A&A 616, A97 (2018)
https://doi.org/10.1051/0004-6361/201833103

Astronomy & Astrophysics

Return of the features

Efficient feature selection and interpretation for photometric redshifts ⋆

A. D'Isanto[1,2], S. Cavuoti[3,4,5], F. Gieseke[6], and K. L. Polsterer[1]

[1] Astroinformatics Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
e-mail: antonio.disanto@h-its.org; kai.polsterer@h-its.org
[2] Zentrum für Astronomie der Universität Heidelberg, Astronomisches Rechen-Institut, Heidelberg, Germany
[3] Department of Physics "E. Pancini", University Federico II, via Cinthia 6, 80126 Napoli, Italy
[4] INAF – Astronomical Observatory of Capodimonte, via Moiariello 16, 80131 Napoli, Italy
[5] INFN – Section of Naples, via Cinthia 9, 80126 Napoli, Italy
e-mail: cavuoti@na.infn.it
[6] Machine Learning Group Image Section, Department of Computer Science, University of Copenhagen, Sigurdsgade 41, 2200 København N, Denmark

Received 26 March 2018 / Accepted 23 April 2018

ABSTRACT

Context. The explosion of data in recent years has generated an increasing need for new analysis techniques in order to extract knowledge from massive data-sets. Machine learning has proved particularly useful to perform this task. Fully automatized methods (e.g. deep neural networks) have recently gathered great popularity, even though those methods often lack physical interpretability. In contrast, feature based approaches can provide both well-performing models and understandable causalities with respect to the correlations found between features and physical processes.
Aims. Efficient feature selection is an essential tool to boost the performance of machine learning models. In this work, we propose a forward selection method in order to compute, evaluate, and characterize better performing features for regression and classification problems. Given the importance of photometric redshift estimation, we adopt it as our case study.
Methods. We synthetically created 4520 features by combining magnitudes, errors, radii, and ellipticities of quasars, taken from the Sloan Digital Sky Survey (SDSS). We apply a forward selection process, a recursive method in which a huge number of feature sets is tested through a k-Nearest-Neighbours algorithm, leading to a tree of feature sets. The branches of the feature tree are then used to perform experiments with the random forest, in order to validate the best set with an alternative model.
Results. We demonstrate that the sets of features determined with our approach improve the performances of the regression models significantly when compared to the performance of the classic features from the literature. The found features are unexpected and surprising, being very different from the classic features. Therefore, a method to interpret some of the found features in a physical context is presented.
Conclusions. The feature selection methodology described here is very general and can be used to improve the performance of machine learning models for any regression or classification task.

Key words. methods: data analysis – methods: statistical – galaxies: distances and redshifts – quasars: general

1. Introduction

In recent years, astronomy has experienced a true explosion in the amount and complexity of the available data. The new generation of digital surveys is opening a new era for astronomical research, characterized by the necessity to analyse data-sets that fall into the Tera-scale and Peta-scale regime. This is leading to the need for a completely different approach with respect to the process of knowledge discovery. In fact, the main challenge will no longer be obtaining data in order to prove or disprove a certain hypothesis, but rather to mine the data in order to find interesting trends and unknown patterns. The process of discovery will not be driven by new kinds of instrumentation to explore yet unobserved regimes, but by efficient combination and analysis of already existing measurements. Such an approach requires the development of new techniques and tools in order to deal with this explosion of data, which are far beyond any possibility of manual inspection by humans. This necessity will become urgent in the next years, when surveys like the Large Synoptic Survey Telescope (LSST; Ivezić et al. 2008), the Square Kilometer Array (SKA; Taylor 2008), and many others, will become available. Therefore, machine learning techniques are becoming a necessity in order to automatize the process of knowledge extraction from big data-sets. In the last decade, machine learning has proved to be particularly useful to solve astrophysical complex non-linear problems, both for regression (see for instance Hildebrandt et al. 2010; Bilicki et al. 2014; Cavuoti et al. 2015; Hoyle 2016; Beck et al. 2017) and classification tasks (see Mahabal et al. 2008; Rimoldini et al. 2012; Cavuoti et al. 2013a; D'Isanto et al. 2016; Smirnov & Markov 2017; Benavente et al. 2017). These techniques find nowadays many applications in almost all the fields of science and beyond (Hey et al. 2009). In the literature, two main machine learning branches can be found that deal with the selection of the most relevant information contained in the data. The first traditional way consists in the extraction and selection of manually crafted features, which

⋆ The three catalogues are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/qcat?J/A+A/616/A97

are theoretically more suitable to optimize the performance. In Donalek et al. (2013) feature selection strategies are compared in an astrophysical context.

The second option is using automatic feature selection models and became more popular in more recent years. For example, Athiwaratkun & Kang (2015) delegate this task to the machine by analysing the automatically extracted feature representations of convolutional neural networks. In convolutional neural networks, during the training phase the model itself determines and optimizes the extraction of available information in order to obtain the best performance. The challenge of feature selection is fundamental for machine learning applications, due to the necessity of balancing between overfitting and the curse of dimensionality (Bishop 2006), which arises when dealing with very high-dimensional spaces. Therefore a clever process of feature selection is needed to overcome this issue. In this setting, a different strategy was chosen for this work, in which a forward selection algorithm (Guyon & Elisseeff 2003) is adopted to identify the best performing features out of thousands of them. We decided to apply this procedure in a very important field: photometric redshift estimation. Due to the enormous importance that this measurement has in cosmology, great efforts have been lavished by the astronomical community on building efficient methods for the determination of affordable and precise photometric redshifts (Richards et al. 2001; Hildebrandt et al. 2008, 2010; Ball et al. 2008). Photometric redshifts are of extreme importance with respect to upcoming missions, for example the forthcoming Euclid mission (Laureijs et al. 2011), which will be based on the availability of photometric redshift measures, and the Kilo Degree Survey (KiDS; de Jong et al. 2017), which aims to map the large-scale matter distribution in the Universe, using weak lensing shear and photometric redshift measurements (Hildebrandt et al. 2016; Tortora et al. 2016; Harnois-Déraps et al. 2017; Joudaki et al. 2017; Köhlinger et al. 2017). Furthermore, photometric redshifts estimation is crucial for several other projects, the most important being the Evolutionary Map of the Universe (EMU; Norris et al. 2011), the Low Frequency Array (LOFAR; van Haarlem et al. 2013), Dark Energy Survey (Bonnett et al. 2016), the Panoramic Survey Telescope and Rapid Response System (PANSTARRS; Chambers et al. 2016), and the VST Optical Imaging of the CDFS and ES1 Fields (VST-VOICE; Vaccari et al. 2016). In light of this, we propose to invert the task of photometric redshift estimation. That is to say, having stated the possibility to determine the redshift of a galaxy based on its photometry, we want to build a method that allows us to investigate the parameter space and to extract the features to be used to achieve the best performance. As thoroughly analysed in D'Isanto & Polsterer (2018), the implementation of deep learning techniques is providing an alternative to feature based methods, allowing the estimation of photometric redshifts directly from images. The main concerns when adopting deep learning models are related to the amount of data needed to efficiently perform the training of the networks, the cost in terms of resources and computation time, and the lack of interpretability related to the features automatically extracted. In fact, deep learning models can easily become like magic boxes and it is really hard to assign any kind of physical meaning to the features estimated by the model itself. Therefore, a catalogue-based approach still has great importance, due to the gains in time, resources, and interpretability. In particular, this is true if a set of significant features is provided, in order to concentrate the important information with respect to the problem in a reduced number of parameters. Both methods, based on automatically extracted features or on selected features, constitute the starting point to build an efficient and performing model for redshift estimation, respectively. The topic of feature selection is a well-treated subject in the literature (see for example Rimoldini et al. 2012; Tangaro et al. 2015; Hoyle et al. 2015; D'Isanto et al. 2016). The forward selection approach we used (Gieseke et al. 2014) is meant to select between thousands of features generated by combining plain photometric features as they are given in the original catalogue. No matter what selection strategy is applied, the final results have to be compared to those obtained with the traditional features from the literature (D'Abrusco et al. 2007; Richards et al. 2009; Laurino et al. 2011) and with automatically extracted features. The aim is to find the subsets that give a better performance for the proposed experiments, mining into this new, huge feature space and to build a method useful to find the best features for any kind of problem. Moreover, we propose to analyse the obtained features, in order to give them a physical explanation and a connection with the processes occurring in the specific category of sources. Such an approach also demands a huge effort in terms of computational time and resources. Therefore, we need an extreme parallelization to deal with this task. This has been done through the intensive use of graphics processing units (GPU), a technology that is opening new doors for Astroinformatics (Cavuoti et al. 2013b; Polsterer et al. 2015; D'Isanto & Polsterer 2018), allowing the adoption of deep learning and/or massive feature selection strategies. In particular, in this work, the feature combinations are computed following Gieseke et al. (2014) and Polsterer et al. (2014), using a GPU cluster equipped with four Nvidia Pascal P40 graphic cards[1]. Likewise for Zhang et al. (2013), the k-Nearest-Neighbours (kNN; Fix & Hodges 1951) model is used, running recursive experiments in order to estimate the best features through the forward selection process. This choice has been done because the kNN model scales very well with the use of GPU, with respect to performance and quality of the prediction, as shown in Heinermann et al. (2013). In this way, for each run of the experiment, the most contributing features are identified and added to previous subsets. Thereby, a tree of feature groups is created that afterwards can be compared with the traditional ones. The validation experiments are performed using a random forest (RF) model (application in astronomy Carliles et al. 2010). We will show that this approach can strongly improve performance for the task of redshift estimation. The improvement is due to the identification of specific feature subsets containing more information and capable of better characterizing the physics of the sources. In the present work, we perform the experiments on quasar data samples extracted from the Sloan Digital Sky Survey Data Release 7 (SDSS DR7; Abazajian et al. 2009) and Data Release 9 (SDSS DR9; Ahn et al. 2012). The proposed approach is very general and could be also used to solve many other tasks in astronomy, including both regression and classification problems.

In Sect. 2 the methodology and models used to perform the experiments are described together with the statistical estimators used to evaluate the performance. The strategy adopted for the feature selection is also explained. Section 3 is dedicated to the data used and the feature extraction process. In Sect. 4 the experiments performed and the results obtained are described. Finally, in Sect. 5 the results are discussed in detail and in Sect. 6 some conclusions are drawn.

[1] https://images.nvidia.com/content/pdf/tesla/184427-Tesla-P40-Datasheet-NV-Final-Letter-Web.pdf

2. Methods

The main purpose of this work is to build an efficient method capable of generating, handling and selecting the best features for photometric redshift estimation, even though the proposed method is also able to deal with any other task of regression or even classification. We calculate thousands of feature combinations of photometric data taken from quasars. Then, a forward selection process is applied, as will be explained in more detail in the next sections. This is done to build a tree of best performing feature subsets. This method has to be considered as an alternative to the automatic features extraction used in D'Isanto & Polsterer (2018). Both methods can be useful and efficient, depending on the nature of the problem, and on the availability of data and resources. For this reason, the results obtained with both methods will be compared. The experimental strategy is based on the application of two different machine learning models and evaluated on the basis of several statistical tools. In the following these models, kNN and RF, are presented. The strategy used to perform the feature selection is then depicted in detail and we give a description of the statistical framework used for the experiments' evaluation and of the cross validation algorithm.

