





# The probabilistic random forest applied to the selection of quasar candidates in the QUBRICS survey

Francesco Guarneri <sup>1,2</sup>★, Giorgio Calderone <sup>2</sup>, Stefano Cristiani<sup>2,3,4</sup>★, Fabio Fontanot <sup>2</sup>,  
Konstantina Boutsia <sup>5</sup>, Guido Cupani,<sup>2</sup> Andrea Grazian<sup>6</sup> and Valentina D’Odorico<sup>2,3,7</sup>

<sup>1</sup>*Dipartimento di Fisica, Sezione di Astronomia, Università di Trieste, via G.B. Tiepolo 11, I-34131 Trieste, Italy*

<sup>2</sup>*INAF – Osservatorio Astronomico di Trieste, Via G.B. Tiepolo, 11, I-34143 Trieste, Italy*

<sup>3</sup>*IFPU – Institute for Fundamental Physics of the Universe, via Beirut 2, I-34151 Trieste, Italy*

<sup>4</sup>*INFN – National Institute for Nuclear Physics, via Valerio 2, I-34127 Trieste, Italy*

<sup>5</sup>*Las Campanas Observatory, Carnegie Observatories, Colina El Pino, Casilla 601, La Serena, Chile*

<sup>6</sup>*INAF – Osservatorio Astronomico di Padova, Vicolo dell’Osservatorio 5, I-35122 Padova, Italy*

<sup>7</sup>*Scuola Normale Superiore, Piazza dei Cavalieri, I-56126 Pisa, Italy*

Accepted 2021 June 24. Received 2021 June 23; in original form 2021 March 15

## ABSTRACT

The number of known, bright ( $i < 18$ ), high-redshift ( $z > 2.5$ ) QSOs in the Southern hemisphere is considerably lower than the corresponding number in the Northern hemisphere due to the lack of multiwavelength surveys at  $\delta < 0$ . Recent works, such as the QUBRICS survey, successfully identified new, high-redshift QSOs in the South by means of a machine-learning approach applied on a large photometric data-set. Building on the success of QUBRICS, we present a new QSO selection method based on the Probabilistic Random Forest (PRF), an improvement of the classic Random Forest algorithm. The PRF takes into account measurement errors, treating input data as probability distribution functions: this allows us to obtain better accuracy and a robust predictive model. We applied the PRF to the same photometric data-set used in QUBRICS, based on the SkyMapper DR1, *Gaia* DR2, 2MASS, *WISE*, and *GALEX* databases. The resulting candidate list includes 626 sources with  $i < 18$ . We estimate for our proposed algorithm a completeness of  $\sim 84$  per cent and a purity of  $\sim 78$  per cent on the test data-sets. Preliminary spectroscopic campaigns allowed us to observe 41 candidates, of which 29 turned out to be  $z > 2.5$  QSOs. The performances of the PRF, currently comparable to those of the CCA, are expected to improve as the number of high- $z$  QSOs available for the training sample grows: results are however already promising, despite this being one of the first applications of this method to an astrophysical context.

**Key words:** methods: data analysis – methods: statistical – surveys – quasars: general.

## 1 INTRODUCTION

Luminous quasars, especially at high redshift, play the paramount role of cosmic beacons for a variety of studies on the formation and evolution of galaxies and supermassive black holes (SMBH), dark matter, primordial elements, reionization, cosmological parameters, fundamental constants, and General Relativity. However, finding quasars at high- $z$  is not a trivial task, due to their relative scarcity with respect to other sources with the same apparent luminosity. The advent of the Sloan Digital Sky Survey (SDSS) survey (e.g. Ahumada et al. 2020) has represented a significant improvement in this respect, at least in the Northern hemisphere. At present, the SDSS has delivered more than  $10^5$  (Lyke et al. 2020) spectroscopically confirmed QSOs at  $0 < z < 6.5$ , with a large fraction at absolute magnitudes  $M_{1450} \leq -26$ .

In the Southern hemisphere, due to the lack of wide multiwavelength surveys at  $\delta \leq 0^\circ$ , the situation used to be significantly less favourable. Comparing QSO surface densities (e.g. Véron-Cetty &

Véron 2010) in different parts of the sky, of the 22 known QSOs with  $z > 3$  and  $V < 17$ , only 5 have been found at  $\delta < 0^\circ$ , and all the 3 QSOs with  $V < 16$  are in the North. Besides, recent studies point out that even an exquisite survey as the SDSS can suffer from incompleteness due to colour selection (e.g. Fontanot et al. 2007; Schindler et al. 2019). As a consequence, also in the Northern hemisphere high- $z$  QSO densities could be biased towards lower numbers due to the adoption of efficient but relatively incomplete selections.

In Calderone et al. (2019) and Boutsia et al. (2020), we presented the first results of the QUBRICS survey, aimed at finding  $z \geq 2.5$  QSOs at bright  $i$ -band magnitudes ( $i \leq 18$ ) in the Southern hemisphere, taking advantage of the recent availability of new multiwavelength public databases. The candidate selection in QUBRICS has been based on the Canonical Correlation Analysis (CCA; Anderson 2003) and its success rate in finding  $z > 2.5$  QSOs is estimated to be around 70 per cent, with the predominant contaminants being lower- $z$  QSO at  $z < 2.5$ . Its completeness, evaluated against the presently known bright QSOs at  $z > 2.5$ , turns out to be of the order of 90 per cent (Calderone et al. 2019).

In this paper, we explore the possibility to use other selection methods in order to further increase the purity and completeness of

\* E-mail: francesco.guarneri@inaf.it (FG); stefano.cristiani@inaf.it (SC)

the QUBRICS sample, and to fully exploit the information content of the multiband photometric data bases on which QUBRICS is based. In particular, in this paper we will present and discuss a selection procedure based on the Probabilistic Random Forest (PRF; Reis, Baron & Shahaf 2019), an improvement of the Random Forest that makes it possible to properly include measurement errors in the predictive model and to handle missing data in the data-set.

The paper is organized as follows: in Section 2, we will briefly describe the Random and Probabilistic Random Forest; Section 3 will describe how the initial data-set has been prepared, while in Section 4, the results obtained from our tests will be presented; in Section 5, we will attempt to characterize the spectroscopic sample and compare our results with those obtained with the CCA method, while Section 6 will describe the results of a small spectroscopic campaign. Conclusions are drawn in Section 7.

## 2 MACHINE LEARNING TECHNIQUES

Modern astronomical data-sets are rapidly growing both in size and complexity, thanks to recent multiwavelength and multipoch surveys such as the *SDSS* (Ahumada et al. 2020), *GAIA* (Gaia Collaboration et al. 2020), *DES* (Abbott et al. 2021), or *Pan-STARRS* (Chambers et al. 2016); machine learning (ML) methods are becoming increasingly popular as automatic tools to perform a variety of tasks on these data bases (Baron 2019).

Machine learning (ML) techniques are generally classified into two broad groups/categories: supervised and unsupervised methods. The former are used to map a set of features to a target *label* or quantity, which is provided by a third party actor (another algorithm or a human expert) while the latter are used to infer existing relationships in the data-set, without relying on external labels.

Given a data-set, individual elements are called *objects*, and data associated with a single object *features*. As a practical example, a data-set may be a large photometric collection, where each object is an observed source and each feature is a magnitude measurement. Target labels and quantities differ depending on the specific task, as supervised learning can be used both for classification and regression: in the first scenario (classification) the label is discrete; in the second (regression) it is continuous. Examples for the two cases are, respectively, the classification of a source as a star or a quasar and the estimate of the redshift given a number of photometric measurements. Supervised methods also have model parameters and hyper-parameters: the former are learnt from the data which the model is trained on and are required in the prediction stage; the latter are instead set by the user and fine-tuned to obtain the best performances out of an ML algorithm.

To assess the capabilities of an ML algorithm, it is common practice to subdivide the available data-set into three sub-samples: a training, validation, and testing data-set. The first sample is used to train the algorithm; the validation data-set is used to find the optimal hyper-parameters for the specific task of interest and to gain finer control of the learning process (e.g. to prevent overfitting). The last data-set is finally used to estimate the predictive capabilities of the algorithm, as the learnt model is applied to an unseen data-set. Training, validation, and testing sets should be independent to obtain an unbiased evaluation of the performances of the algorithm. An alternative approach, especially useful in case of a limited data-set, is the *k*-fold cross validation: the original data-set is split in two parts, a training/validation and testing data-set. Training and validation are carried out at the same time: the training data-set is split in *k* subsets; in turn, one of these *k* subsets is used as validation set, while the algorithm is trained on the remaining *k* - 1 subsets. This

allows to perform the validation process without requiring additional subdivisions in the base data-set.

