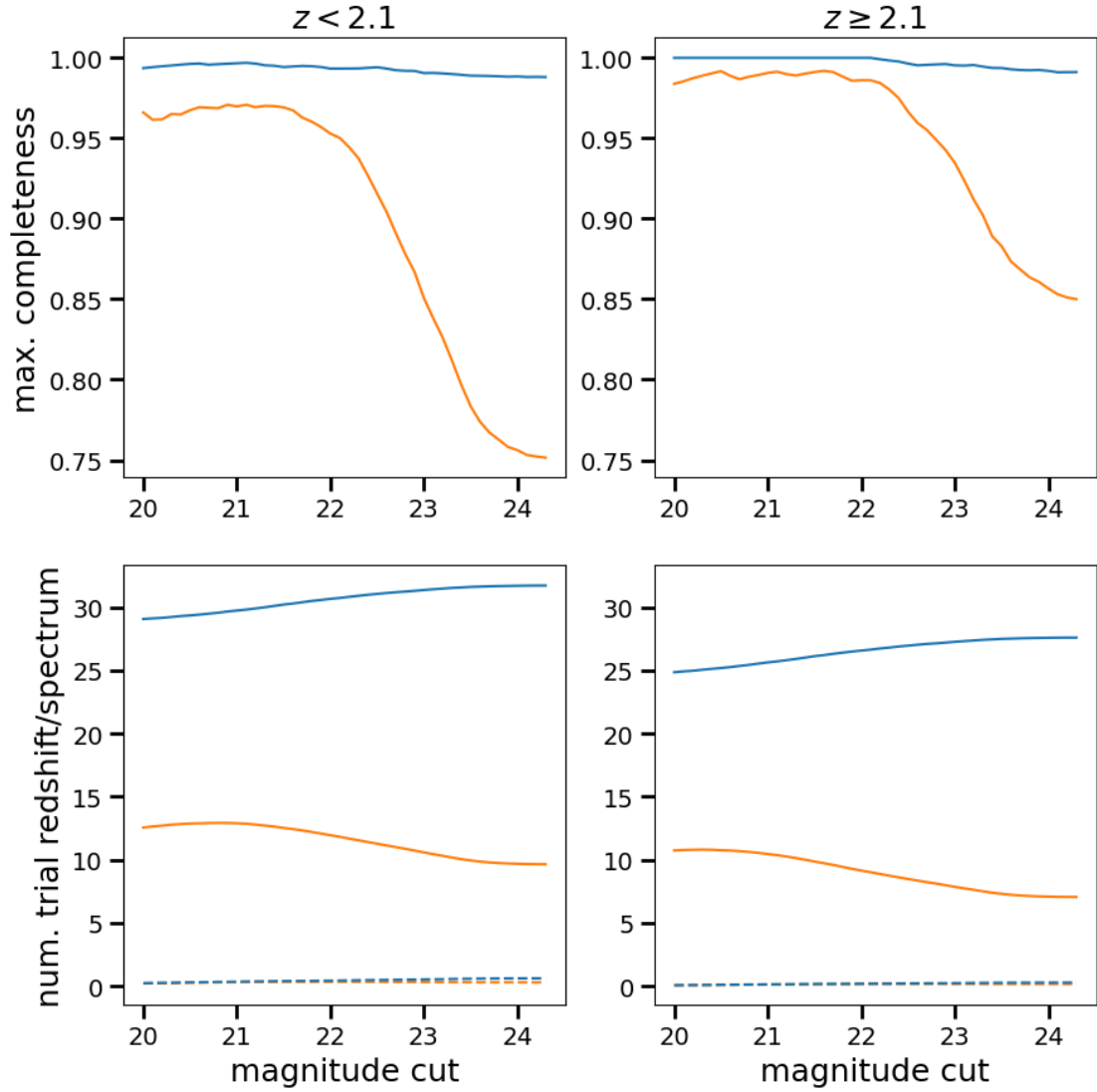
Fig. B.2. Top panels show level of completeness after peak finder step. The bottom panels show the number of peaks per spectrum as solid lines and the number of correct peaks per spectrum as dashed lines. To compute the latter, only spectra of quasars are counted, whereas to compute the former all spectra are considered. The results for the old (new) peak finder are shown via blue (orange) lines.
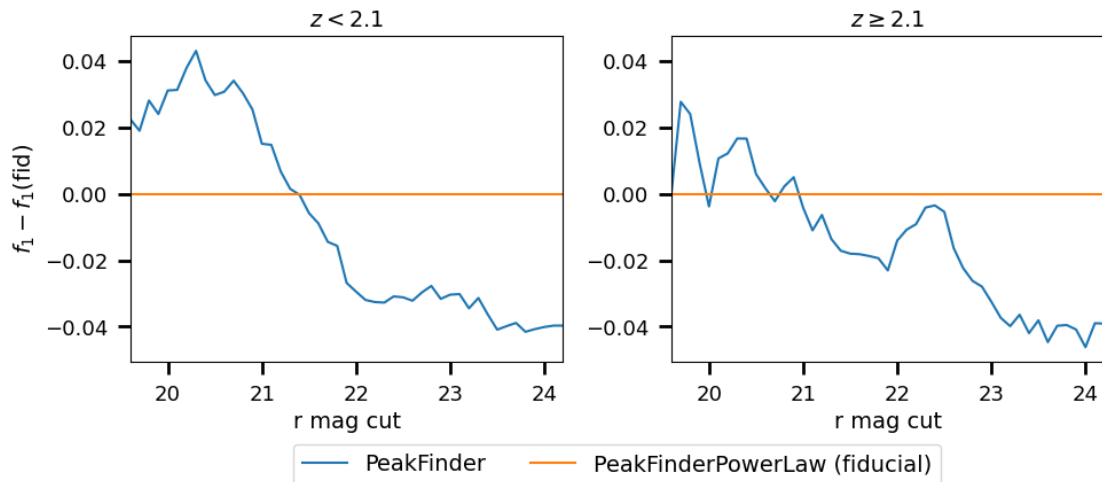