2.1. Regression models

As mentioned above, our method makes use of kNN and RF models, which are described in detail in the following subsections, while the details regarding the deep convolutional mixture density network (DCMDN) used to compare the results with an automatic features extraction based model can be found in D'Isanto & Polsterer (2018).

2.1.1. kNN

The kNN (Fix & Hodges 1951) is a machine learning model used both for regression and classification tasks (Zhang et al. 2013). This model explores the feature space by estimating the k nearest points (or neighbours) belonging to the training sample with respect to each test item. In our case the distance involved is calculated through a Euclidean metric. In the case of a regression problem (like redshift estimation), the kNN algorithm is used to find a continuous variable averaging the distances of the k selected neighbours. The efficiency of the algorithm is strongly related to the choice of the parameter k, which represents the number of neighbours to be selected from the training set. The best choice of this parameter is directly related to the input data, their complexity, and the way in which the input space is sampled. Clearly, the most simple case is a model with $k = 1$. In this case, a prediction equal to the target of the closest pattern in the training set is associated to each pattern. Increasing the k parameter could improve the precision of the model (this is due to the increasing generalization capability), but can also generate overfitting (Duda et al. 2000). In our experiments, the choice of the k parameter was part of the learning task by evaluating a set of possible values. The kNN is one of the simplest machine learning algorithms, but even if it could be outperformed by more complex models, it has the advantage of being very fast and in any case quite efficient. Another possible problem concerning the use of the kNN model is given by possible differences in the range of the input features. This could generate problems and misleading results in the estimation of distances in the parameter space. For this reason, all the features used in this work have been normalized using the min-max normalization technique (Aksoy & Haralick 2000).

2.1.2. Random forest

The RF (Breiman et al. 1984) is one of the most popular ensemble-based machine learning models, and could be used for regression and classification tasks (see Carliles et al. 2010, for an application to photometric redshift estimation). It is an ensemble of decision trees, where each tree is meant to partition the feature space in order to find the best split that minimizes the variance. Each decision tree is built by adding leaf nodes where the input data are partitioned with respect to a different chosen feature, repeating the process for all the possible choices of variables to be split. In case of a regression problem, the root mean square error (RMSE) is computed for each possible partition, and the partition which minimizes the RMSE is chosen. The RF averages the results provided by many decision trees, each trained on a different part of the training set through the bagging technique (Breiman 1996). This avoids overfitting due to single decision trees growing too deep. Moreover, the decision tree makes use of the bootstrapping technique (Breiman 1996) in order to increase the performance and stability of the method and reduce overfitting at the same time. This consists in giving, as input, a different random sub-sample of the training data to each decision tree. The RF uses the feature bagging during the training phase. This consists in selecting a random subset of features at each split. Bootstrapping and bagging help to avoid correlations between single decision trees, which could appear when training them on the same training set and in the presence of strong features selected multiple times.

2.2. Features selection strategy

The huge number of features evaluated, as described in Sect. 3, imposes the need to establish an efficient feature selection process. In fact, in order to estimate a subset of the best $f = 10$ features[2], starting with $r = 4520$ features, would imply, if we want to test all the possible combinations, the following number of experiments:

$$n = \frac{r!}{f! * (r - f)!} = 9.7 \times 10^{29}. \quad (1)$$

Assuming that a nonillion experiments are too many to be performed, a more efficient approach had to be chosen. Therefore, we decided to apply a forward selection process (Mao 2004) as described in the following. The number of features used for the experiment was iteratively increased. In other words, to select the first best feature a kNN model for each of the $r = 4520$ features was trained in a one-dimensional feature space. Due to the memory limitations of the hardware architecture used, the feature selection was done by performing 100 kNN experiments, selecting for each of them a random subset of 20 000 data points and using a five-fold cross validation (see Sect. 2.4 for more details). The repeated experiments on different training samples were meant to generate statistics of the features in order to identify the most frequently selected ones. This was done to minimize the biases introduced by the random extraction of the training data. Since 100 runs were performed, sometimes more than one feature was selected. The basic idea behind the proposed strategy is to select a limited number of best-performing features per step. The number of features which were actually selected were chosen by evaluating the occurrence of each of them as the best feature in all of the 100 runs. Therefore, for

[2] The reason for selecting 10 features is discussed in Sect. 4.3 and Fig. 6.

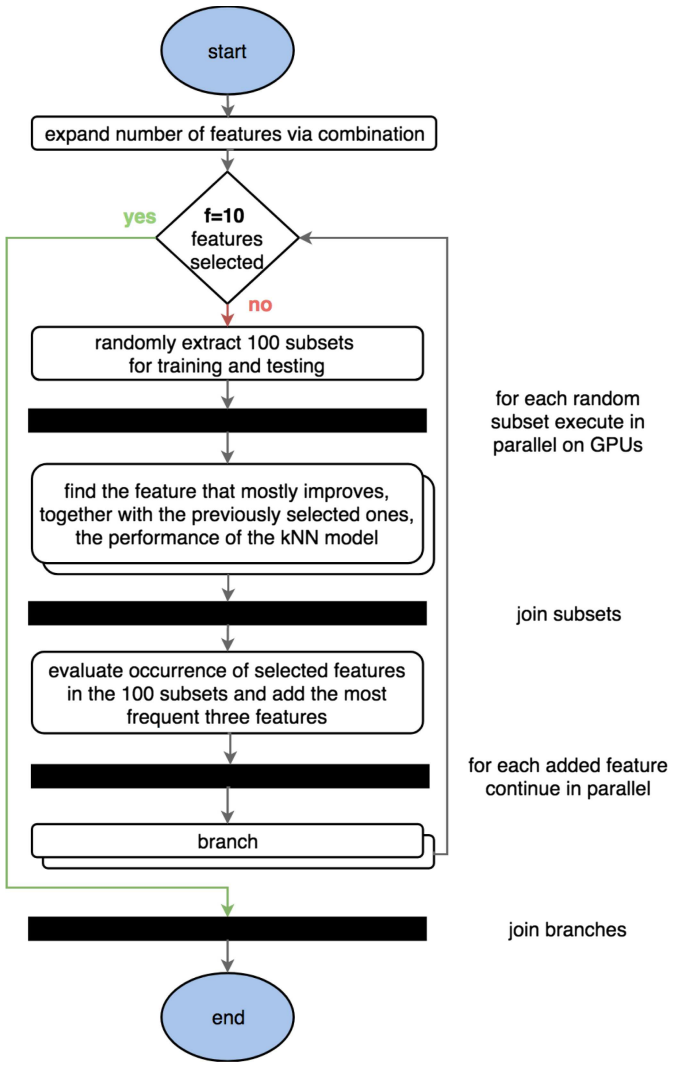


Fig. 1. Workflow used to generate tree structure. The black boxes represent states where multiple operations are started in parallel or parallel operations are joined. The iteration is stopped when each branch of the tree has a depth of 10. A five-fold cross validation is applied for every model evaluation step.

each iteration a minimum of one and a maximum of three features were selected. After choosing the best features, they were fixed and the next run was performed in order to choose the subsequent features. This method was iterated until the tenth feature was selected. A tree with a maximum branching number of three was derived, because in every step a maximum number of three features that best improve the model were chosen. Each branch can be seen as a set of best-performing-feature combinations. The necessity of performing a high number of experiments on different data subsets is caused by the slightly varying behaviour of the kNN model with respect to different input patterns. The whole workflow is summarized in Fig. 1. The cross validation, moreover, was used in order to further reduce any risk of overfitting.

2.3. GPU parallelization for kNN

The feature selection is done by parallelizing the experiments on a GPU cluster. The massive use of GPUs proved to be mandatory in order to deal with such an amount of data, features, k values, and runs on randomly sampled data-sets. Following Heinermann et al. (2013) and Gieseke et al. (2014), the kNN algorithm has been parallelized by using GPUs. Typically, GPU-based programs are composed by a host program running on central processing unit (CPU) and a kernel program running on the GPU itself, which is parallelized on the GPU cores in several threads or kernel instances. This scheme is particularly adapted to kNN models, due to the advantages obtained by parallelizing matrix multiplications. In the code used for this work (Gieseke et al. 2014) the calculation is performed by generating matrices containing the distances of the selected features from the query object. This calculation is entirely performed on the GPU, while the CPU is mainly used for synchronization and for updating a vector containing the selected features at every step. The approach based on this method proved to speed up the calculation by a factor of ∼150. We modified the given code to start the selection process with a given set of already selected features. This was done to enable the generation of the feature trees based on 100 random subsets.

2.4. Statistical estimators and cross validation

The results have been evaluated using the following set of statistical scores for the quantity $\Delta z = (z_{\rm spec} - z_{\rm phot})/(1 + z_{\rm spec})$ expressing the estimation error[3] on the objects in the blind validation set:

- bias: defined as the mean value of the normalized residuals Δz;
- RMSE: root mean square error;
- NMAD: normalized median absolute deviation of the normalized residuals, defined as $\mathrm{NMAD}(\Delta z) = 1.48 \times \mathrm{median}(|\Delta z_i - \mathrm{median}(\Delta z)|)$;
- CRPS: the continuous rank probability score (Hersbach 2000) is a proper score to estimate how well a single value is represented by a distribution. It is used following D'Isanto & Polsterer (2018).

The prediction of redshifts in a probabilistic framework has many advantages. The ability of reporting the uncertainty is the most important one to mention. In order to correctly evaluate the performance of the features in a probabilistic setting, the CRPS was added to the set of scores. By using the RF as a quantile regression forest and fitting a mixture of Gaussians to the predictions of the ensemble members, a probability distribution can be generated and the CRPS can be calculated. The DCMDN, by definition, predicts density distributions that are represented by their mean when calculating the scores used for point estimates.

As stated before, all the indicators are then averaged on the k folds of the cross validation. Through this approach, the standard deviation is also obtained as a measure of the error on each statistical estimator. We do not report those values as the errors were small enough to be considered negligible. Cross validation (Kohavi 1995) is a statistical tool used to estimate the generalization error. The phenomenon of overfitting arises when the model is too well adapted to the training data. In this case, the performance on the test set will be poor as the model is not general enough. A validation set is defined, in order to test this generalization of the model, with respect to the training data, on an unseen and omitted set of data. In particular, cross validation becomes necessary when dealing with small training sets or high-dimensional feature spaces.

[3] We note that Δz denotes the normalized error in redshift estimation and not the usually used plain error.

In this kind of approach, the data-set is divided into k subsets and each of them is used for the prediction phase, while all the $k-1$ subsets constitute the training set. The training is then repeated k times, using all the subsets. The final performance is obtained by averaging the results of the single folds and the error on the performance is obtained by evaluating the standard deviation of the results coming from the different folds. In this work, we adopt a k-fold cross validation approach, with $k = 5$ for the kNN experiments and $k = 10$ for the RF experiments.