Despite their widespread use and proved success in Astronomy (e.g Carrasco et al. 2015, and references therein), machine learning algorithms in general are not designed to deal with data-sets in which the features have different uncertainties. However, the performances of ML algorithms strongly depend on the signal to noise of the input data (Reis et al. 2019), suggesting that noise and measurement errors play an important part in the learning and predictive process. Available algorithms can be modified to account for uncertainties during the training process, but simple methods are unsuited to extract all available information: for instance, in a Random Forest algorithm (which will be described in Section 2.1) uncertainties in the data-set can be used as additional features; the association between measurements and errors is however indirect, as there is not an explicit probability distribution function involved.

An alternative approach, the Probabilistic Random Forest (PRF), has been recently developed by Reis et al. (2019), who modified the Random Forest technique to directly account for measurement errors.

### 2.1 The original Random Forest

The Random Forest is an ensemble learning method – an algorithm that uses multiple learning algorithms to obtain better predictive performance than any of the constituent learning algorithm alone – that operates by creating a large number of decision trees during the training process (Breiman 2001).

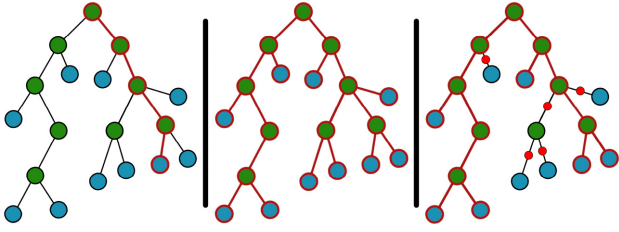
Decision trees are predictive models described by a tree-like graph, used both for classification and regression tasks (Reis et al. 2019). Examples of both employments can be found in Bai et al. (2019) and Silva, Cao & Hayes (2018); in the following, however, we will focus on classification tasks.

Each decision tree is built out of a set of consecutive nodes, and each node is a condition on a feature of the data-set. Conditions are in the form of decision branch:

$$X_i > X_{i,\text{th}}, \quad (1)$$

in which  $X_i$  is the *i*th feature for objects in the data-set and  $X_{i,\text{th}}$  is a threshold value. Both the feature and the threshold value for a node are determined during the training process based on the minimization of a cost function, commonly based on the Gini impurity. The Gini impurity of a given subset is the probability of misclassifying an object, if it is assigned a label randomly drawn from the label distribution of that same subset (Breiman et al. 1984).

We consider as an example a simple two-class (A, B) classification task: the training process starts with the whole training set and a single node, the root of the tree. The algorithm searches for the feature and threshold value that produces the best split, i.e. the one that minimized the aforementioned cost function, determining the condition for the root node. Objects in the training set are then split in two subset, one for which equation (1) is satisfied, one for which equation (1) is not. For both of these, a new best-splitting feature is searched: the process continues iteratively as long as the combined impurity of the resulting two child nodes is lower than the impurity of the parent node. If this condition is not satisfied the current node becomes a terminal node (*leaf*) which does not carry a condition but rather a label: this is determined according to the most common label in the subset associated to the terminal node itself. During the classification process an unlabelled object is propagated along a decision tree according to its feature values and is finally classified based on the terminal node it reaches. An example of a decision tree



**Figure 1.** An object propagates through a decision tree. Terminal nodes (leaves) are light blue coloured. The path of an object along a tree is marked by red lines, black lines show all possible paths. The left-hand panel represents the propagation in the classic RF approach: an object propagates either to the left or right node for each split. The middle panel shows an ideal PRF model: an object propagates along the whole tree, reaching all terminal nodes at the same time, and each object may reach several leaves (although with different probabilities). The right-hand panel shows a ‘pruned’ PRF (i.e. with a set probability threshold): red dots represent nodes which can not be reached due to low probabilities associated with those splits. Adapted from Reis et al. (2019).

can be found, for instance, on the scikit learn (Pedregosa et al. 2011) website.<sup>1</sup>

A simple, unpruned, decision tree is not limited in its size, and shows perfect performances on the training set, while typically showing worse performance when applied on new, unseen data: this behaviour is generally referred to as overfitting (Breiman 2001). The random forest mitigates the issue using numerous decision trees and introducing randomness in the training process. This is usually done using two complementary approaches: each tree is trained on a randomly extracted subset of the original data-set (a technique which is also called bootstrap), and for each node the best splitting feature is chosen from a random subset of all available features; the dimension of the subset is one of the hyper-parameters set by the user. During the prediction process each tree independently classifies each object; the final class is determined by majority vote, i.e. the most common class among all trees is chosen.

## 2.2 Probabilistic Random Forest

The Probabilistic Random Forest (PRF; Reis et al. 2019) is an improvement of the original RF designed to properly handle measurement errors. The main difference between the RF and the PRF consists in the treatment of input data: a ‘classic’ RF algorithm maps feature and labels, while, on the other hand, the PRF also takes feature and labels uncertainties ( $\Delta X$  and  $\Delta y$ ) into account in order to identify the optimal mapping function.

Uncertainties arise both from measurements ( $\Delta X$ ) and classification labels ( $\Delta y$ ). In the PRF implementation, the two are treated quite differently: features are considered as probability distribution functions (PDF), with expectation value equal to the feature value and variance equal to the associated error squared. On the other hand, labels are treated as probability mass functions (i.e. discrete density functions): each object has a fixed chance of belonging to each class.

This simple change has an important effect on decision trees: in an RF an unlabelled object in a given node propagates either to the subsequent right or left node. In the PRF, instead, each object propagates into both nodes with a given probability (Fig. 1); the probability of propagating to the left or right branch are given by the cumulative distribution function for a particular feature; in the current

PRF implementation<sup>2</sup> the PDF is chosen for all objects as a Gaussian, but the choice can be arbitrary. Moreover, as all objects propagate along the whole tree, all leaves contribute to the classification solution of each object.

In principle, any object can always propagate to the next node, even if the probability to do so is small. To optimize the algorithm a probability threshold is introduced: this is implemented in the PRF as an adjustable parameter (*keep\_proba*), with a default value of 0.05 (i.e. an object does not propagate to subsequent nodes if the probability to do so is less than 5 per cent).

The PRF has several advantages over the classic RF:

(i) noise robustness: Reis et al. (2019) tested various noise injections in both the training and testing data, finding that in almost all cases the PRF outperforms the original RF. Improvements in the performance of the algorithm depend on the noise characteristics: noise which produces a clear distinction in objects with poorly and well-measured features leads to negligible improvements; complex noise, that does not result in a clear distinction between feature quality leads to a greater boost in the classification abilities. This is even more noticeable when the noise is different in the training and testing data-set, which is a possible occurrence in astronomy (for instance when measurements are taken from different catalogues);

(ii) missing values: these are rather common in astronomical data-sets, and sometimes many objects are missing measurements for at least one of the selected features. The PRF can naturally handle missing data: an object with a non-measured feature will just propagate both to the left and the right of a node with 50 per cent probability.

## 3 THE PREPARATION OF THE SAMPLE: THE QUBRICS SURVEY

The PRF needs a training set large enough to produce a robust predictive model. The data-set should include sources of interest – high redshift QSOs – together with those that should be excluded by the selection process: in our case typically non-active galaxies, stars, and low-redshift QSOs.

In this work, we have used the same data-set described in the papers by Calderone et al. (2019) and Boutsia et al. (2020) in order to have a direct comparison of the performances of the PRF with other well-established techniques, e.g. the Canonical Correlation Analysis (CCA) used in the QUBRICS survey.

### 3.1 The QUBRICS survey

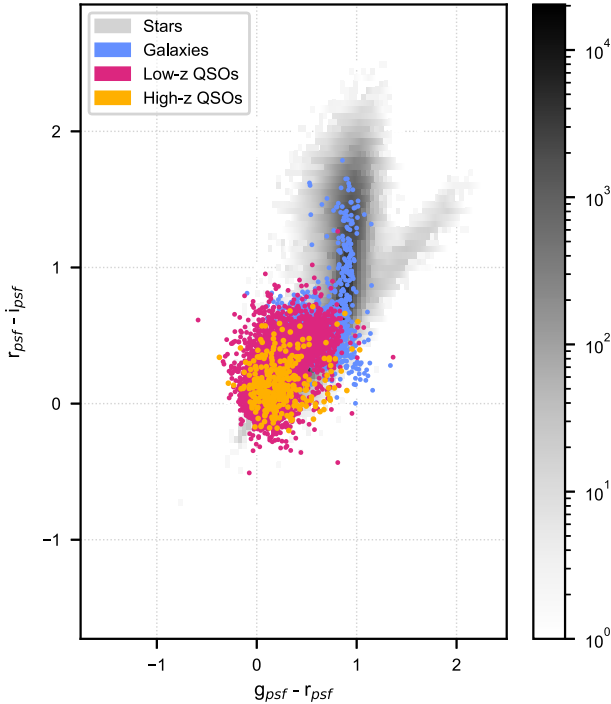
The QUBRICS *Main Sample* (hereafter MS) contains objects with photometric measurements from three catalogues:

- (i) The  $i, z$  magnitudes from the SkyMapper survey (Data Release 1.1 Wolf et al. 2018);
- (ii) The  $G$  magnitude from the *Gaia* survey (Data Release 2 Gaia Collaboration 2016, 2018);
- (iii) The  $W1, W2, W3$  magnitudes from the *WISE* survey (Wright et al. 2010).