Fig. B.3. Changes in performance (f_1 score) of SQUEzE when training with the different peak finders. The orange line shows the performance of the original peak finder, `PeakFinder`, compared to the peak finder developed here, `PeakFinderPowerLaw`, our fiducial choice. The left (right) panel shows the performance for the low (high) redshift quasars. For the original peak finder, we use the following parameters: four magnitude bins, no smoothing, a minimum significance of zero, using the default set of lines, using two random forests, and adding columns $z_{\rm try}$ and magnitude r. These parameters are selected following the approach described in Appendix C for the new peak finder.

Appendix C: Robustness of the chosen `SQUEzE` parameters

Section 4 shows the results of SQUEzE on the miniJPAS mocks. Then, in Section 6.1 we discuss the reasons behind the decrease in the performance compared to the previous estimates from Pérez-Ràfols & Pieri (2020). The performance decrease is attributed to the data being noisier than assumed by the rough estimates of Pérez-Ràfols & Pieri (2020). In the subsequent subsections, we review the main choices for the different parameters and conclude that only minor tweaks to SQUEzE parameters are required in order to achieve the optimal configuration and that this configuration is only marginally better than the default choices. This supports the universality of SQUEzE models stated in Pérez-Ràfols & Pieri (2020).

Throughout this section, we use the validation set to justify the different choices we made. We change one parameter at a time to evaluate the effect of changing this parameter, fixing the rest to our fiducial choice, as described in Section 3. Comparisons are made using the f_1 score as defined in Equation 1. We checked that in all cases the purity and completeness remain roughly stable, i.e. that the f_1 score does not increase with a significantly higher level of purity at the cost of lower completeness level or vice-versa. However, for clarity, we only show the f_1 score here.

C.1. Training in magnitude bins

In Section 3.3 we classify objects separately based on their r-band magnitude. Here, we explore the reasons behind this choice. The main argument to justify splitting the sample is that brighter objects have higher signal-to-noise ratios and thus emission lines are easier to detect. On top of this, faint objects are substantially more numerous, thus dominating the training set. It is therefore reasonable to think that the random forest could learn to identify the low signal-to-noise quasars better and lower the performance at the bright end.

To test if this is indeed the case, we ran SQUEzE in three scenarios. First, we took all the objects in a single magnitude bin, $r \in (17.0, 24.3]$. Second, we split the bin in two: $r \in (17.0, 22.5]$ and $r \in (22.5, 24.3]$. Third, we further split each of the bins: $r \in (17.0, 20.0]$, $r \in (20.0, 22.5]$, $r \in (22.5, 23.6]$, and $r \in (23.6, 24.3]$. We trained SQUEzE in each of these magnitude bins and combined the results into the larger single bin to compare.