3. Data

In the following subsections the details about the data-set used and the feature combinations performed for the experiments are outlined.

3.1. Data-sets

The experiments are based on quasar data extracted from the SDSS DR7 (Abazajian et al. 2009) and SDSS DR9 (Ahn et al. 2012). Three catalogues have been retrieved for the experiments. Moreover, images for the DCMDN experiments have been downloaded making use of Hierarchical Progressive Survey (HiPS; Fernique et al. 2015).

Catalogue DR7a. Catalogue DR7a is the most conservative with respect to the presence of bad data or problematic objects. It is based on DR7 only, with clean photometry and no missing data; the query used is reported in Appendix D. Furthermore, to be more conservative, we checked the spectroscopic redshifts in two different data releases (9 and 12) and we decided to cut all the objects with a discrepancy in $z_{\rm spec}$ not fulfilling the given criteria

$|z_{\rm DR7} - z_{\rm DR9}| < 0.01$, and
$|z_{\rm DR7} - z_{\rm DR12}| < 0.01$, and
$|z_{\rm DR12} - z_{\rm DR9}| < 0.01$.

The final catalogue contains 83 982 objects with a spectroscopically determined redshift.

Catalogue DR7b. Catalogue DR7b has been obtained using the same query used for Catalogue DR7a, but removing the image processing flags. This has been done in order to verify if the presence of objects previously discarded by the use of these flags could affect the feature selection process. The catalogue has been cleaned by removing all the objects with NaNs and errors bigger than a value of one, ending with a catalogue containing 97 041 objects.

Catalogue DR7+9. Catalogue DR7+9 has been prepared mixing quasars from DR7 and DR9 in order to perform the feature selection with a different and more complete redshift distribution. The difference in the redshift distribution of the two catalogues can be seen from the histogram in Fig. 2. The catalogue has been cleaned with the same procedure adopted for Catalogue DR7b and the common objects between DR7 and DR9 have been used only once. This produced a catalogue of 152 137 objects. In the following sections, the results obtained with this catalogue are discussed in depth.

3.2. Classic features

In classic redshift estimation experiments for quasars and galaxies, as can be found in the literature (e.g. D'Abrusco et al. 2007), for SDSS data colours are mainly used as features. To be comparable, we decided to use a set of ten features as our benchmark feature set. Colours of the adjacent filterbands for the point spread function (PSF) and model magnitudes are used together with the plain PSF and model magnitudes. In SDSS, the model magnitudes are the best fitting result of an exponential or de Vaucouleurs model. All $\mathsf{Classic}_{10}$ features can be found in the first column of Table 2.

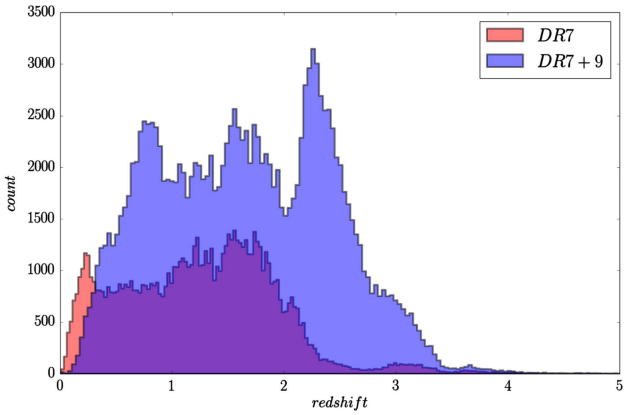


Fig. 2. Histogram showing the redshift distribution of the catalogues with objects from DR7 only and DR7 plus DR9. The distribution for the catalogue DR7b is not reported here because the difference with respect to catalogue DR7a is practically negligible.

Table 1. Types of features downloaded from SDSS and their combinations in order to obtain the final catalogue used for the experiments.

	Magnitudes	σ	Radii	Ellipticities
	modelMag / Extinction	✓	devRad	devAB
	petroMag / Extinction	✓	expRad	expAB
	psfMag / Extinction	✓	petroRad	
	devMag / Extinction	✓	petroR50	
	expMag / Extinction	✓	petroR90	
Plain	25 + 25 dereddened	25	25	10
Combined	1 225 Differences 2450 Ratios	300 Pairs	300 Differences	45 Differences 90 Ratios
Total 4520	3725	325	325	145

Notes. The number of each feature type is given alongside with the final number of synthetically derived features.

3.3. Combined features

For each of the three catalogues, the features concerning magnitudes and their errors, radii, ellipticities, and extinction are retrieved. An overview of the features is shown in Table 1. Magnitudes that have been corrected for extinction are denoted with an underline indicating that, for example, $\underline{u_{\rm model}}$ is equivalent to $u_{\rm model} - u_{\rm extinction}$. The parameter space has been enriched by performing several combinations of the original features (Gieseke et al. 2014). A similar feature generation approach was applied also in Polsterer et al. (2014) but with a limited set of plain features and combination rules. In other words, the magnitude features were combined obtaining all the pairwise differences and ratios, both in the normal and dereddened version. The errors on the magnitudes have been composed taking their quadratic sums. Finally, radii and ellipticities have been composed through pairwise differences with ratios only for the ellipticities. The final catalogue consists of 4520 features for each data item. It has to be noted that the $\mathsf{Classic}_{10}$ features are of

course included in this set of features. In Table 1, the types and amounts of the features obtained following this strategy are specified. As appears from the table, the feature combinations can be divided into several groups:

- simple features: magnitudes, radii, and ellipticities as downloaded from the SDSS database;
- differences: pairwise differences of the simple features; colour indexes are a subset of this group utilizing only adjacent filters;
- ratios: ratios between the simple features; an important subset of this group is the one containing ratios between different magnitudes of the same filter; we will define this subset as photometric ratios;
- errors: errors on the simple features and their propagated compositions.

As we will see in the following, the ratios group, and its subgroup, the photometric ratios, are particularly important for the redshift estimation experiments.

4. Experiments and results

The feature selection was performed applying the forward selection strategy, as described in Sect. 2.2, on the three catalogues. The verification of the resulting feature sets was performed using the RF. This algorithm is widely used in literature, and therefore the results obtained here can be easily compared to those with different feature selection strategies.

In addition, experiments using the classic features were performed, in order to compare their performances with the proposed selected features. Already at an early stage of the experiments, it turned out that only four selected features are sufficient to achieve a performance comparable to classic features. Therefore the scores are always calculated separately for the full set of ten selected features (Best_{10}) and the first four (Best_4) features only. To compare the results with a fully automated feature extraction and feature selection approach, a DCMDN was also used for the experiments.

It has to be noted that in some cases the same features sets have been found but exhibiting a different ordering. In these cases, all the subsets have been kept for the sake of correctness. In the next subsections the three experiments and the corresponding results are shown. The two experiments with Catalogue DR7a and DR7b are designed to provide results that are comparable to the literature. For a scientifically more interesting interpretation, the less biased, not flagged, and more representative Catalogue DR7+9 was used for the main experiment. Therefore, only the results and performances of the first two experiments are given in a summarized representation, reserving more space for a detailed description of Experiment DR7+9. Further details concerning the results obtained with Catalogue DR7a and DR7b are shown in Appendix A.

4.1. Experiment DR7a

The feature selection on the Catalogue DR7a produced 22 subsets of ten features each. Only 20 features, of the initial 4520, compose the tree. The three features,

- $\underline{g_{psf}}/\underline{u_{model}}$
- $\underline{i_{psf}}/\underline{z_{model}}$
- z_{model}/z_{psf},

appear in all the possible branches. For all presented feature sets, the RF experiments were performed. The best performing ten features are indicated in the second column of Table 2 (DR7a subset) in the order of their occurrence. The performances are compared with the results of the $\mathsf{Classic}_{10}$ features presented in the first column of the same table. A summary of the most important results is shown in the first section of Table 3. As shown in Table 3, the experiment with the Best_{10} subset outperforms the experiment with the $\mathsf{Classic}_{10}$ features with respect to all the statistical scores.

Moreover, in Table 3 the results obtained using the DCMDN are shown in order to compare the predictions with a model based on automatic features selection. The DCMDN model automatically extracts 65536 features from images in the five filters *ugriz* of size 16×16 pixel^2. This model is meant to generate probability density functions (PDFs) in the form of Gaussian mixtures instead of point estimates. Therefore, in order to calculate the scores, the weighted mean of every PDF with respect to the mixture components has been estimated. As shown in the table, the performance is superior with respect to the $\mathsf{Classic}_{10}$ features and the Best_4 subset, but it is outperformed by the Best_{10} subset of features. The performances of these four sets have been compared using the CRPS score, as reported in the left section of Table 4. Those results are consistent with the

Table 2. Classic and best feature subsets obtained by the feature selection process of the experiments on the three catalogues.

$\mathsf{Classic}_{10}$	DR7a Best_{10}	DR7b Best_{10}	DR7+9 Best_{10}
r_{psf}	i_{psf}/i_{model}	i_{psf}/i_{model}	i_{petro}/i_{psf}
r_{model}	$\underline{g_{psf}}/\underline{u_{model}}$	$\underline{g_{psf}}/\underline{u_{model}}$	$\underline{g_{psf}} - \underline{u_{model}}$
$u_{psf} - g_{psf}$	$\underline{r_{psf}}/\underline{i_{model}}$	$\underline{r_{psf}}/\underline{i_{model}}$	$\underline{i_{exp}}/\underline{r_{psf}}$
$g_{psf} - r_{psf}$	$\underline{i_{dev}}/\underline{i_{psf}}$	$\underline{i_{dev}}/\underline{i_{psf}}$	$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{r_{dev}}}$
$r_{psf} - i_{psf}$	$\underline{r_{psf}}/\underline{g_{model}}$	$\underline{z_{psf}}/\underline{i_{model}}$	$\underline{r_{psf}}/\underline{g_{exp}}$
$i_{psf} - z_{psf}$	$\underline{i_{psf}}/\underline{z_{model}}$	$\underline{r_{psf}}/\underline{g_{exp}}$	$\underline{i_{psf}}/\underline{z_{model}}$
$u_{model} - g_{model}$	$r_{psf} - r_{petro}$	$r_{psf} - r_{petro}$	$i_{psf} - i_{dev}$
$g_{model} - r_{model}$	$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{exp}}}$	$i_{psf} - i_{petro}$	r_{petro}/r_{psf}
$r_{model} - i_{model}$	z_{model}/z_{psf}	$\underline{z_{model}}/\underline{z_{psf}}$	$\underline{i_{psf}} - \underline{r_{model}}$
$i_{model} - z_{model}$	$i_{psf} - i_{petro}$	$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$	z_{exp}/z_{psf}

Notes. After the selection process, the RF was used to identify the feature branches of the corresponding trees that show the best performance.

Table 3. Summary of the scores obtained with the RF and DCMDN models in the three experiments.