In order to be included in the MS an object must have a measured magnitude in *all* these six bands. Further additional constraints introduced in Calderone et al. (2019) are: (i)  $14 < i_{\text{psf}} < 18$ , (ii) galactic latitude  $|b_{\text{gal}}| > 25^\circ$ , (iii) no photometric flags in the  $i$  and  $z$

<sup>1</sup><https://scikit-learn.org/stable/modules/tree.html>

<sup>2</sup><https://github.com/ireis/PRF>



**Figure 2.**  $r - i$  versus  $g - r$  SkyMapper point spread function magnitudes for stars, galaxies, low- and high-redshift QSOs included in the main sample. While there are 4 visible clumps, there is not an efficient way to separate them. Stars were binned with an arbitrary bin-size to better visualize their distribution and the number counts for each bin is shown by the colourmap on the right.

**Table 1.** Number for all sources in the MS, including the most recent QSO candidate sample.

Source type	Number
All	1014 875
Unclassified sources	162 118
<i>Bona fide</i> stars	843 690
Known non-active galaxies	4024
Known QSOs (all $z$ )	5043
CCA QSO candidates (all)	1412
CCA QSO candidates (not yet observed)	818

band<sup>3</sup> ( $iv$ ) matching *GAIA* DR2 and *WISE* sources within 0.5 arcsec and ( $v$ ) SNR > 3 in the first three *WISE* bands; constraints have been designed to reduce contamination, leading to a total of 1014 875 sources over approximately 12 400 square degrees, mostly in the Southern hemisphere.

When available, additional data have been added:  $J$ ,  $H$ , and  $K_s$  magnitudes from the 2MASS survey (Skrutskie et al. 2006),  $u$ ,  $v$ ,  $g$ ,

<sup>3</sup>Photometric flags indicate issues during the image processing; pipeline specific flag are represented as power of two: 1, for instance, indicates that two sources are close enough to bias their respective photometry, 2 that distinct sources were initial blended, 4 the presence of saturated pixels in an object; the complete list is available in the SkyMapper DR1.1 documentation at this web-page. Pipeline flags are combined with a bit-wise OR and the results are given in the published catalogue per source and photometric band, allowing end-users to exclude poorly processed objects (Wolf et al. 2018).

$r$  SkyMapper magnitudes, *Gaia*  $G_{RP}$ ,  $G_{BP}$  measurements and the  $W4$  magnitude from *WISE*. *GALEX* (Bianchi, Shiao & Thilker 2017) data have been added by using the Mikulski Archive for Space Telescopes<sup>4</sup> (MAST). Sources have been cross-matched with a 5 arcsec matching radius; in the rare case of multiple matches the closest has always been retained. Only data produced as part of the All-Sky and Medium-Sky surveys (AIS and MIS, respectively) have been selected.

This additional photometric information is valuable to machine learning algorithms. As shown in Fig. 2 and discussed in Carrasco et al. (2015) it is not trivial to apply a simple colour–colour plot to separate QSOs, especially at high redshift, from contaminants (e.g. non-active galaxies or stars). Including additional photometric information such as infrared magnitudes from *WISE* helps in disentangling different populations, but it is not simple to devise appropriate colour cuts in a multidimensional colour space.

Parallax and proper motion information have been used to identify *bona fide* stars: 83.1 per cent of sources in the MS have been classified as such. The remaining entries have been matched with catalogues of known QSOs and extragalactic sources, in particular the *SDSS* DR14Q (Pâris et al. 2018), the 13th edition of the Véron-Cetty catalogue (Véron-Cetty & Véron 2010) and the 2dFGRS (Colless et al. 2001). The matching process identified 4666 confirmed QSOs and 3665 non-active galaxies. Matching against these catalogues, together with *bona fide* stars identified through *Gaia* parallaxes and proper motion measurements, provided a source type classification for 84 per cent of the original MS. The remaining 16 per cent are unlabelled sources: as described in Calderone et al. (2019) and Boutsia et al. (2020) they are given an estimated classification (non-active galaxies, stars, low- and high-redshift QSOs) and redshift using the CCA.

The CCA method produces a linear transformation matrix: when multiplied with an appropriate magnitude matrix a new label is obtained (hereafter CCA). The CCA procedure ensures that the CCA label is maximally correlated with the classification labels. The same transformation matrix, obtained on known sources, can be applied on unclassified objects: this allows to select the most favourable QSO candidates. The same procedure can be applied in order to obtain a redshift estimate and further exclude contaminants.

Extended objects were discarded to produce a sample of higher purity: this was accomplished following the same approach of Calderone et al. (2019). A measure of the size of an object,  $ext_{iz}$  was obtained by comparing point spread function and petrosian magnitudes from the SkyMapper survey. The difference between the two was initially calculated for stars in the main sample, in order to derive a typical value, per magnitude interval, for point-like sources. The same quantity was then derived for unlabelled sources: those with  $ext_{iz} > 3$  were excluded *a priori* from the selection. In this way, we consider it safe to use psf magnitudes as features, since our focus is on searching for point-like objects like quasars and most of the extended targets are excluded *a priori*.

Finally, as part of the QUBRICS survey, various spectroscopic campaigns provided a secure identification for  $\sim 500$  targets, thus raising the number of known QSOs in the MS to 5043; of these, 428 are at  $z \geq 2.5$ . Additional catalogue matching provided an identification for  $\sim 400$  non-active galaxies. The most recent number of sources in the MS, used in this work, is shown in Table 1.

<sup>4</sup>The website is accessible at: <https://doi:10.17909/T9H59D> MAST.

### 3.2 The training set

Before applying the PRF to a selected data-set a few preliminary operations are needed:

(i) in order to have a balanced number of stars, QSOs and galaxies for the training set we considered only a subset of all available stars. The latter have been chosen in order to evenly sample the available  $i - z$  colour space: sources have been subdivided in bins (0.15-mag wide) based on their colour, and for each bin up to 600 stars were chosen. All objects in bins with less than 600 entries have been kept; bins with more than 600 entries have been randomly sampled: this produces a set of 5814 stars. The data-set built out of the 5814 stars, 5043 QSOs and 4024 galaxies will be referred to as *Reduced Main Sample* and will be the primary training sample for the PRF;

(ii) oversampling the high-redshift QSO sub-sample of the *reduced main sample*. This is necessary to ensure an appropriate training test for the PRF when distinguishing high- and low-redshift sources. We used the `imbalanced-learn` PYTHON module (Lemaître, Nogueira & Aridas 2017), in particular employing the `RandomOverSampler` method. This is the simplest oversampling method available: new samples are obtained by randomly drawing with replacement already available objects;

(iii) distinguishing between nulls and non-detections: in particular an appropriate treatment of the flux upper limits can provide useful data to the algorithm, improving its predictive capabilities (Section 3.3).

### 3.3 Non-detections and missing data

Missing data in photometric data-sets are common and can be the result of two different occurrences: a measurement may be missing because a particular area of the sky has not been observed in a given pass-band or because the target is too faint to be detected in the pass-band. Despite apparently producing the same result in the final data-set – a missing value – the two cases should be treated differently in the implementation of the PRF algorithm, because the information content is different.

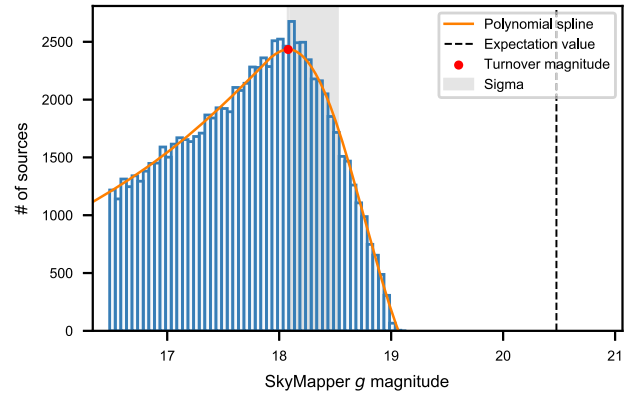
In the following, we will refer to missing data due to the first scenario as `Null` and to non-detections as `ND`. If fed to the machine learning algorithm, both `Null` and `ND` can provide additional information: in the following (Section 4.2.2) it will be shown that supplying `ND`s to the algorithm produces slightly higher completeness (from 77 to 84 per cent) and lower contamination (from 28 to 22 per cent).