Results of this exercise are shown in Figure C.1, where we compare the performance of the models with one and two bins to that of the fiducial model with four bins. When objects at all magnitudes are included, the performance drops for low-z quasars by ~ 0.015 and ~ 0.005 when using one or two bins, respectively. When cutting at different limiting magnitudes, we find the performance to be generally higher in the four-bin model by ~ 0.02 and ~ 0.01 for low-z and high-z quasars, respectively. An even finer magnitude split might yield even better results, but larger amounts of data would be necessary, so we leave this for future studies.

C.2. Significance threshold for peak finding

In Section 3.1 we explain that peaks are identified by selecting outliers with a minimum significance from a power-law fit of the spectrum continua. Here, we justify the choice of $N = 2$ in our standard runs.

Table C.1. Intervals added to help with identified line confusions.

Real line	Assumed line	Interval(s) added
Lyα	C IV	LC1 (Mg II)
C IV	C III]	LC1 (Mg II), LC5 (Hα)
C IV	Lyα	LC2 (C III]), LC7 (Hα)
C III]	C IV	LC2 (C III]), LC6 (Hα)
C IV	Mg II	LC4 (Hα)
Mg II	Hβ	LC4 (Hα)
Lyα	Mg II	-
C III]	Hβ	LC3 (Hα)

The first column gives the real line and the second column gives the assumed line. The third column gives the intervals added to help remove the line confusion. In parentheses, we show the line whose peak is expected to appear in the interval if the assumed line is correct.

We compare the performance when taking different cuts in the peak significance. We explore a cut of 1.5 to 3.0 (both included) in steps of 0.5. We compare the performance of those models against the fiducial model, with a significance cut of 2.0. Results of this exercise are shown in Figure C.2. The performance using different significant cuts is showing some fluctuations around a 0.01 change in the performance level. The fiducial choice seems to be marginally better than the other cases studied.

C.3. Optimisation of the line bands

As shown in Section 3.2, for each of the trial redshifts we computed a set of line metrics based on the predicted position of the emission lines of interest. The line bands used on the main results were optimised for BOSS data and our j-spectra have a significantly different resolution. We tested if the same line bands should be used here, as seems to be indicated by the results in Pérez-Ràfols & Pieri (2020).

We changed the line bands with the following criteria. First, we removed the weak emission lines that cannot be resolved in the mean spectrum of miniJPAS quasars (Martínez Uceta et al. In prep.). Then, we expanded the bands to include more than one filter. This is important as there are non-detections in some of the filters, more so at the faint end. Having more than one filter in each of the bands allowed us to measure the line metrics even when a few of these faulty measurements are present in the spectra. We label this set of lines as 'wide'.

On top of widening the bands, we also added a few new 'emission lines'. These are added in intervals where we do not expect an emission line, but where we would expect it if we had line confusion (see Table C.1). With these, we not only increase the number of filters used to classify the j-spectra, but we also could potentially improve the existing line confusion. We label this set of lines as 'wide+extra'.

The chosen set of line bands is given in Table C.2. To compare with the number of filters used in our fiducial choice (Figure 1) we give this quantity for the default bands (left panel of Figure C.3) and when we only use the wider bands (right panel of Figure C.3). In both the 'wide' and the 'wide+extra' line bands the number of filters used is higher than in the default case (see Figure 1).