Exp	Set	# Features	Mean	RMSE	NMAD
DR7a	$\mathsf{Classic}_{10}$	10	−0.024	0.163	0.051
	Best_4	4	−0.023	0.163	0.080
	Best_{10}	10	−0.014	0.124	0.044
	DCMDN	65 536	−0.020	0.145	0.043
DR7b	$\mathsf{Classic}_{10}$	10	−0.030	0.180	0.059
	Best_4	4	−0.027	0.183	0.087
	Best_{10}	10	−0.019	0.145	0.050
	DCMDN	65 536	−0.024	0.171	0.032
DR7+9	$\mathsf{Classic}_{10}$	10	−0.033	0.207	0.073
	Best_4	4	−0.032	0.206	0.100
	Best_{10}	10	−0.023	0.174	0.060
	DCMDN	65 536	−0.027	0.184	0.037

Notes. The DCMDN automatically extracted 65 536 features for each experiment. The resulting scores are also given.

Table 4. Table showing the performance of the different feature subsets with respect to the CRPS score for the three catalogues.

DR7a	CRPS	DR7b	CRPS	DR7+9	CRPS
$\mathsf{Classic}_{10}$	0.110	$\mathsf{Classic}_{10}$	0.131	$\mathsf{Classic}_{10}$	0.167
Best_4	0.154	Best_4	0.172	Best_4	0.203
Best_{10}	0.089	Best_{10}	0.106	Best_{10}	0.140
DCMDN	0.099	DCMDN	0.124	DCMDN	0.146

previously found results. A detailed listing of the results is given in Appendix A with the individual feature tree being visualized as a chord diagram (Krzywinski et al. 2009).

4.2. Experiment DR7b

In the experiment performed with Catalogue DR7b, the proposed model selected 26 features generating 41 subset combinations. Only the following two features appear in all the subsets:

- $i_{\mathrm{psf}}/i_{\mathrm{petro}}$
- $\underline{g_{\mathrm{psf}}/u_{\mathrm{model}}}$.

From the RF validation runs, the subset reported in the third column of Table 2 (DR7b) produces the best performance. The most important results are shown in the second section of Table 3, in which the results obtained with the previous experiment (DR7a) are confirmed. This is valid considering both the RMSE and the CRPS indicators. The CRPS is shown in the middle section of Table 4. Therefore, the performance given using the Best_{10} subset is superior to that using the $\mathsf{Classic}_{10}$ features. The DCMDN model is outperformed too. Several features can be found in both experiments with catalogues DR7a and DR7b and the general structure of the tree between the two experiments is comparable. Therefore, the exclusion of photometric flags seems not to affect substantially the global process of feature selection. It can be noticed, however, that the general performance degrades. This is due to the increased presence of objects characterized by a less clean photometry. The detailed feature selection results for this experiment and the chord diagram are also shown in Appendix A.

4.3. Experiment DR7+9

The feature selected from the Catalogue DR7+9 are shown in Table 5. In Fig. 3 a chord diagram is given to visualize the structure of the individual subsets. In this experiment the model selected 14 individual features grouped in nine subsets. Due to the different redshift distribution, different features are selected with respect to the previous experiments. The following six features are in common between all the subsets:

- $i_{\mathrm{psf}} - i_{\mathrm{dev}}$
- $\underline{i_{\mathrm{psf}}/z_{\mathrm{model}}}$
- $\underline{g_{\mathrm{psf}} - u_{\mathrm{model}}}$
- $\underline{i_{\mathrm{petro}}/i_{\mathrm{psf}}}$
- $\underline{r_{\mathrm{psf}}/g_{\mathrm{exp}}}$
- $\underline{i_{\mathrm{exp}}/r_{\mathrm{psf}}}$.

The best performing subset is shown in the fourth column of Table 2 (DR7+9 subset), while in the third section of Table 3 results obtained with the RF experiments are given. Moreover, in the right section of Table 4 the results with the CRPS as indicator are provided. For this experiment we also report the z_{spec} versus z_{phot} plots in Fig. 4. This classical representation visualizes the better concentration along the ideal diagonal for both the Best_{10} features as well as the features derived through the DCMDN. When using the features in a probabilistic context, the better performance with respect to outliers of the DCMDN can be observed (Fig. 5). The probability integral transform (PIT Gneiting et al. 2005) histograms show very similar performances for all the feature sets that were selected. Besides the outliers, the estimates are sharp and well calibrated, exhibiting no difference in comparison to the results generated with the $\mathsf{Classic}_{10}$ features. This is a good indication that no systematic biases were added through the selection process.

Finally, the performance obtained with the $\mathsf{Classic}_{10}$ features is compared to the ones achieved with the Best_{10} features in a cumulative way. In Fig. 6, the RMSE and the NMAD are plotted with respect to the number of features of the Best_{10} set that were used. This is important in order to show that starting with the 4th feature, the model reaches already a performance comparable with the $\mathsf{Classic}_{10}$ features. Originating in the random data sampling during the selection process, the resulting different feature subsets do not show obvious differences in the quality of the final performance. In fact, the results obtained with the Best_{10} subset are far better with respect to the performance obtained using the $\mathsf{Classic}_{10}$ features and the DCMDN. This is a confirmation of the quality and strength of the proposed method.

5. Discussion

In the following subsections we discuss in detail the features found with the proposed method, the improvement in performance of the photometric redshift estimation models in comparison to the classic features, and the physical interpretation of the selected features.

5.1. Features

The results obtained from the feature selection process for the three experiments demonstrate that most of the information can be embedded in a limited number of features with respect to the initially generated amount of pairwise combinations. The following four features have been selected and are in common between all the three experiments:

- $r_{\mathrm{psf}} - r_{\mathrm{petro}}$
- $i_{\mathrm{psf}} - i_{\mathrm{dev}}$
- $\underline{i_{\mathrm{psf}}/z_{\mathrm{model}}}$
- $\underline{r_{\mathrm{psf}}/i_{\mathrm{exp}}}$.

This is a clear indicator that those features contain some essential information. Besides noting that they encode spatial and morphological characteristics, we have no clear explanation. Some features, as will be analysed in the next sections, can be clearly connected to physical processes occurring in the considered sources. Other features are instead much harder to interpret, which demands a deeper analysis in the future. Given that photometric redshifts are just used as a testbed for the proposed methodology, such an analysis is beyond the scope of this work. A quick and shallow inspection of the features exhibits that the ratios and differences play a major role. In Table 5 for the experiment DR7+9 the different groups of features are highlighted using different background patterns. This visually summarizes the dominant occurrence of those groups. In fact, all the features except the 4th (errors) belong to one of these two groups. Moreover, the individual branches of feature sets employ a feature of the same group for the first seven positions, showing a great

Table 5. Detailed feature branches obtained from the feature selection for the DR7+9 experiment.

id	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1	i_{petro}/i_{psf}	$g_{psf} - u_{model}$	i_{exp}/r_{psf}	$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{r_{dev}}}$	r_{psf}/g_{exp}	i_{psf}/z_{model}	$i_{psf} - i_{dev}$	r_{petro}/r_{psf}	$i_{psf} - i_{dev}$	$z_{psf} - z_{model}$
2*										z_{exp}/z_{psf}
3									$i_{psf} - r_{model}$	
4										$i_{psf} - i_{dev}$
5								$z_{psf} - z_{model}$	r_{petro}/r_{psf}	
6				$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{r_{dev}}}$				r_{petro}/r_{psf}	$i_{psf} - i_{dev}$	$z_{psf} - z_{model}$
7									$z_{psf} - z_{model}$	
8								$z_{psf} - z_{model}$	r_{petro}/r_{psf}	$i_{psf} - i_{dev}$
9									$r_{psf} - r_{petro}$	

Notes. The 2nd branch, indicated with the $*$ symbol, is the best performing subset with respect to the experiments using the RF. The ratios and photometric ratios are indicated, respectively, with vertical lines and dots. The differences are with horizontal lines and the errors are with north west lines. The colour code for the features is the same as shown in the chord diagram in Fig. 3.

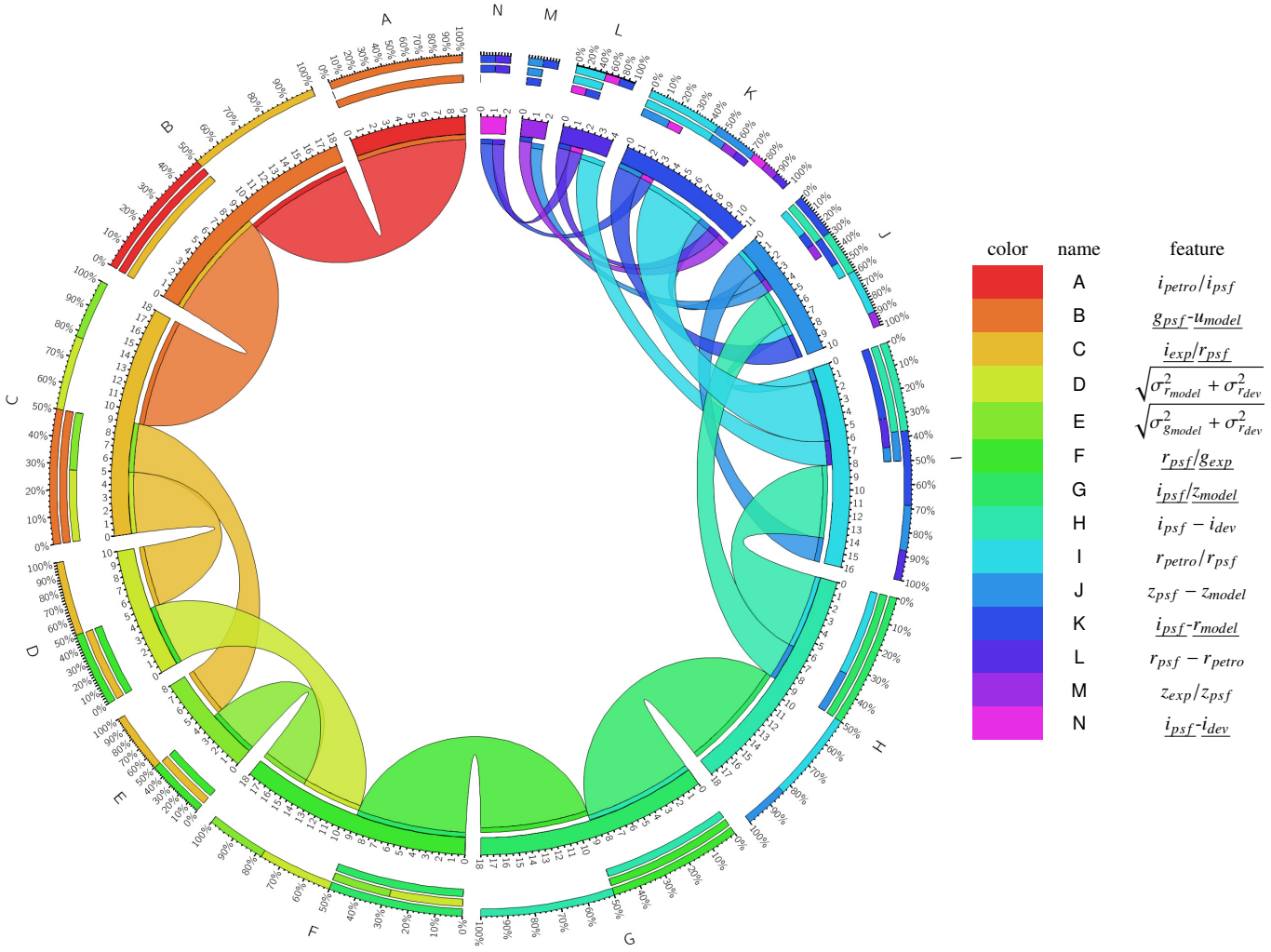


Fig. 3. Chord diagram of the features derived in Experiment DR7+9. Every feature is associated to a specific colour, and starting from the first feature A it is possible to follow all the possible paths of the tree, depicting the different feature subsets. Ordered from outside to inside, the external arcs represent the occurrences of a particular feature: the total percentage of the individual connections, the numbers and sources of connections entering, and the numbers and targets of connections exiting. (Note the branches splitting in feature C and re-joining in feature F).