In the PRF approach, `Null` are easily dealt with: given a node, for which the splitting condition (equation 1) is based on the  $i$ th feature, an object whose corresponding feature is a `Null` propagates to both left and right node with the same probability: 0.5 for the left, 0.5 for the right node.

`ND`, instead, should propagate like a measured feature, with an appropriate probability distribution.

In order to distinguish `ND` and `Null` we have taken advantage of the additional information found in the published catalogues: SkyMapper and *Gaia* DR2, for instance, provide the number of visits per object per photometric band. If the number of visits is larger than zero, a missing value is considered an `ND`, otherwise it is a `Null`. Often catalogues (e.g. *WISE*), already distinguish `ND` and `Null`.

If a given catalogue did not specify a limiting magnitude, we estimated a reference value for `ND` and an appropriate probability distribution from the properties of the catalogue. In fact, for a given band, objects counts are expected to increase as a function of the magnitude till incompleteness sets in (i.e. the probability of a non-



**Figure 3.** Magnitude distribution for the SkyMapper  $g$  band (blue histogram). The yellow line represents the polynomial spline used to estimate the peak of the distribution (red dot at magnitude  $\sim 18$ ). The dashed, black line shows the reference value used for the  $g$  band in the *Reduced Main Sample*, obtained as the expectation value of the appropriate PDF; the shaded area shows the  $\sigma$  interval associated with the reference value.

detection becomes non-negligible and increases as a function of the magnitude until in practice it becomes one).

We followed two different approaches:

(i) a non-detection is assumed to have a magnitude corresponding to a signal-to-noise ratio roughly equal to 1, to which a Gaussian PDF is associated with  $\sigma = 1.085$  (corresponding to  $\text{SNR} = 1$ ). For example, in the case of the SkyMapper bands, assuming a background limited regime, we can determine the magnitude for which a typical SNR is achieved, e.g.  $\text{SNR} = 10$  (and, correspondingly, a  $\sigma_m = 0.1085$ ), and from this magnitude, e.g.  $m_{10}$ , derive the reference value for `ND` as:

$$\text{ND} = m_{10} + 2.5 \quad (2)$$

(ii) the magnitude distribution at the limit of the detections has been used to estimate the probability distribution for `ND` (a low-pass distribution, equation 3). For each photometric band of interest, a large number ( $\sim 10^5$  or more) of sources from the same survey has been collected from randomly selected regions in the Southern hemisphere and a histogram has been built with a 0.05 mag bin. A polynomial spline has been used to interpolate the histogram and obtain a reliable estimate of the magnitude at the turnover of the counts ( $\text{TM}$ , the red dot in Fig. 3). Once the  $\text{TM}$  is determined, we introduced a low-pass distribution, expressed as

$$\begin{cases} f(m) = N[1 - \exp(-\frac{\text{TM}-m}{\sigma})] & \text{TM} \leq m \leq \text{TM} + k\sigma \\ f(m) = 0 & \text{otherwise} \end{cases}, \quad (3)$$

where  $\sigma$  is the 68 per cent percentile of the sources fainter than  $\text{TM}$ ,  $k$  is the upper limit needed to have a finite distribution, and  $N$  is the normalization coefficient needed to ensure that  $\int_{\text{TM}}^{\text{TM}+k\sigma} \text{PDF}(m) dm = 1$ . The parameter  $k$  has been chosen to be 10, which produces values similar to those determined with the approach (i); the final non-detection value is obtained by calculating the expectation value of the distribution (3) as  $\text{ND} = \int_{\text{TM}}^{\text{TM}+k\sigma} m \text{PDF}(m) dm$ . Different choices of  $k$  do not affect significantly the results of the PRF classification, provided that its value is chosen large enough ( $k > 5$ ). This is not surprising, since the PRF, in the training phase, has the capacity to adapt the probability thresholds described in Section 2.2 to our choice of  $k$ .

Turnover values and standard deviations estimated for the *Reduced Main Sample* are listed in Table 2.

**Table 2.** Turnover, associated error  $\sigma$  for each photometric band in the *Reduced Main Sample*.

Survey band	Turnover magnitude ( <i>AB</i> )	$\sigma$ (mag)
SkyMapper <i>u</i>	17.89	0.46
SkyMapper <i>v</i>	17.65	0.42
SkyMapper <i>g</i>	18.08	0.44
SkyMapper <i>r</i>	18.10	0.41
<i>Gaia</i> <i>G<sub>BP</sub></i>	20.81	0.46
<i>Gaia</i> <i>G<sub>RP</sub></i>	19.43	0.47
2MASS <i>J</i>	16.55	0.29
2MASS <i>H</i>	15.84	0.36
2MASS <i>K<sub>s</sub></i>	15.44	0.35
WISE <i>W4</i>	15.64	0.39
GALEX NUV (AIS)	22.56	0.38
GALEX FUV (AIS)	22.16	0.46
GALEX NUV (MIS)	23.78	0.38
GALEX FUV (MIS)	23.53	0.41

**Table 3.** Number counts for objects used in the *Reduced Main Sample*.

Source type	# of sources available
Quasar (all <i>z</i> )	5043
Quasar ( $z \geq 2.5$ )	428
Quasar ( $z < 2.5$ )	4615
Non-active galaxies	4024
Star	5814

Both methods are rather rough approximations. They provide similar results with slightly better results, in term of contamination of the test sample, for the approach (ii), which has been adopted in the following.

## 4 APPLYING THE PRF TO THE REDUCED MAIN SAMPLE

In this section, we test the capabilities of the PRF in finding high-redshift (for our purposes  $z > 2.5$ ) QSOs, aiming at the production of a high-purity sample, in order to minimize the investment of telescope time for spectroscopic follow-up, but also of sufficient completeness for applications such as the calculation of luminosity functions. The algorithm will provide a classification based on the available magnitudes/colours, as the main source of information.

### 4.1 General approach

We first apply the PRF to a sample including all types of sources (i.e. stars, non-active galaxies, and quasars of all redshifts). By construction, the *Reduced Main Sample* is built so that the fraction of each component is roughly one third of the total. Due to the typical surface densities of the various categories, the number of high-redshift quasars with respect to the total is relatively small (roughly 8.5 per cent of all QSOs in the *Reduced Main Sample*). Special care will then be needed when dealing with predictions of the algorithm for this kind of objects. The number of sources for each class is listed in Table 3.

A training data-set for the PRF includes two components: a magnitude matrix (with associated errors) and a label vector, possibly with uncertainties in the classification. The latter are not mandatory, and in this work only feature (i.e. magnitude) errors have been used. Class labels are for the most part assigned by means of information

derived from the literature or based on assumptions – for instance, *bona fide* stars – and in both cases no uncertainties could be given. Two data-sets have been used for the various tests: the *Reduced Main Sample*, and its sub-sample containing only QSOs. Classification labels (class-labels) for all tests are numerical: the association of a class with a number is completely arbitrary, and the results provided by the PRF do not depend on the choice of the label.

The algorithm is trained and validated using a *k*-fold cross validation; results are then checked against an independent test set. Validation+train and test data-set are randomly generated at each run using a defined random state; special care has been taken to ensure that objects in both the train and test data-sets follow the same class distribution; the training+validation data-set has been chosen to be 80 per cent of the available sources; the remaining objects have been used as a test set.

The PRF hyper-parameters, described in Reis et al. (2019) and in the corresponding PRF GitHub repository, have been chosen on the basis of a 5-fold cross-validation test: a higher *k* did not produce meaningful differences. Based on the results of the *k*-fold test we chose to use 200 trees for each test, `sqrt` for `max_parameters` and 0.05 for `keep_proba`; other parameters have been kept at their default values. Finally, each decision tree is built out of a bootstrapped sub-sample of the original training set (i.e. we set `bootstrap = True` during the PRF initialization). To avoid biases due to a small testing data-set we have also chosen to repeat the process 100 times: we have split the original data-set in training and testing using a defined random state – unique for each of the 100 iterations – and checked the consistency of the results.