We test the performance of the new sets of lines (using only the wider bands and using both the wider and extra bands) and
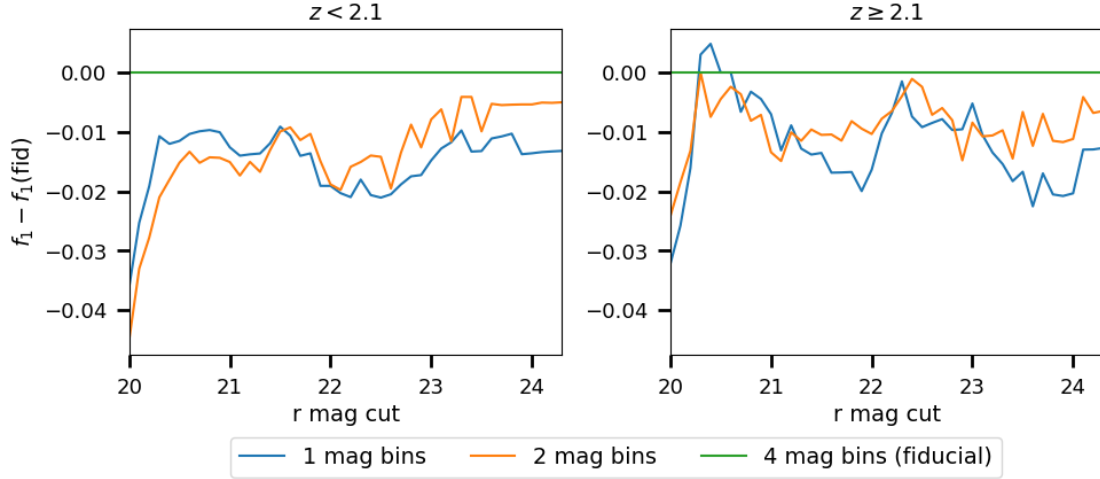
Fig. C.1. Changes in SQUEzE performance (f_1 score) when training with the different number of magnitude bins. The magnitude bins are $r \in (17.0, 24.3]$ for the model with one bin (blue lines), $r \in (17.0, 22.5]$ and $r \in (22.5, 24.3]$ for the model with two bins (orange lines), and $r \in (17.0, 20.0]$, $r \in (20.0, 22.5]$, $r \in (22.5, 23.6]$, and $r \in (23.6, 24.3]$ for the model with four bins (green lines). The left (right) panel shows the performance for the low (high) redshift quasars. The fiducial model, with four bins, is the best performing.
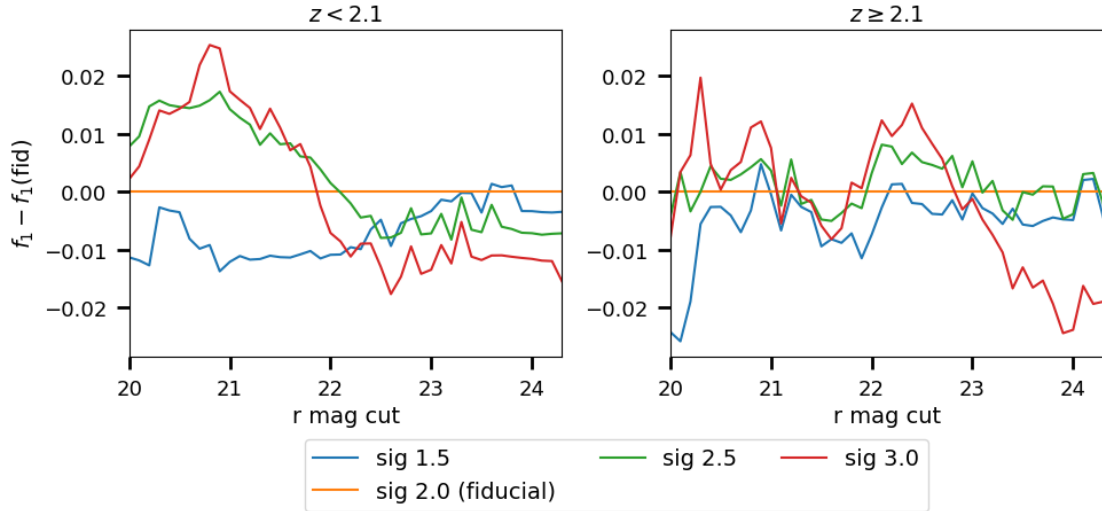