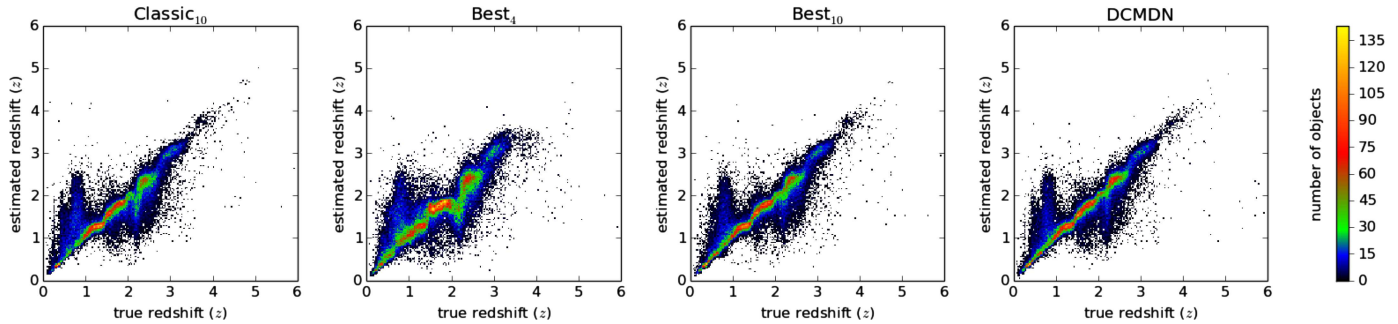


Fig. 4. Comparison of the spectroscopic (true) redshifts (z_{spec}) against the photometrically estimated redshifts (z_{phot}) of the different feature sets in experiment DR7+9.

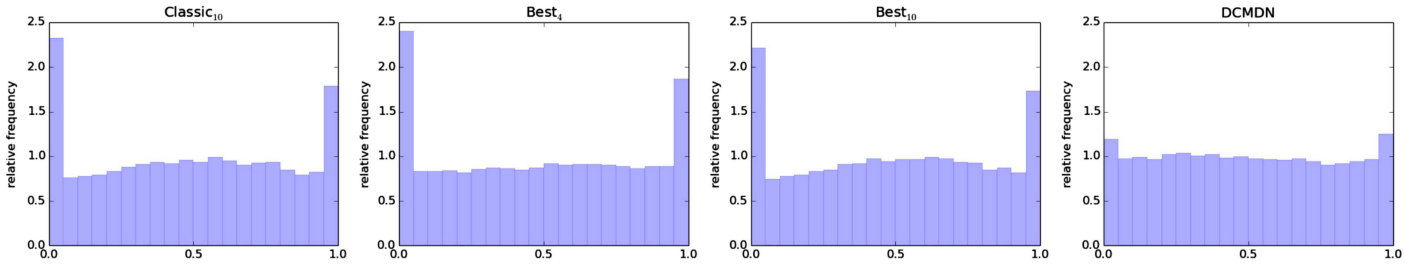


Fig. 5. PIT histograms for experiment DR7+9 for the different features sets, as shown in Table 4. Except the PIT of the DCMDN, all other feature sets generate results with significant outliers at the extrema.

stability in the composition of the branches. The experiment based on the DR7+9 catalogue generates a much less complex structure of the tree of feature sets with respect to experiments DR7a and DR7b. Fewer branches and a reduced number of features are selected. Reasons for this behaviour are the more complete redshift distribution of catalogue DR7+9 with respect to the other two and the improvement in SDSS photometry from DR7 to DR9. This drives the model to find the required information in a reduced number of efficient features. The analysis of the tree composition and features distribution can be done following the chord diagram shown in Fig. 3. The chord diagram is an optimal visualization tool for the description of a complex data structure. In this diagram, every feature is associated to a specific colour, and starting from the first feature (A) it is possible to follow all the possible paths of the tree, depicting the different feature subsets. Ordered from outside to inside, the external arcs represent the occurrences of a particular feature: the total percentage of the individual connections, the numbers and sources of connections entering, and the numbers and targets of connections exiting. Therefore, the chord diagram, coupled with Table 5, gives a clear description of the structure and composition of the tree of features. In addition, in Table 5 the same colour code as in the chord diagram is adopted, to identify the features and their distribution. The chord diagram clearly visualizes that the feature trees split at feature C and later rejoin at feature F. In comparison to the chord diagram obtained for Experiment DR7+9, the two chord diagrams for experiments DR7a and DR7b (see Appendix A) immediately visualize the higher complexity of those trees. From Fig. 3 and Table 5 it appears that, apart from a few exceptions, the selected features follow a precise scheme. No classic colour indexes or any of the Classic_{10} features have been chosen, while only differences between different magnitudes of the same band or differences between different type of magnitudes play a certain role. The ratios have been all selected in the extinction-corrected version, except for the subcategory of the photometric ratios. This can be understood considering that the latter are ratios between magnitudes of the same filter where the contribution of the extinction correction tends to cancel out.

Another relevant aspect in experiment DR7+9 is that all the 15 features in the tree are exclusively a composition of magnitudes and their errors. Neither radii nor ellipticities have been chosen during the selection process. As only quasars have been used in the experiments, this introduces a bias to the selection process in favour of magnitudes and against shape-based features. This is a clear indication that just the magnitudes are required to describe the objects and explore the parameter space in the setting of photometric redshift estimation. Although photometric ratios are shape-related parameters, they express the ratio between the centred and the extended part of a component that can be interpreted as flux of the hosting galaxy. Therefore, here a bias introduced by using quasars for the experiments cannot be observed.

It is remarkable that photometric errors are selected as features, given that there is no obvious physical relation between the redshift of the considered objects and the measurement errors reported by the photometric pipeline of SDSS. Therefore it is important to consider how errors are derived in the SDSS, based on flux measurements (Lupton et al. 1999). Magnitude errors quantify the discrepancy between the fitted photometric model (psf, model, petrosian, etc.) and the observed pixel-wise distribution of spatially correlated fluxes with respect to the applied noise model. Therefore, it is evident that the errors on the single magnitudes appear to be larger for fainter objects, a physical property that is directly correlated to distance. In addition, the deviation of spatial flux distributions from the applied spatial photometric models are good morphological indicators; for example, the shape and size of the hosting galaxy are correlated with redshift. The workflow adopted is able to capture these dependencies, selecting a composition of errors as an important feature of the best set.

Even though 4520 features were synthetically created by combining base features, only 15 were selected in experiment

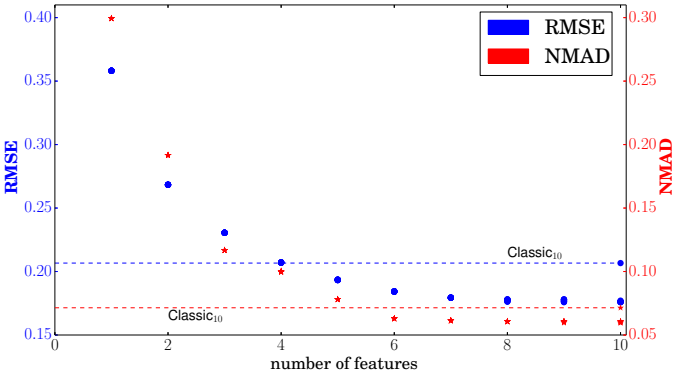


Fig. 6. Comparison of model performance with regard to the number of used features. The root mean square error and normalized median absolute deviation of the results from the DR7+9 RF experiments are presented. As reference line the performance achieved with the $\mathrm{Classic}_{10}$ features is shown. As it can be seen, from the fourth feature on, the performance of the subsets outperforms the $\mathrm{Classic}_{10}$ features. After the ninth feature, the improvement settles. When adding many more features, the performance will start to degrade.

DR7+9 (19 and 26 for experiments DR7a and DR7b, respectively). Furthermore, some features encode the same type of information with just subtle differences in composition. It is remarkable that every feature that is built on magnitudes incorporates a PSF magnitude. Moreover, the model and exponential magnitude in the SDSS are related[4], with the model magnitude being just the better fitting model when comparing an exponential and a de Vaucouleurs profile. In the first stages of the selection process, the proposed algorithm does not select differing branches but identifies essential features to produce good results when photometrically estimating redshifts. These observations are also valid for the results found in experiments DR7a and DR7b.

5.2. Comparison of performance

Using the RF, the validation experiments were carried out on every feature set. The second subset, indicated as Best_{10}, gave a slightly better performance than the others. Even though we would not consider this as a substantial effect, we decided to choose this as our reference set. It can be noticed from Fig. 6 that from the 4th feature on, every subset delivers a performance comparable to the performance of all ten features in the $\mathrm{Classic}_{10}$ set, with respect to the RMSE. Consistently, the use of more than four features outperforms the $\mathrm{Classic}_{10}$, independently of the subset used. Adding more features improves further the performance and the trend becomes asymptotic around the 9th feature. At a certain point, adding many more features results in a degradation of the redshift estimation performance. After the 8th feature, the contribution is of a minor nature. Just to have a fair comparison to the $\mathrm{Classic}_{10}$ features, we decided to pick the same number of ten features, even though a smaller number is sufficient to outperform the $\mathrm{Classic}_{10}$ features. The performance improvement is evident seeing the results reported in Table 3 and Fig. 4. It is important to note that the CRPS results (Table 4) confirm the performance shown with respect to the other scores. When predicting PDFs instead of point estimates, the PIT histograms (Fig. 5) indicate the DCMDN as the best calibrated model. This result is reasonable because the DCMDN is the only model trained using the CRPS as loss function, which is focused on the PDFs calibration. The kNN and the RF are instead based on the optimization of point estimates using the RMSE. Therefore, the calibration of the PDFs estimated using the DCMDN is superior. The use of such a probabilistic model is helpful to handle the presence of extreme outliers, since it is not based on the minimization of the RMSE, as discussed in D'Isanto & Polsterer (2018). The usage of PDFs allows us to identify objects with an ambiguous redshift distribution, while in a point estimation scenario, where just the mean of such a distribution would be considered, the estimates of those objects would result in extreme outliers.

[4] http://classic.sdss.org/dr7/algorithms/photometry.html#mag_model

Table 6. Cross experiments performed with the RF, using the Best_{10} sets obtained from every experiment with all the three catalogues.