The predictions produced by the algorithm have been evaluated on the basis of contamination (i.e. the complementary of the precision, the fraction of relevant instances among the retrieved instances) and completeness (i.e. the recall, the fraction of relevant instances that are retrieved), defined as

(i) **Contamination**: the number of undesired, but selected, sources over the total number of selected sources. The definition can be restated as  $\frac{FP}{TP+FP}$ , where TP is the number of true positives and FP the number of false positives;

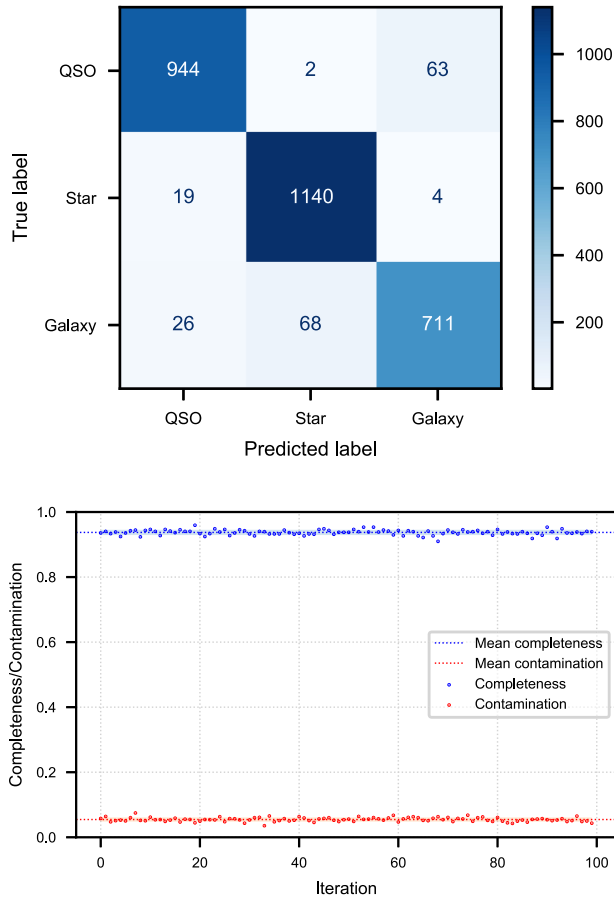
(ii) **Completeness**: the ratio of the number of sources of interest identified by the algorithm over the total number of sources of interest (independently known). The definition can be restated as  $\frac{TP}{TP+FN}$ , where TP is the number of true positives and FN the number of false negatives. It should be noted that in this context the definition of completeness does not take into account ( $i < 18$ ) QSOs that may be existing in the sky and are not present in the QUBRICS MS because they were absent in some of the key data bases, e.g. skymapper.

### 4.2 The PRF as classifier

#### 4.2.1 QSOs, stars, and galaxies

The algorithm has been first used to distinguish QSOs, stars, and non-active galaxies; the three classes have been labelled respectively as 1, 2, and 3. Training, validation, and testing data-sets have been all extracted from the whole *Reduced Main Sample*. A confusion matrix is used to visualize the prediction of the algorithm (Fig. 4, upper panel).

The test data-set is composed of 1009 QSOs, 1163 stars, and 805 galaxies: as shown in Fig. 4 roughly 93 per cent of QSOs are correctly classified by the algorithm. Moreover, most (62 out of 63) of misclassified QSOs are at  $z < 0.5$  and are identified as galaxies: this is not surprising, as the spectral energy distribution of low-redshift



**Figure 4.** Top panel: Prediction of the algorithm for the test data-set when used to separate QSOs, stars, and galaxies represented as a confusion matrix. Bottom panel: completeness and contamination calculated over 100 iterations. The plot refers to QSOs only.

quasars can be dominated by the host galaxy, and a spectroscopic observation is required to reveal the presence of the QSO. Both stars and galaxies contaminate the QSO sample, even if the latter are more commonly selected as QSOs. Considering the QSO as the target class the algorithm scores a recall of 93 per cent with a contamination of 5 per cent. We note, however, that our final aim is to identify high-redshift ( $z > 2.5$ ) sources: out of the 89 in the test data-set, 88 ( $\sim 99$  per cent) are correctly classified by the algorithm as quasars.

This result is tied to a particular choice of train/testing data-sets: different test sources might provide slightly different results; we thus evaluated the performance of the algorithm 100 times, using a different test data-set at each iteration, averaging the results at the end. Results are shown in Fig. 4 lower panel, where both completeness and contamination are calculated with respect to the QSO class: the average completeness (contamination) is 93.7 per cent (5.5 per cent) with a scatter of  $\sigma = 0.8$  per cent ( $\sigma = 0.7$  per cent); the percentage of high-redshift QSOs identified by the algorithm is, in each run,  $\sim 98.5$  per cent with a large scatter,  $\sim 1.5$  per cent.

#### 4.2.2 Low- and high-redshift QSOs

As the final goal of this work is to identify QSOs with  $z > 2.5$ , having a reliable classification as a QSO is not sufficient: we still need to exclude low-redshift QSOs that in Calderone et al. (2019) and Boutsia

et al. (2020) have been found to be the major contaminant. To this end we have applied a second time the PRF to the objects classified as QSOs in the previous step, trying to discriminate whether their redshift is higher or lower than a threshold, initially chosen to be  $z = 2.5$  in accordance with the QUBRICS definition of high- $z$  QSOs. This choice results in an unbalanced training data-set (as shown in Table 3), as the number of objects above  $z = 2.5$  is just the 8.5 per cent of the total. Moreover, more than 2/3 of sources available for training at  $z < 2.5$  are at  $0 < z < 1.5$ : in both cases the redshift distribution of sources in the training sample is not suitable for our purposes and negatively impact on the performances of the algorithm. In order to mitigate the issue we have chosen to apply a simple oversampling method: in our approach, new samples are generated by sampling with repetition the available data. We experimented with different oversampling strategies, in order to find the best compromise in term of completeness versus contamination. The best results were obtained by applying the oversampling algorithm twice: once for objects with redshift between 0 and 3, in order to produce a ratio of sources with  $2 \leq z < 3$  to those with  $0 \leq z < 2$  equal to 0.5; the second for sources at  $z \geq 2.5$ , in order to match the number of low-redshift objects. This choice has been adopted to preserve all the available information in the data-set: undersampling the majority class would remove precious information which could instead be used by the algorithm. Moreover, we have not used a more advanced oversampler (e.g. SMOTE; Chawla et al. 2002) due to the missing values in our data-set: SMOTE creates synthetic instances by searching for nearest neighbours and averaging over their corresponding feature values.

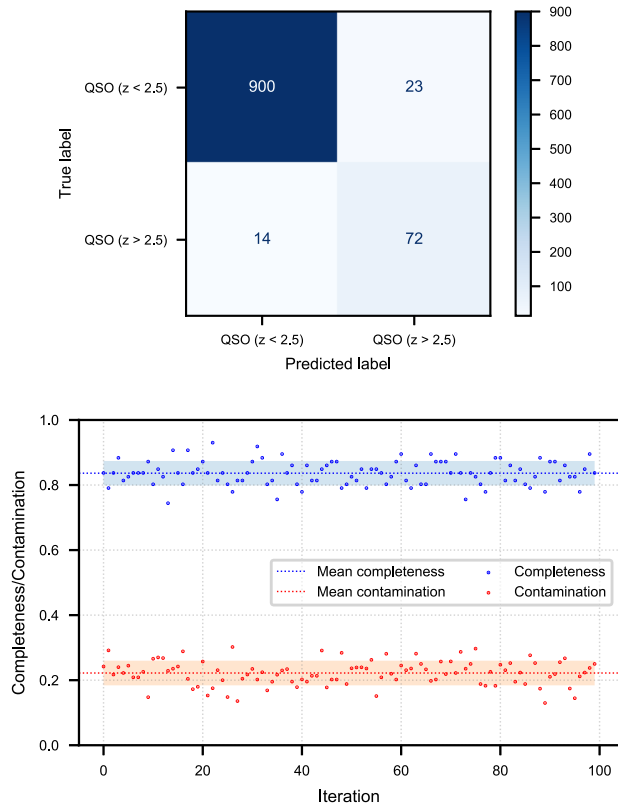
These results also suggest that the data-set currently available is small with respect to the complexity of the problem: a better and more ample training set would greatly benefit the performances of the PRF.

The available data have been subdivided once again in training + validation and testing, with the same ratio used in the previous test (80–20 per cent). Results for the classification process on a test data-set are shown in Fig. 5.

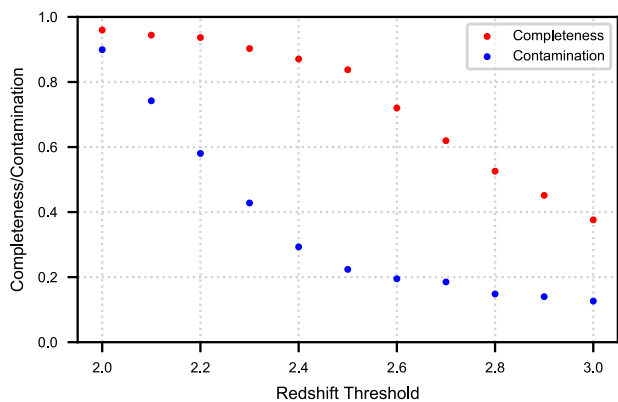
We repeated the same procedure described in the previous section, in order to avoid biases due to the particular training/testing sample. In this case, we achieve an average completeness (contamination) of  $\sim 84$  per cent ( $\sim 22$  per cent), with higher scatter with respect to what we observed before: 3 per cent for both completeness and contamination. These values can be compared to those obtained for the QUBRICS survey for a similar threshold (fig. 5 in Calderone et al. 2019, green line): the CCA produces a data-set with a slightly higher completeness, but higher contamination ( $\sim 37$  per cent).