Fig. C.2. Comparison of SQUEzE performance (f_1 score) when the minimum peak significance is 1.5 (blue lines), 2.0 (orange lines), 2.5 (green lines), and 3.0 (red lines). The left (right) panel show the performance change for the low (high) redshift quasars. We see fluctuations in the performance at the percentage level.
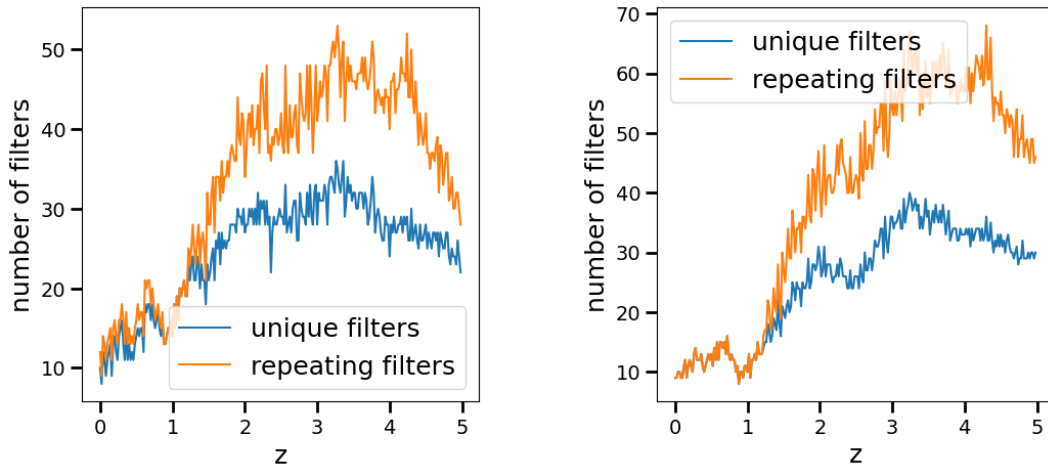

Fig. C.3. Same as Figure 1, but using the default bands (left panel) and using only the wider bands (right panel). Both cases use a smaller number of filters compared to the fiducial set of lines (Figure 1).

Table C.2. Alternative line bands used by SQUEzE to compute the metrics.

LINE	WAVE	START	END	BLUE_START	BLUE_END	RED_START	RED_END
Lyβ	1,033.03	1,000.00	1,080.00	890.00	990.00	1,103.00	1,159.00
Lyα	1,215.67	1,190.00	1,260.00	1,103.00	1,159.00	1,295.00	1,350.00
Si IV	1,396.76	1,370.00	1,417.00	1,295.00	1,350.00	1,432.00	1,494.50
C IV	1,549.06	1,515.00	1,575.00	1,432.00	1,494.50	1,603.00	1,668.00
C IV blue	1,549.06	1,515.00	1,549.06	1,432.00	1,494.50	1,603.00	1,668.00
C IV red	1,549.06	1,549.06	1,575.00	1,432.00	1,494.50	1,603.00	1,668.00
C III]	1,908.73	1,865.00	1,940.00	1,756.00	1,835.00	1,970.00	2,060.00
Mg II	2,798.75	2,720.00	2,816.00	2,570.00	2,690.00	2,851.00	2,984.00
Hβ+O [III]	4,862.68	4,800.00	5,020.00	4,640.00	4,740.00	5,060.00	5,145.00
Hα	6,564.61	6,480.00	6,650.00	6,320.00	6,460.00	6,750.00	6,850.00
LC1	2,233.88	2,140.00	2,315.00	2,060.00	2,110.00	2,335.00	2,385.00
LC2	2,392.05	2,280.00	2,470.00	2,200.00	2,260.00	2,490.00	2,550.00
LC3	2,576.78	2,520.00	2,630.00	2,420.00	2,500.00	2,640.00	2,710.00
LC4	3,705.86	3,600.00	3,790.00	3,510.00	3,580.00	3,800.00	3,900.00
LC5	5,327.61	5,280.00	5,400.00	5,160.00	5,250.00	5,480.00	5,570.00
LC6	8,088.82	7,980.00	8,230.00	7,870.00	7,960.00	8,280.00	8,550.00
LC7	8,364.91	8,280.00	8,550.00	8,120.00	8,260.00	8,650.00	8,790.00