Exp.	Catalogue DR7a	Catalogue DR7b	Catalogue DR7+9
DR7a	0.124	0.146	0.176
DR7b	0.125	0.145	0.176
DR7+9	0.124	0.147	0.174

Notes. The results are expressed using the RMSE. It can be noticed the negligible difference of performance, for every catalogue, independently from the feature set used.

Six features of the best subset are ratios of different magnitudes. Three of them are plain ratios, while three are photometric ratios. Analysing the fourth column of Table 2, it appears that one of the components of these features is always a PSF magnitude, coupled with a model, petro, or exp magnitude. Therefore, from the analysis of the results obtained, we can state that the reason for the performance improvement is not in the choice of some specific features, or in a particular subset of features, but in their type and in the combination of certain groups.

All these aspects are clear indicators to demonstrate the following two conclusions. The proposed method is highly stable, enabling us to derive subsets of features that are equivalently well-performing and similar, based on a common structure. In this sense, the improvement with respect to the use of $\mathrm{Classic}_{10}$ features is clear. In order to prove the robustness of the proposed method, we performed some experiments using for each data-set the Best_{10} features obtained with the other two catalogues, as shown in Table 6, and the results were almost as good as in the other cases. The method captures the inherent structure of the physical properties of the sources, which is essential to provide good photometrically estimated redshifts for quasars.

5.3. Physical interpretation

In contrast to deep learning models, feature-based approaches have the advantage of allowing an interpretation in a physical context. Therefore the features selected by our approach are discussed in the following. By analysing the importance of each feature of the Best_{10} set in smaller redshift bins, the contribution of certain spectral features can be understood. In Fig. 7 the importance is presented for sliding bins of $\Delta z = 0.2$ based on the Gini index (Breiman et al. 1984). The Gini index is used in the RF to perform the segmentation of the parameter space orthogonally to its dimensions at every node. As all ten features contribute individually, the total contribution is normalized to one and the individual lines are presented in a cumulative way. The relative importance of each feature clearly does not reflect their ordering, as they have been assembled by a forward feature selection algorithm. In particular, the first feature of the best

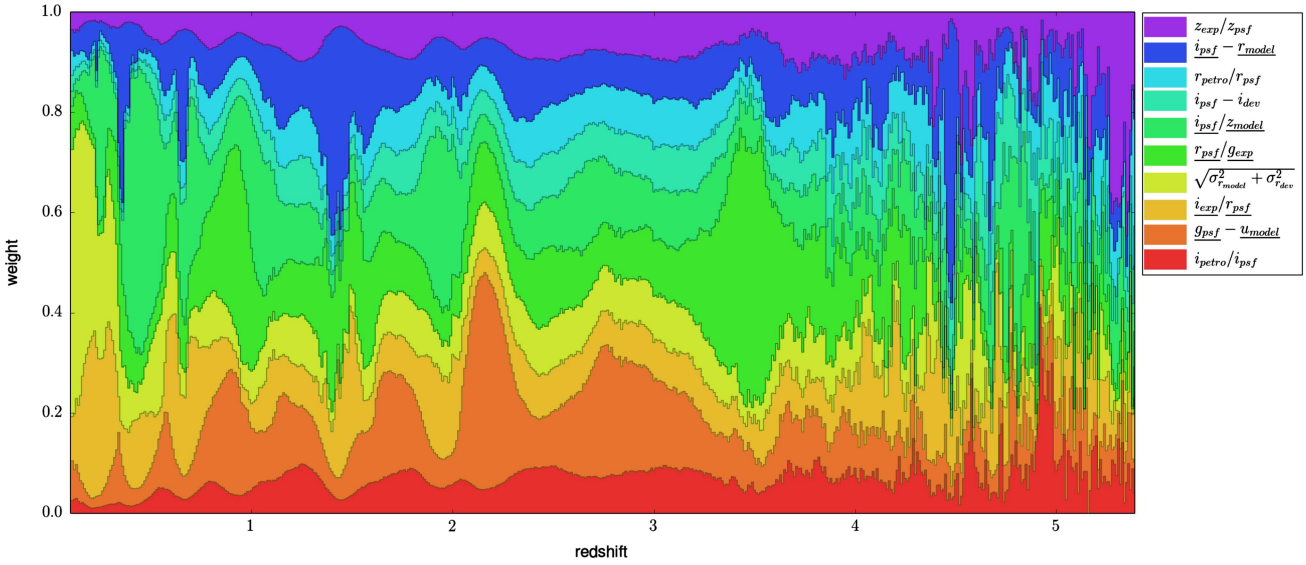


Fig. 7. Importance of every feature of the Best_{10} subset from experiment DR7+9. For a sliding redshift bin of $\Delta z = 0.2$, the importance of every feature was calculated in a localized regression model based on the Gini index as utilized by the RF. The colour code used is the same adopted for the chord diagram in Fig. 3.

set does not show a dominant role when using multiple features. When building a photometric regression model based on just a single feature, the concentration index in the *i* band provides the best tracer for distance. Therefore a concentration index in the *i* band is consequently chosen in all the three experiments. This selection is of course heavily biased by the distribution of our training objects with respect to redshift and by the fact that objects for training are selected based on the classification of the spectral template fitting of SDSS.

As soon as more photometric features are used, the spectral energy distribution and distinct spectral features are the dominant source of information for estimating the redshifts. Those features are mainly ratios. To use ratios instead of colours is a surprising fact, as in the literature colours are the usual choice for photometric redshift estimation models. In Fig. 7 one can inspect how the different features contribute at different redshift bins, building a well-performing model that covers the full redshift range. Besides some very narrow redshift regions, no clear structure with preference of some photometric features can be observed at higher redshifts ($z > 4$). This is due to the poor coverage of the training and validation data in that range. The ordering of the features in the Best_{10} set and their importance as shown in Fig. 7 can be compared with the global feature importance as obtained from the RF experiment (Table 7). The feature importance calculated on the overall redshift distribution gives different indications with respect to the bin-wise analysis, but it is quite consistent with the original order obtained from the feature selection. This is a further demonstration of the stability and robustness of the proposed method.

The different behaviours and importance found for the features in the individual redshift bins can be partially explained by analysing distinct features in the spectral energy distribution. By carefully inspecting the emission lines of quasars as reported by the SDSS spectral pipeline, a connection between some photometric features and emission lines could be found. Those features that are composed of adjacent filter bands are very sensitive to spectral lines that are in the vicinity of the overlapping area of filter transmission curves. This can be explained by a flipping of the feature, for example positive or negative for colours and above or below one for ratios. Already a little shift of an emission line with respect to the redshift is enough to create a significant change in the feature space that is detected and utilized by the machine learning model. Five features of the Best_{10} share this characteristic. Therefore the discussion with respect to emission lines is focused on selected features that are composed of magnitudes from neighbouring filter bands. Using the well known relation

$$z = \frac{\lambda_{\text{observed}}}{\lambda_{\text{emitted}}} - 1 = \frac{\lambda_{\text{filter intersection}}}{\lambda_{\text{qso emission line}}} - 1, \quad (2)$$

it is possible to calculate the redshift at which a specific emission line becomes traceable when using a certain filter combination. The proposed features capture many distinct emission lines, showing peaks in the redshift bins where the lines appear. This is shown in Figs. 8 and 9, where the feature importance has been compared with the classic features of the corresponding bands. To understand better the influence of the usage of magnitudes describing extended objects, both the PSF and the model magnitudes of the classic features where used for comparison. In Fig. 8

Table 7. Features of the Best_{10} set from experiment DR7+9, ordered by decreasing importance as expressed by the score of the RF based on the Gini criterion.

Position		Feature	Score
1	▲1	$g_{\text{psf}} - u_{\text{model}}$	0.424
2	▼1	$i_{\text{petro}}/i_{\text{psf}}$	0.121
3	▲1	$\sqrt{\sigma^2_{r_{\text{model}}} + \sigma^2_{r_{\text{dev}}}}$	0.092
4	▼1	$i_{\text{exp}}/r_{\text{psf}}$	0.072
5	=	$r_{\text{psf}}/g_{\text{exp}}$	0.071
6	=	$i_{\text{psf}}/z_{\text{model}}$	0.064
7	▲2	$i_{\text{psf}} - r_{\text{model}}$	0.062
8	▼1	$i_{\text{psf}} - i_{\text{dev}}$	0.042
9	▲1	$z_{\text{exp}}/z_{\text{psf}}$	0.026
10	▼2	$r_{\text{petro}}/r_{\text{psf}}$	0.025

Notes. The change with respect to the initially found ordering of the presented approach, and the RF score are reported, too.