During the training process, it is possible to set a lower redshift threshold in order to produce higher completeness – at the expense of higher contamination – with respect to the original,  $z = 2.5$  threshold. Conversely, higher redshift thresholds will produce lower contamination and, at the same time, lower completeness. In order to test the effect on the selection process different thresholds have been selected sampling the redshift interval  $z = 2-3$  with steps of 0.1 in redshift units. The results are shown in Fig. 6.

As expected, both completeness and contamination rise as the threshold value becomes lower. In order to obtain the same completeness expected for QUBRICS, one should take a redshift limit of  $z \simeq 2.3$ ; the contamination is similar to that achieved with the CCA (roughly 40 per cent). The completeness in the test sample rapidly decreases (40 per cent from  $z = 2.5$  to  $z = 3.0$ ) as redshift thresholds get higher, whereas the contamination decreases at a slower pace (roughly 10 per cent in the same redshift interval): we thus chose to use the  $z = 2.5$  threshold in the application on the unclassified sample.



**Figure 5.** Results of the classification process high- versus low-redshift QSOs in the case of a redshift threshold of  $z = 2.5$ . Top panel: confusion matrix for this test. Bottom panel: completeness and contamination calculated over 100 runs with different test and train data-set. Dashed lines mark the average values, dots results for a particular run while the shaded regions denote the  $1\sigma$  interval.



**Figure 6.** Completeness (red dots) and contamination (blue dots) as a function of the delimiting redshift for the low- and high-QSO classes. Each completeness-contamination pair in the plot is the average of 100 iterations on different test data-sets, and is calculated with respect to the  $z = 2.5$  threshold.

#### 4.2.3 Analysis of the contaminants

The contaminants affecting the final sample of high-redshift QSOs will be of two types: lower redshift ( $z < 2.5$ ) QSOs wrongly classified at high redshift and galaxies/stars misclassified as QSOs in the first step that survive the subsequent classification as high-redshift QSOs.

The performance of the algorithm for the first type of contaminants has been treated in the previous subsection. The probability of a QSO with  $z < 2.5$  to be classified as a high- $z$  QSO ( $P_Q$  in the following), as shown in Fig. 5, is on average 2 per cent.

To quantify the global performance concerning non-QSO contaminants, it is necessary to combine the two previous tests. To this end, the algorithm has been initially trained on part (80 per cent) of the *Reduced Main Sample* and applied on a test data-set (the remaining 20 per cent); both train and test data-sets contain stars, galaxies, and QSOs, and this produces a QSO candidate sample at all redshifts, which includes misclassified sources, i.e. stars and galaxies predicted to be QSOs. These are then re-classified as low- or high-redshift sources, allowing to verify how many non-QSO contaminants are picked up at the end by the algorithm. The process has been repeated 100 times, each with different train-test data-sets.

On average 4 per cent ( $\sim 35$ ) of the galaxies and 2 per cent ( $\sim 20$ ) of the stars are classified as generic QSOs. Of these, 0.1 per cent of the galaxies ( $PG$  in the following) and 0.1 per cent (1) of the stars on average are classified as high redshift sources.

In order to avoid issues tied to the small number of objects in the test set, we repeated the process using all *bona fide* stars in the MS, excluding those part of the *Reduced Main Sample* and used during the training process. Out of the 837 876 initial *bona fide* stars, on average  $\sim 2400$  (0.3 per cent) are picked up by the algorithm as generic QSOs at all redshifts. Of these only  $\sim 400$  (0.05 per cent,  $PS$  in the following) are selected as QSO candidates at  $z \geq 2.5$ .

## 5 CHARACTERIZATION AND COMPARISON WITH THE CCA SELECTION

In order to obtain a list of high-redshift QSO candidates, the PRF has first been applied on the unclassified sources in the MS, trained as described in Section 4.2.1. As described in Section 3.1, extended objects have been discarded *a priori*, leaving a total of 58 782 objects (Unclassified Dataset, UD).

The PRF, applied to the UD, has produced a list of 22113 QSO candidates (at all redshifts), 18 573 stars and 18 096 potential galaxies.

In order to select high-redshift ( $z \geq 2.5$ ) QSOs the PRF has been trained on the QSO sub-sample of the *Reduced Main Sample*, as described in Section 4.2.2, and then has been applied to the previously selected generic QSO candidates. The threshold defining the high-versus low-redshift class has been chosen at  $z = 2.5$ . This second selection has identified a final sample of 626 high- $z$  QSO.

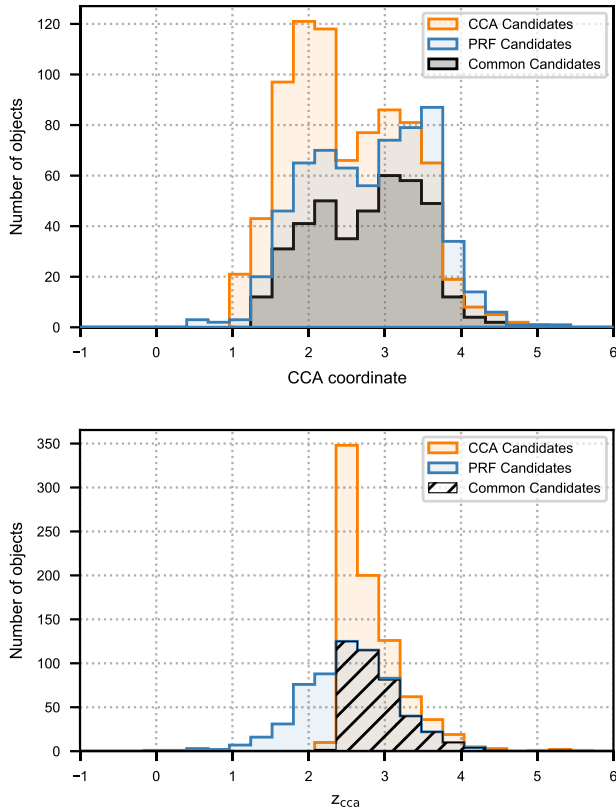
The completeness of this list, for  $z \geq 2.5$  quasars is expected to be, according to Section 4.2, 83 per cent.

In order to estimate the success rate of a spectroscopic follow-up we can assume that the partition of the unclassified sources in the MS in stars, galaxies and generic quasars is described by the results of the PRF classifier, i.e. 31 per cent of stars, 31 per cent of galaxies, and 38 per cent of generic QSOs. The number of expected contaminants, due to the misclassification of stars and galaxies, turns out to be low:  $18573 \times PS = 18$  and  $18096 \times PG = 9$ , respectively.

To estimate the more significant contamination of misclassified low- $z$  QSOs and, conversely, the number of high- $z$  QSOs not selected, we have convoluted the expected redshift distribution of  $i \leq 18$  QSOs, derived from Shen et al. (2020), with the probability, as a function of the redshift, for a  $z < 2.5$  QSO to be classified at high- $z$  and for a  $z \geq 2.5$  QSO to be classified at low- $z$  (computed as in Section 4.2).

As a result, in the list of 626 high- $z$  QSO candidates we expect to have about 66 per cent (411) true  $z \geq 2.5$  QSOs, 30 per cent (188)  $z < 2.5$  QSOs, 4 per cent (27) galaxies or stars. 16 per cent



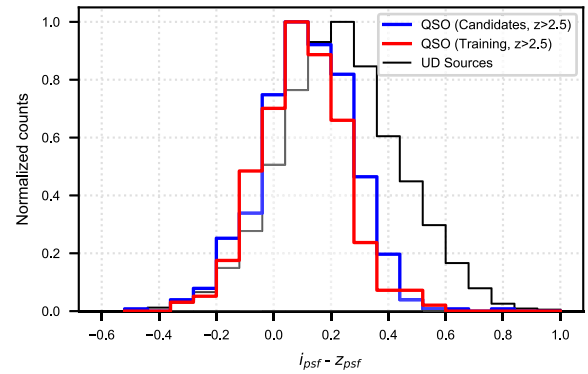


**Figure 7.** CCA coordinate (top panel) and CCA redshift estimate  $z_{\text{CCA}}$  (bottom panel) for sources selected by the PRF (blue), CCA (orange) or both algorithms (green). Five sources at  $z_{\text{CCA}} > 10$  have been excluded from the plots, and are most likely contaminants. The CCA coordinate corresponds to an object-type classification, while the  $z_{\text{CCA}}$  is an estimate for the redshift of the object. As described in Calderone et al. (2019), non active galaxies are spread around  $\text{CCA} \sim -1$ , stars around  $\text{CCA} \sim 0$ , low- $z$  QSOs around  $\text{CCA} \sim 2$ , and high-redshift QSOs have  $z_{\text{CCA}} \gtrsim 3$ .

of the  $z \geq 2.5$  QSOs, mainly around  $z = 2.5$ – $2.9$ , are expected to have been missed in the selection. The lower- $z$  QSO misclassified as high-redshift sources with  $z > 2.5$ , are actually expected to have a redshift  $\langle z \rangle \sim 2.1$  and all at  $z > 1.5$ .