In the first block, intervals of major lines have been widened to ensure continuous coverage at a given range of redshift. Minor lines (e.g. NeIV and Ne V) have been removed. OIII and Hβ lines have been merged into a single wider line. In the second block, additional line intervals have been added to help deal with identified line confusion (see Table C.1). All quantities are given in Å.
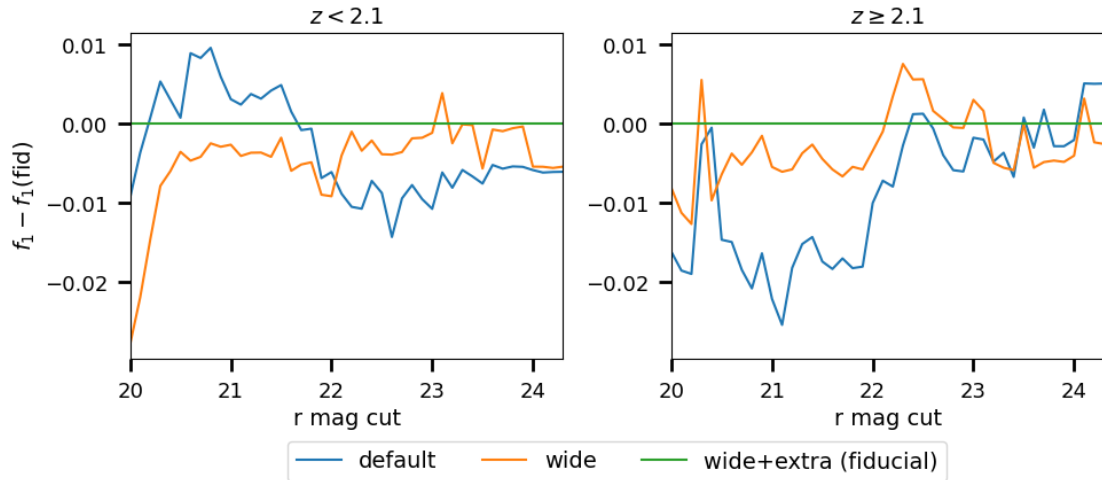

Fig. C.4. Comparison of SQUEzE performance (f_1 score) when using the default set of line bands and the two new sets of line bands (using wider bands and using both wider bands and some extra bands), specified in Table C.2. Both cases are computed with a minimum peak significance of 2. The left (right) panel shows the performance for the low (high) redshift quasars. The fiducial model, using the wide+extra line bands is the best performing.

compare it to the default lines. This comparison is shown in Figure C.4. The 'wide+extra' lines are superior, but only marginally, with an increase of the order of 0.01-0.02, seen mostly for bright magnitudes. This fact supports the predictions by Pérez-Ràfols & Pieri (2020) on the universality of their model, where only marginal improvements are expected when changing the parameters of the model.

C.4. Single classifier versus redshift split classifier

In Pérez-Ràfols et al. (2020), they argue that using two different random forest classifiers, one for high redshift, z_{try} and one for low redshift, z_{try}, they obtain a better level of performance as opposed to using a single classifier. Since our datasets have significantly different properties (in particular

magnitude and resolution), this statement does not necessarily hold here. In order to check this, we redid our analysis using a single random forest classifier. The options passed to SQUEzE are {"criterion": "entropy", "max_depth": 10, "n_jobs": 3, "n_estimators": 1000}. We compare the results from this new run with our fiducial in Figure C.5. There is a consistent increase of $\sim 0.03 - 0.04$ in the performance at low redshift when using two random forest classifiers. For high-redshift quasars, the improvement in the performance is smaller, reaching only ~ 0.02. Thus, we decided to stick with using two random forest classifiers.
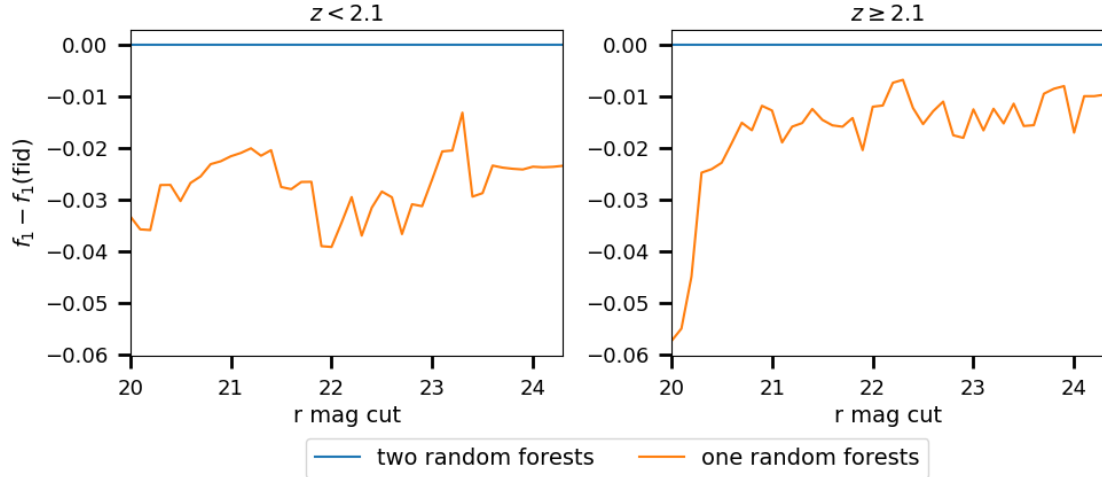
Fig. C.5. Comparison of SQUEzE performance (f_1 score) when using one or two random forest classifiers. When two are used, the first one is charged with classifying the low-z quasars and the second one, the high-z quasars. The left (right) panel shows the performance for the low (high) redshift quasars. The fiducial model, using two random forests, is the best performing.