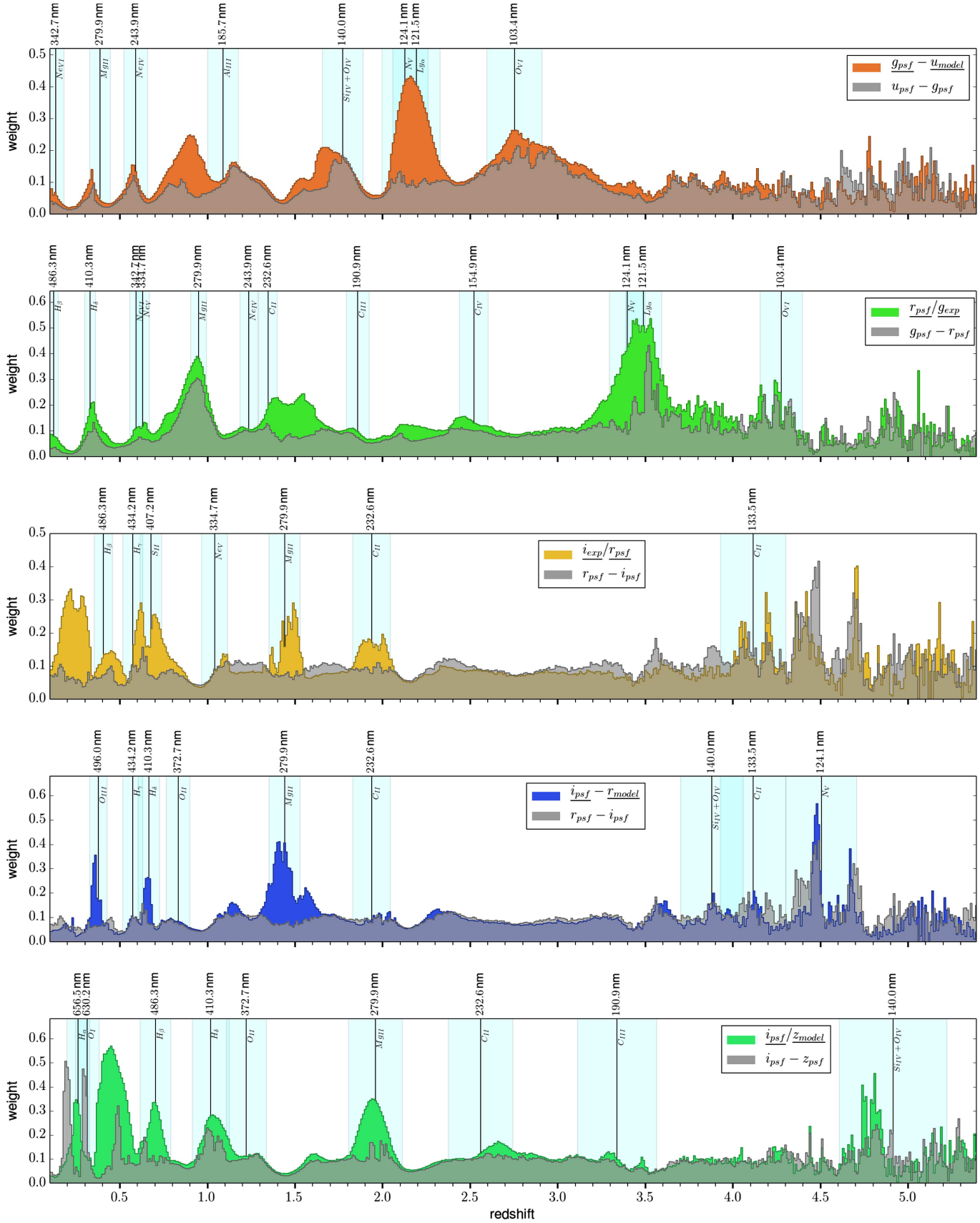


Fig. 8. Feature importance of the five features from the Best_{10} set composed by magnitudes from neighbouring bands. As in Fig. 7, for a sliding redshift bin of $\Delta z = 0.2$, the importance of every feature was calculated. The results are compared to the classic features using PSF magnitudes of the same bands. Based on the characteristics of the *ugriz* filters, the wavelengths indicating the start, centre, and end of the overlapping regions are used to overplot the positions of particular quasar emission lines using Eq. (2). The used colour code is the same as in Fig. 3, while corresponding features of the $\mathsf{Classic}_{10}$ set are always shown in grey.

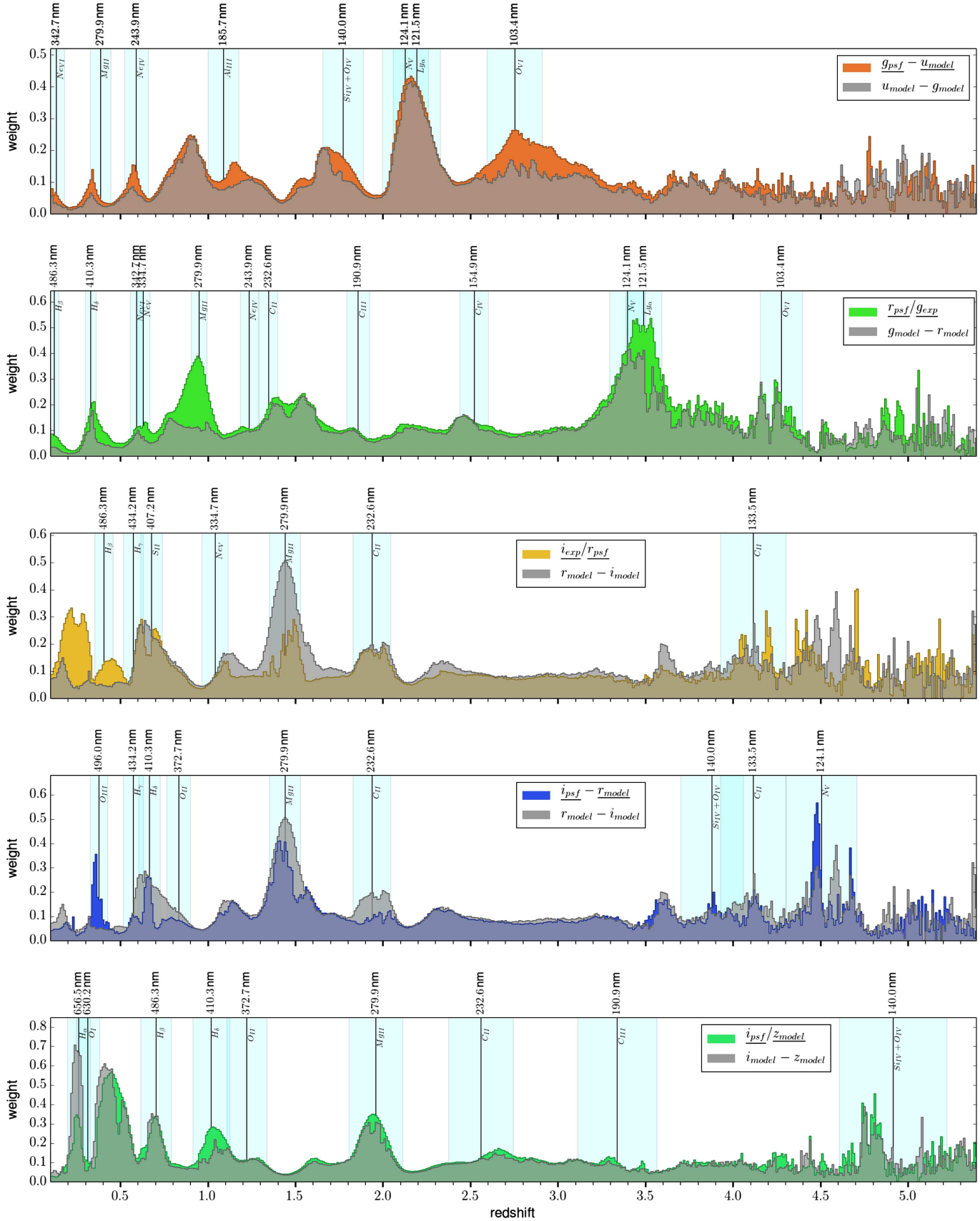


Fig. 9. Feature importance of the five features from the Best_{10} set composed by magnitudes from neighbouring bands. As in Fig. 7, for a sliding redshift bin of $\Delta z = 0.2$, the importance of every feature was calculated. The results are compared to the classic features using model magnitudes of the same bands. Based on the characteristics of the *ugriz* filters, the wavelengths indicating the start, centre, and end of the overlapping regions are used to overplot the positions of particular quasar emission lines using Eq. (2). The used colour code is the same as in Fig. 3, while corresponding features of the $\mathsf{Classic}_{10}$ set are always shown in grey.

the comparison is performed with respect to PSF colours, while in Fig. 9 the same comparison is done with respect to model colours. By using Eq. 2, a selected set of spectral emission lines of quasars has been convolved with the corresponding filter characteristics to annotate the plots. Besides the maximum of the overlapping region, the start and the end of the intersection are depicted. We defined the upper and lower limits as the points at which the sensitivity of the filter curve is equal to 0.001 in quantum efficiency. It can be seen that many emission lines perfectly correspond to peaks in importance exhibited by the features of the Best_{10} set. This can be observed only partially for the classic features.

In particular, purely PSF or model magnitude-based colours have a different and often complementary contribution for several spectral lines. This is due to the fact that either concentrated or extended characteristics of the analysed objects are considered. The proposed features are more suitable than classic features to describe the peaks at distinct emission lines. Considering the $N_V - Ly_\alpha$ lines for the $\underline{g_{psf}} - \underline{u_{model}}$ feature, the comparison between the extended and concentrated classic features clearly indicates that an extended component of the source is captured via this feature. Keeping in mind that a pixel size of 0.4″ of the SDSS camera corresponds[5] at a redshift of $z \approx 2.2$ to ≈3.4 kpc, this is a clear indicator that the hosting galaxy is significantly contributing to the solution of the photometric redshift estimation model. A similar behaviour can be observed for the $N_V - Ly_\alpha$ lines in the $\underline{r_{psf}}/\underline{g_{exp}}$ feature, while the Mg_{II} emission line mainly appears in the PSF colour. Therefore the Mg_{II} emission line can be considered to be more prominent in the central region of the objects. Between the most notable lines, the Lyman-α and the Balmer series can be identified. Other important lines found are the C_{II}, C_{III}, C_{IV}, O_I, O_{II}, O_{III}, O_{VI} , and the Mg_{II} lines. Besides the identified peaks caused by specific emission lines, some peaks in weight stay unexplained. Even though it is possible to distinguish between mostly spatially extended or concentrated characteristics of the objects, an association of a single emission line fails. In those cases not the transition of a line between two filters but an overall shape relation is captured by the selected parameters. As the selected features combine the strength of identifying line transitions as well as morphological characteristics, the resulting boost in performance of the photometric redshift estimation model can be well explained. To explain the meaning of the selected features that use a combination of features extracted from the same photometric band and thereby describe a morphological structure of the source, further image-based investigations are necessary. This proves that a model using the proposed feature selection approach is better able to exploit the information that represents the underlying physical and morphological structure as well as the processes going on in the sources.

6. Conclusions

In this work a method to select the best features for photometric redshift estimation is proposed. The features are calculated via a greedy forward selection approach, in which the features are selected from a set of 4520 combinations based on the photometric and shape information stored in the SDSS DR7 and DR9 catalogues. By randomly sampling the training data and running multiple kNN experiments, trees in which every branch constitutes a subset of features were generated for all the experiments. The obtained branches were then validated using a RF model and compared to the results obtained using classic sets of features. Moreover, the results were compared with a convolutional neural network based model, meant to automatically perform the feature extraction and selection. Three experiments, based on different catalogues, were carried out. The first catalogue was obtained selecting quasars from SDSS DR7 and applying photometric flags. The second catalogue was composed of quasars from SDSS DR7 too, but without using photometric flags. Finally, the third catalogue was made by mixing SDSS DR7 and DR9 quasars, in order to extend the redshift distribution. We have shown that all the sets obtained in all the experiments outperform the $\mathsf{Classic}_{10}$, and in particular a best-performing branch has been identified for each catalogue. The best sets also gave a better performance with respect to the automatic model (even though the latter typically shows a better calibration and is less affected by outliers when predicting PDFs instead of point estimates). The new best features obtained in the present work are not immediately comprehensible. Further analysis shows a relation between the dominant features of the Best_{10} set and the emission lines of quasars, which correspond to the peaks of importance of the different features along the redshift distribution. The same analysis carried out on the $\mathsf{Classic}_{10}$ features proves that the latter are not able to capture the same physical information as compactly as the selected features. This explains why the results obtained with the proposed method are outstanding with respect to the ones obtained with the $\mathsf{Classic}_{10}$ features. Moreover, we demonstrate that the proposed features fill the redshift space in a complementary way, each adding information that is relevant in different redshift ranges. The proposed method is highly stable, as shown from the distribution of the features and the groups to which they belong. The experiments show that the useful information is concentrated in a reduced number of features, which are typically very different from the $\mathsf{Classic}_{10}$. Furthermore, we verified that the difference in terms of performance with respect to the various sets is almost negligible. This demonstrates that the true advantage with respect to the $\mathsf{Classic}_{10}$ features is not given by the selected features themselves, but from their distribution and type in the specific set. Therefore, the stability shown from the different branches, for example the common distribution scheme of the features, and the ability to better capture the underlying physical processes, explains the superior performance obtained. The method is very general and could be applied to several tasks in astrophysics (and not only in astrophysics). In the future we propose to apply it to different sources (i.e. galaxies with and without an active nuclei) in order to verify if the obtained features are general or if they are only related to the fine structure of the data itself and to this specific population of sources. This includes the question of how much the processes of the active galactic nuclei dominate with respect to the processes in the surrounding galaxy the feature selection approach. It goes without saying that this first step made in the interpretation of the new features could open new doors in the understanding of the physics of quasars with respect to distance and age by providing better and more precise tracers. On the other hand, the method shows a different approach alternative to the application of deep learning, but also employing GPUs intensively. Both approaches are meant to establish an affordable and well-performing method to precisely predict photometric redshifts, in light of the upcoming missions and instruments in the near future.