### 5.1 Comparison with the CCA predictions

The sample of candidates obtained using the PRF has been compared with the equivalent list produced using the CCA method (Calderone et al. 2019): 401 sources turned out to be in common, while 417 and 225 objects are exclusively selected by the CCA and PRF, respectively. Fig. 7 shows the comparisons between the CCA coordinate (top panel) and the  $z_{\text{CCA}}$  estimate for PRF (CCA) and common candidates. Most (>95 per cent) of the PRF selected sources are at  $\text{CCA} > 1$ , corroborating the choice made in Calderone et al. (2019) to exclude from the candidate list unclassified sources with  $\text{CCA} < 1$ . On the other hand, a significant (190 out of 227) part of PRF selected objects is at  $z_{\text{CCA}} < 2.26$ : 61 per cent of these are however within the  $1\sigma$  scatter found for the CCA selection estimate. Taking the corresponding estimates for the completeness and success rate of the CCA method from Boutsia et al. (2020), and with the rough assumption that the CCA and PRF selection are statistically independent, we would expect the intersection of the two selections to have a 89 per cent success rate and a 75 per cent completeness.



**Figure 8.** The  $i_{\text{psf}} - z_{\text{psf}}$  colour for QSOs at  $z > 2.5$ . The red histogram shows to the colour distribution for objects in the training sample, the blue one for candidates and the grey one for all unclassified sources (UD data-set). Each histogram has been normalized by its maximum in order to show all of them in the same plot.

### 5.2 Colour comparison

The PRF uses magnitudes as features to provide a label for unclassified objects. It is interesting to compare the magnitude or colour distribution of the newly classified objects to that of the training: we expect the colour distribution of the two to be similar. The  $i - z$  colour was chosen in this case, as all sources in the MS are required to have a reliable magnitude in these two bands.

As shown in Fig. 8 there is good agreement between the distribution of colours for the training and candidate sample, especially for high-redshift QSOs, the main interest for this work. This holds true for non-active galaxies as well, while the relation is less tight with stars: the training set was built to fully sample the available  $i - z$  colour space, while the UD occupies a slightly smaller range.

The effect of Nulls was verified as well, considering sources with at least one Null in a photometric band: these are rather scarce, and the results consequently noisy. Due to their small number and random nature (e.g. a switched-off detector or a corrupted image), Nulls are not expected to produce significant deviations in the colour distributions. Considering the SkyMapper  $u_{\text{psf}}$  as an example (one of the bands where Nulls are more numerous), there are 26 of 5043 QSOs in the training sample and 215 of 22113 QSO candidates with a Null value. The sample mean  $i - z$  for QSOs in the training sample (QSO candidates) is 0.08 (0.13) with a standard deviation of 0.17 (0.18): hence, there is not a significant difference in the two distributions.

## 6 SPECTROSCOPIC VALIDATION

Since the creation of the PRF candidate list, observations of 41 sources have been carried out at Las Campanas Observatory and at Telescopio Nazionale Galileo (TNG, La Palma) using LDSS-3 (Clay Telescope), IMACS (Baade Telescope), and Dolores.

Observations at LDSS-3 have been obtained with the VPH-all grism, 1 arcsec-central slit and no blocking filter, covering a wavelength range between 4000 and 10000 Å and a low ( $\sim 800$ ) resolution. Observations at IMACS used the #300 grism with a 17.5 deg blaze angle with a dispersion of  $1.34 \text{ \AA pixel}^{-1}$  and the same wavelength range of LDSS-3. Exposures at TNG have been taken during the AOT41 and AOT42 periods under two proposals (PI: G. Calderone and A. Grazian); the LR-B grism with a 1 arcsec slit aperture have been used.

**Table 4.** Observed, PRF selected sources with a robust spectroscopic redshift estimate. Sources with a \* show strong Broad Absorption Lines (BAL QSOs).

	QUBRICS ID	RA (J2000)	Dec.	$m_i$ (AB mag)	$z_{\text{spec}}$	Class	Obs. date	Instrument	CCA selected
1	814160	13:01:18.31	-08:10:14.81	17.68	3.281	QSO	2021-01-29	LDSS3	Y
2	824362	12:33:22.24	-11:53:39.53	17.90	3.183	QSO	2021-01-29	LDSS3	Y
3	831008	20:43:56.68	-00:39:08.48	17.86	2.894	QSO	2020-10-23	DOLORES	Y
4	831970	20:44:59.66	-02:54:38.28	17.79	2.851	QSO	2020-10-23	DOLORES	Y
5	842834	12:11:20.09	-33:14:27.46	17.73	3.826	QSO	2021-01-29	LDSS3	Y
6	859489	21:58:00.41	-07:18:05.50	17.80	2.538	QSO	2020-10-23	DOLORES	Y
7	859798	21:33:26.01	-03:23:32.70	17.78	2.518	QSO	2020-10-23	DOLORES	N
8	861290	21:31:58.81	-04:54:39.47	17.97	3.381	QSO	2020-10-23	DOLORES	Y
9	861881	21:54:15.85	-04:45:21.89	17.65	2.366	QSO	2020-09-13	DOLORES	Y
10	864254	23:34:54.76	-69:30:42.84	17.86	3.894	QSO	2021-01-02	IMACS	Y
11	866799*	02:10:51.46	-84:54:37.57	17.17	3.685	QSO	2020-11-27	IMACS	Y
12	893444	02:53:56.09	-20:26:39.68	17.86	3.022	QSO	2021-01-28	LDSS3	Y
13	908786	23:22:27.67	-04:48:45.16	17.78	2.589	QSO	2020-10-23	DOLORES	Y
14	992797*	01:07:10.57	-62:36:48.35	17.28	2.833	QSO	2020-10-07	LDSS3	Y
15	995059	23:32:46.67	-08:23:44.03	17.31	2.894	QSO	2020-10-23	DOLORES	Y
16	1005746	04:14:29.34	-05:56:14.38	17.33	2.417	QSO	2020-09-13	DOLORES	N
17	1007009*	15:07:26.88	-16:25:42.30	17.86	3.015	QSO	2021-02-27	IMACS	Y
18	1013347	04:42:30.12	-26:32:19.05	17.55	2.914	QSO	2020-11-27	IMACS	Y
19	1018840	03:27:24.51	-52:38:58.20	17.79	3.771	QSO	2020-12-31	IMACS	Y
20	1026400	06:36:44.21	-63:40:33.04	17.69	2.410	QSO	2020-11-24	LDSS3	Y
21	1030917	00:36:25.37	-32:23:36.55	17.80	3.512	QSO	2020-11-26	LDSS3	Y
22	1031280	22:33:47.55	-04:02:04.60	17.89	2.092	QSO	2020-10-23	DOLORES	N
23	1031462	22:36:06.12	-16:10:34.21	17.80	2.109	QSO	2020-10-23	DOLORES	Y
24	1031929	02:10:25.34	-38:17:17.96	17.42	3.308	QSO	2020-11-26	LDSS3	Y
25	1032609	00:33:11.87	-40:51:49.35	17.82	1.963	QSO	2020-11-26	LDSS3	Y
26	1033197	23:41:33.99	-20:24:08.81	17.56	2.603	QSO	2020-09-13	DOLORES	Y
27	1034040	00:58:23.18	-09:04:34.98	17.82	2.775	QSO	2020-10-23	DOLORES	Y
28	1034851	01:18:08.11	-23:07:56.31	17.36	3.096	QSO	2020-10-08	LDSS3	Y
29	1035092	00:29:55.80	-22:26:28.53	17.73	2.782	QSO	2020-10-23	DOLORES	Y
30	1039886	22:40:56.35	-08:03:58.41	17.71	2.463	QSO	2020-09-13	DOLORES	N
31	1040503	01:24:36.61	-31:26:23.61	17.20	1.902	QSO	2020-10-07	LDSS3	Y
32	1041074	00:55:53.34	-23:07:43.84	17.90	2.218	QSO	2020-10-08	LDSS3	N
33	1041119	23:09:35.13	-15:13:14.27	17.95	2.339	QSO	2020-10-23	DOLORES	N
34	1044054	00:31:50.88	-18:20:21.84	17.75	3.519	QSO	2020-10-23	DOLORES	Y
35	1044577*	00:21:11.30	-17:29:01.04	17.68	1.888	QSO	2020-09-13	DOLORES	Y
36	1059422	01:54:48.05	-10:49:40.61	17.43	2.556	QSO	2020-10-08	LDSS3	Y
37	1080395	03:12:52.40	-31:38:33.21	17.83	3.879	QSO	2020-11-27	IMACS	Y
38	1086629	04:39:25.68	-43:49:17.87	17.68	3.516	QSO	2021-01-31	IMACS	Y
39	1094391	04:55:55.50	-64:58:35.35	17.48	2.444	QSO	2021-01-31	IMACS	Y
40	1101726	14:59:01.01	-02:51:05.79	17.75	3.354	QSO	2021-02-27	IMACS	Y
41	1122453	02:35:57.55	-34:48:56.45	17.79	3.737	QSO	2020-11-24	LDSS3	Y