Table C.3. Statistics of the redshift precision for different redshift precision requirements.

Δz	mag bin	$\overline{\Delta v}$ (km/s)	$\sigma_{\Delta v}$ (km/s)	σ_{NMAD} (per cent)	N
	$17.0 < r \leq 20.0$	224.30	4,282.73	0.77	520
	$20.0 < r \leq 22.5$	968.87	3,834.35	0.99	2209
0.15	$22.5 < r \leq 23.6$	1,349.70	6,016.32	1.66	1061
	$23.6 < r \leq 24.3$	1,603.68	6,355.87	1.86	210
	$17.0 < r \leq 20.0$	457.03	2,870.53	0.74	517
	$20.0 < r \leq 22.5$	711.28	3,121.52	0.95	2198
0.10	$22.5 < r \leq 23.6$	1,015.71	4,209.44	1.42	957
	$23.6 < r \leq 24.3$	889.00	4,802.76	1.55	184
	$17.0 < r \leq 20.0$	463.10	2,188.32	0.69	491
	$20.0 < r \leq 22.5$	461.34	2,274.05	0.79	1998
0.05	$22.5 < r \leq 23.6$	352.09	2,642.76	1.12	747
	$23.6 < r \leq 24.3$	5.51	2,624.00	1.10	125

Mean redshift offset and dispersion for the different magnitude bins. The mean offset indicates potential biases of our redshift estimate and the dispersion indicates our typical redshift error. Each block shows the results when running with different values for Δz. We note that our fiducial model has a significance level of $\Delta z = 0.15$.

C.5. Impact of redshift tolerance

Section 4 shows that as we go fainter in magnitude, the redshift precision decreases. This suggests that the redshift tolerance used could be too large. On the other hand, Pérez-Ràfols & Pieri (2020) used an even larger redshift tolerance (0.15). Here, we discuss the possible effect of using different cuts on the redshift tolerance. We ran SQUEzE using $\Delta z = \left| z_{\mathrm{true}} - z_{\mathrm{try}} \right| < 0.15$, $\Delta z < 0.10$ (our fiducial choice) and $\Delta z < 0.05$.

Table C.3 shows the mean redshift offset and the dispersion measured in each of these three cases. As expected, the redshift precision increases as tighter constraints in Δz are used. However, when the constraints are too tight, performance is impacted. For instance, Figure C.6 shows that using Δz of 0.10 and 0.15 results in very similar performance levels, but using Δz of 0.05 implies a performance drop of ~ 0.10 for the high-z quasars. Naturally, the comparison here is not so simple, as the criteria for correct classifications change with the chosen Δz. While this is true, this should only affect those quasars in which the trial redshift is close to the true redshift. In cases where line confusion occurs, then the chosen Δz is irrelevant as the redshift error is much larger. The line confusion plots (Figure 5) indicate that the latter is driving the performance. We conclude that the fiducial $\Delta z = 0.10$ is the optimal threshold choice as it performs similarly to $\Delta z = 0.15$, but with better redshift precision.