Out of the 41 PRF selected and observed sources, 29 turned out to be genuine high- $z$  ( $z > 2.5$ ) QSOs and 12 QSOs with  $1.88 < z < 2.5$ ; no stars nor galaxies have been selected by the algorithm. The results of the spectroscopic observations are summarized in Table 4. In these preliminary observations, we achieved a success rate of  $\sim 70$  per cent that becomes 80 per cent if we consider the candidates in common with the CCA selection. The  $z < 2.5$  contaminant QSOs turn out to be 12, with an average redshift  $\langle z \rangle = 2.2$  and a minimum redshift of  $z = 1.888$  (for a BAL QSO, ID = 1044 577), in good agreement with the predictions of Section 5, based on the characterization of our selection method. As observed in Boutsia et al. (2020) and detailed in Cupani et al. (2021) lower  $z$ , extremely strong BAL QSOs are picked up because their huge absorption troughs tends to mimic the colours of higher redshift QSOs.

It should be noted that the number of observed targets is still low ( $\sim 5$  per cent of the whole candidate list), and in these exploratory runs targets have not been chosen in a systematic way and might not be entirely representative of the final performance of the PRF method. In any case, the results appear encouraging and further observations worth pursuing, possibly in parallel with other selection techniques, in order to better evaluate the capabilities of the PRF method and

enlarge and make more complete the sample of bright high-redshift quasars in the Southern hemisphere.

## 7 CONCLUSIONS

Searching for QSOs is a challenging task, even more so if relatively high-redshift sources are the goal. In this paper, we presented a selection method based on a machine learning algorithm, the Probabilistic Random Forest, and used it to select relatively bright ( $i < 18$ ) high-redshift ( $z > 2.5$ ) QSOs. The PRF has been applied to the same initial data-set used for the QUBRICS survey (Calderone et al. 2019), including photometric estimates from the SkyMapper DR1, *Gaia* DR2, *WISE*, 2MASS, and *GALEX* surveys. We have first used the PRF algorithm to select QSOs (at all redshifts), in order to remove stars and non-active galaxies; we then re-classified the QSO candidates, in low- and high- $z$  QSO candidates. Our tests show that, when applied to the QUBRICS sample, the PRF selection has a completeness of  $\sim 83$  per cent in selecting high-redshift sources, with a relatively low contamination of  $\sim 22$  per cent. Similarly to what observed in Calderone et al. (2019), the main responsible for

contamination turn out to be low- $z$  QSOs (93 per cent); stars and non-active galaxies are of secondary importance ( $\sim 3$  and  $\sim 4$  per cent, respectively).

When applied to the unclassified data-set of QUBRICS, which contains 58 782 sources, the algorithm produces a list of 626 high redshift QSO candidates: of these, 401 are in common with the equivalent CCA sample of Calderone et al. (2019), while the remaining 225 are exclusively selected by the PRF.

With preliminary observations of 41 PRF candidates we have been able to confirm 29 new high- $z$  sources, with a success rate close to our expectations. Further spectroscopic identifications are needed to better assess the capabilities of the PRF method.

The relatively small number of high-redshift QSOs available for the training ( $< 10$  per cent of the total) likely hampers the PRF performances, and we have had to resort to oversampling in order to obtain reasonably uniform training sets, as described in Section 4.2. None the less, the PRF has proven to be a powerful and flexible technique to select high-redshift quasars, competitive with respect to other techniques such as the CCA.

We are refining the selection methods and continuing the spectroscopic campaigns, in order to further improve the completeness and success rate of the QUBRICS survey and to extend, with the growth of the training sets, the predictive capabilities to more specific QSO categories (e.g. Boutsia et al. 2021; Cupani et al. 2021).

## ACKNOWLEDGEMENTS

We thank Società Astronomica Italiana (SAIt), Ennio Poretti, Gloria Andreuzzi, Marco Pedani, Vittoria Altomonte, and Andrea Cama for the observation support at TNG. Part of the observations discussed in this work are based on observations made with the Italian Telescopio Nazionale Galileo (TNG) operated on the island of La Palma by the Fundacion Galileo Galilei of the INAF (Istituto Nazionale di Astrofisica) at the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias.

The national facility capability for SkyMapper has been funded through ARC LIEF grant LE130100104 from the Australian Research Council, awarded to the University of Sydney, the Australian National University, Swinburne University of Technology, the University of Queensland, the University of Western Australia, the University of Melbourne, Curtin University of Technology, Monash University and the Australian Astronomical Observatory. SkyMapper is owned and operated by the Australian National University's Research School of Astronomy and Astrophysics. The survey data have been processed and provided by the SkyMapper Team at ANU. The SkyMapper node of the All-Sky Virtual Observatory (ASVO) is hosted at the National Computational Infrastructure (NCI). Development and support the SkyMapper node of the ASVO has been funded in part by Astronomy Australia Limited (AAL) and the Australian Government through the Commonwealth's Education Investment Fund (EIF) and National Collaborative Research Infrastructure Strategy (NCRIS), particularly the National eResearch Collaboration Tools and Resources (NeCTAR) and the Australian National Data Service Projects (ANDS).

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (<https://www.cosmos.esa.int/gaia>), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the Univer-

sity of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation.

This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration.

This paper includes data gathered with the 6.5-m Magellan Telescopes located at Las Campanas Observatory, Chile.

This research is based on observations made with the Galaxy Evolution Explorer, obtained from the MAST data archive at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS526555.

## DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

## REFERENCES

- Abbott T. M. C. et al., 2021, preprint([arXiv:2101.05765](https://arxiv.org/abs/2101.05765))
- Ahumada R. et al., 2020, *ApJS*, 249, 3
- Anderson T. W., 2003, *An Introduction to Multivariate Statistical Analysis*, 3 edn. Wiley, New York
- Bai Y., Liu J., Wang S., Yang F., 2019, *AJ*, 157, 9
- Baron D., 2019, preprint([arXiv:1904.07248](https://arxiv.org/abs/1904.07248))
- Bianchi L., Shiao B., Thilker D., 2017, *ApJS*, 230, 24
- Boutsia K. et al., 2020, *ApJS*, 250, 26
- Boutsia K. et al., 2021, *ApJ*, 912, 111
- Breiman L., 2001, *Mach. Learn.*, 45, 5
- Breiman L., Friedman J.H., Olshen R.A., Stone C.J., 1984, *Classification And Regression Trees*, Routledge & CRC Press
- Calderone G. et al., 2019, *ApJ*, 887, 268
- Carrasco D. et al., 2015, *A&A*, 584, A44
- Chambers K. C. et al., 2016, preprint([arXiv:1612.05560](https://arxiv.org/abs/1612.05560))
- Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2002, *J. Artif. Int. Res.*, 16, 321
- Colless M. et al., 2001, *MNRAS*, 328, 1039
- Cupani G. et al., 2021, *MNRAS*
- Fontanot F., Cristiani S., Monaco P., Nonino M., Vanzella E., Brandt W. N., Grazian A., Mao J., 2007, *A&A*, 461, 39
- Gaia Collaboration, 2016, *A&A*, 595, A1
- Gaia Collaboration, 2018, *A&A*, 616, A1
- Gaia Collaboration, 2021, *A&A*, 649, A1
- Lemaître G., Nogueira F., Aridas C. K., 2017, *J. Mach. Learn. Res.*, 18, 1
- Lyke B. W. et al., 2020, *ApJS*, 250, 8
- Pâris I. et al., 2018, *A&A*, 613, A51
- Pedregosa F. et al., 2011, *J. Mach. Learn. Res.*, 12, 2825
- Reis I., Baron D., Shahaf S., 2019, *AJ*, 157, 16
- Schindler J.-T. et al., 2019, *ApJ*, 871, 258
- Shen X., Hopkins P. F., Faucher-Giguère C.-A., Alexander D. M., Richards G. T., Ross N. P., Hickox R. C., 2020, *MNRAS*, 495, 3252
- Silva P., Cao L., Hayes W., 2018, *Galaxies*, 6, 95
- Skrutskie M. F. et al., 2006, *AJ*, 131, 1163
- Véron-Cetty M. P., Véron P., 2010, *A&A*, 518, A10
- Wolf C. et al., 2018, *Publ. Astron. Soc. Aust.*, 35, e010
- Wright E. L. et al., 2010, *AJ*, 140, 1868

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.