C.6. Passing extra features to the random forest classifiers

To finalise the revision of our choices, we assessed the impact of adding extra features to the random forest classifiers. In particular, we explored adding the trial redshift and/or the r-band magnitude to the list of parameters fed to the classifiers. We compared our fiducial choice (adding both parameters) with the cases where only one of the two features is passed, and when none is. The results of this exercise, shown in Figure C.7, indicate that the code does indeed perform optimally when both features are added. However, the performance change is at the percentage level, as in previous cases.
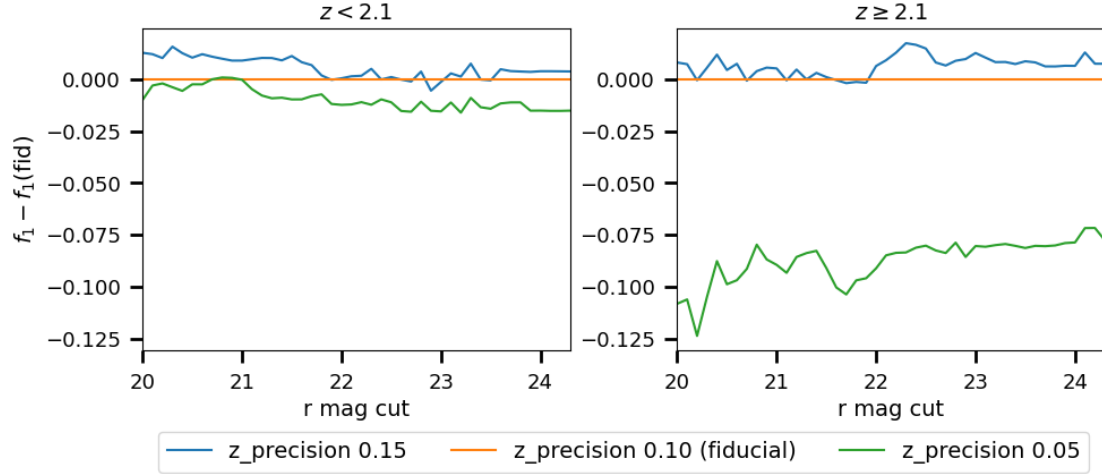
Fig. C.6. Comparison of SQUEzE performance (f_1 score) when using different redshift requirements to define correct classifications. Our fiducial choice is 0.10. The left (right) panel shows the performance for the low (high) redshift quasars. The performance of the fiducial model, with $\Delta z = 0.10$, is marginally worse than the performance with $\Delta z = 0.15$, but the redshift errors are significantly smaller (see Table C.3).
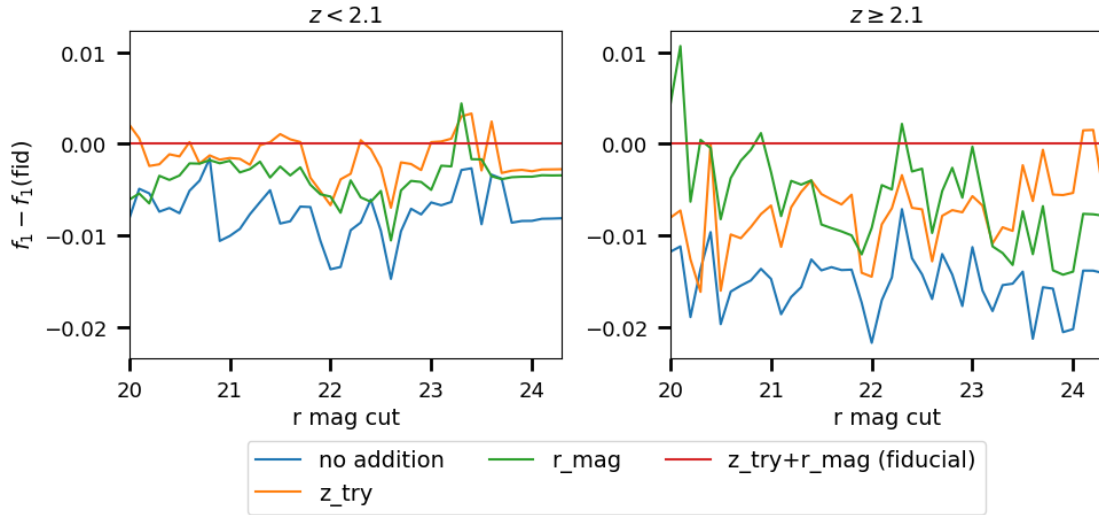

Fig. C.7. Comparison of SQUEzE performance (f_1 score) when we feed different columns to the random forests. We compare SQUEzE default choice, i.e. only feeding the metrics with the cases where we also feed it the trial redshift, the r-band magnitude, or both (our fiducial choice). The left (right) panel shows the performance for the low (high) redshift quasars. The fiducial model is the best performing.