Acknowledgements. The authors gratefully acknowledge the support of the Klaus Tschira Foundation. SC acknowledges support from the project "Quasars at high redshift: physics and Cosmology" financed by the ASI/INAF agreement 2017-14-H.0. We would like to thank Nikos Gianniotis and Erica

[5] Using Wright (2006) with $H_0 = 69.6, \Omega_M = 0.286, \Omega_{DE} = 0.714$.

Hopkins for proofreading and commenting on this work. Topcat has been used for this work (Taylor 2005). The DCMDN model has been developed using Theano (The Theano Development Team et al. 2016). To extract the image for the DCMDN we made use of HIPS (Fernique et al. 2015) and Montage (Berriman et al. 2004). Montage is funded by the National Science Foundation under Grant Number ACI-1440620, and was previously funded by the National Aeronautics and Space Administration's Earth Science Technology Office, Computation Technologies Project, under Cooperative Agreement Number NCC5-626 between NASA and the California Institute of Technology. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2012, ApJS, 203, 21
Aksoy, S., & Haralick, R. M. 2000, Pattern Recognit. Lett. 22, 563
Athiwaratkun, B., & Kang, K. 2015, ArXiv e-prints [arXiv:1507.02313]
Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, ApJ, 683, 12
Beck, R., Lin, Ishida, E., et al. 2017, Mon. Notes Astron. Soc. S. Afr., 468, 4323
Benavente, P., Protopapas, P., & Pichara, K. 2017, ApJ, 845
Berriman, G. B., Good, J. C., Laity, A. C., et al. 2004, ASP Conf. Ser., 314, 593
Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., & Steward, L. 2014, ApJS, 210, 9
Bishop, C. M. 2006, Pattern Recognition and Machine Learning (Information Science and Statistics) (Secaucus, NJ: Springer-Verlag New York, Inc.)
Bonnett, C., Troxel, M. A., Hartley, W., et al. 2016, Phys. Rev. D, 94, 042005
Breiman, L. 1996, Mach. Learn. 24, 123
Breiman, L., Friedman, J., Olshen, R., & Stone, C. 1984, Classification and Regression Trees (Monterey, CA: Wadsworth and Brooks)
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511
Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2013a, MNRAS, 437, 968
Cavuoti, S., Garofalo, M., Brescia, M., et al. 2013b, Smart Innov. Syst. Technol. 19, 29
Cavuoti, S., Brescia, M., De Stefano, V., & Longo, G. 2015, Exp. Astron. 39, 45
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, ArXiv e-prints [arXiv:1612.05560]
de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., et al. 2017, A&A, 604, A134
D'Abrusco, R., Staiano, A., Longo, G., et al. 2007, ApJ, 663, 752
D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111
D'Isanto, A., Cavuoti, S., Brescia, M., et al. 2016, MNRAS, 457, 3119
Donalek, C., Arun Kumar, A., Djorgovski, S. G., et al. 2013, ArXiv e-prints [arXiv:1310.1976]
Duda, R. O., Hart, P. E., & Stork, D. G. 2000, Pattern Classification, 2nd Edition (New York: Wiley-Interscience)
Fernique, P., Allen, M. G., Boch, T., et al. 2015, A&A, 578, A114
Fix, E., & Hodges, J. L. 1951, in US Air Force School of Aviation Medicine, Technical Report 4, 477
Gieseke, F., Polsterer, K. L., Oancea, C. E., & Igel, C. 2014, in 22th European Symposium on Artificial Neural Networks, ESANN 2014
Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. 2005, Mon. Weather Rev. 133, 1098
Guyon, I., & Elisseeff, A. 2003, J. Mach. Learn. Res., 3, 1157
Harnois-Déraps, J., Tröster, T., Chisari, N., et al. 2017, MNRAS, 471, 1619
Heinermann, J., Kramer, O., Polsterer, K., & Gieseke, F. 2013, Lect. Notes Comput. Sci. Ser., 8077, 86
Hersbach, H. 2000, Weather Forecasting, 15, 559
Hey, T., Tansley, S., & Tolle, K., eds. 2009, The Fourth Paradigm: Data-Intensive Scientific Discovery (Redmond, WA: Microsoft Research)
Hildebrandt, H., Wolf, C., & Benítez, N. 2008, A&A, 480, 703
Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31
Hildebrandt, H., Viola, M., Heymans, C., et al. 2016, MNRAS, 465, 1
Hoyle, B. 2016, Astron. Comput. 16, 34
Hoyle, B., Rau, M. M., Zitlau, R., Seitz, S., & Weller, J. 2015, MNRAS, 449, 1275
Ivezić, v., Tyson, J. A., Acosta, E., et al. 2008, ArXiv e-prints [arXiv:0805.2366v4]
Joudaki, S., Mead, A., Blake, C., et al. 2017, MNRAS, 471, 1259
Kohavi, R. 1995, in Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI'95 (San Francisco, USA: Morgan Kaufmann Publishers Inc.), 2, 1137
Köhlinger, F., Viola, M., Joachimi, B., et al. 2017, MNRAS, 471, 4412
Krzywinski, M. I., Schein, J. E., Birol, I., et al. 2009, Genome Res., 19, 1639
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Laurino, O., D'Abrusco, R., Longo, G., & Riccio, G. 2011, MNRAS, 418, 2165
Lupton, R. H., Gunn, J. E., & Szalay, A. S. 1999, AJ, 118, 1406
Mahabal, A., Djorgovski, S. G., Turmon, M., et al. 2008, Astron. Nachr., 329, 288
Mao, K. 2004, IEEE Trans. Syst. Man Cybern. Part B Cybern. 34, 629
Norris, R. P., Hopkins, A. M., Afonso, J., et al. 2011, PASA, 28, 215
Polsterer, K. L., Gieseke, F., Igel, C., & Goto, T. 2014, ASP Conf. Ser., 485, 425
Polsterer, K., Gieseke, F., & Igel, C. 2015, ASP Conf. Ser., 495, 81
Richards, G. T., Weinstein, M. A., Schneider, D. P., et al. 2001, AJ, 122, 1151
Richards, G. T., Myers, A. D., Gray, A. G., et al. 2009, ApJS, 180, 67
Rimoldini, L., Dubath, P., Süveges, M., et al. 2012, MNRAS, 427, 2917
Smirnov, E., & Markov, A. 2017, MNRAS, 469, 2024
Tangaro, S., Amoroso, N., Brescia, M., et al. 2015, Comput. Math. Methods Med. 2015
Taylor, A. R. 2008, IAU Symp., 248, 164
Taylor, M. B. 2005, ASP Conf. Ser., 347, 29
The Theano Development Team, Al-Rfou, R., Alain, G., et al. 2016, ArXiv e-prints [arXiv:1605.02688]
Tortora, C., La Barbera, F., Napolitano, N., et al. 2016, MNRAS, 457, 2845
Vaccari, M., Covone, G., Radovich, M., et al. 2016, in Proceedings of the 4th Annual Conference on High Energy Astrophysics in Southern Africa (HEASA 2016), online at http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=275, id.26, 26
van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2
Wright, E. L. 2006, PASP, 118, 1711
Zhang, Y., Ma, H., Peng, N., Zhao, Y., & Wu, X.-b. 2013, AJ, 146, 22

Appendix A: Additional tables and figures

In this section, the additional tables for the features selection and the tree structure, together with the related chord diagrams for the experiments DR7a and DR7b are given. A brief explanation of how to read a chord diagram follows.

A.1. Chord diagram: how to read

The chord diagram is a tool to visualize complex structures and relations in multidimensional data, which is arranged in a matrix shape. The data are disposed in a circle and each element, in our case the features, is associated with a different colour. The relations between the elements are expressed by ribbons which connect them, with a specific width related to the importance of that specific connection. Therefore, the different ribbons can enter or exit from every arc, representing the features. The chord diagrams utilized for this work are characterized by three external arcs for each feature. Ordered from outside to inside, the external arcs represent the occurrences of a particular feature: the total percentage of the individual connections, the numbers and sources of connections entering, and the numbers and targets of connections exiting. Therefore, starting from the first features indicated in the captions, it is possible to follow all the possible paths of the tree, depicting the different feature subsets and their global scheme. Splitting points, joints, and complex interplay between feature groups can thereby be analyzed intuitively.

Table A.1. Detailed feature branches obtained from the feature selection for the experiment DR7a.

id	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1								$\sigma_{g_{model}}$	z_{model}/z_{psf}	$i_{psf} - i_{dev}$
2					g_{psf}/i_{exp}	i_{psf}/z_{model}		z_{model}/z_{psf}	$\sigma_{g_{model}}$	
3										g_{petro}/r_{petro}
4				i_{psf}/i_{model}				$\sigma_{g_{model}}$	z_{model}/z_{psf}	i_{dev}/i_{psf}
5					i_{psf}/z_{model}	g_{petro}/r_{petro}	$i_{psf} - i_{petro}$			
6									$\sigma_{g_{model}}$	$i_{psf} - i_{dev}$
7								z_{model}/z_{psf}	$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{dev}}}$	
8	i_{psf}/i_{exp}		r_{psf}/i_{exp}							i_{dev}/i_{psf}
9								i_{dev}/i_{psf}	$\sigma_{g_{model}}$	
10					g_{psf}/i_{exp}	i_{psf}/z_{model}	i_{dev}/i_{psf}	$\sigma_{g_{model}}$	$r_{psf} - r_{petro}$	g_{petro}/r_{petro}
11		g_{psf}/u_{model}						$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{dev}}}$		
12				z_{model}/z_{psf}				$\sigma_{g_{model}}$	z_{model}/z_{psf}	$i_{psf} - i_{dev}$
13									$\sigma_{g_{model}}$	$r_{psf} - r_{petro}$
14							$i_{psf} - i_{petro}$	i_{dev}/i_{psf}		r_{dev}/r_{psf}
15					i_{psf}/z_{model}	g_{petro}/r_{petro}			$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{dev}}}$	$r_{psf} - r_{petro}$
16										r_{dev}/r_{psf}
17								$\sigma_{g_{model}}$		
18								$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{dev}}}$	z_{model}/z_{psf}	$i_{psf} - i_{petro}$
19	i_{psf}/i_{model}		r_{psf}/i_{model}	i_{dev}/i_{psf}	r_{psf}/g_{model}		$r_{psf} - r_{petro}$			
20*						i_{psf}/z_{model}		$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{exp}}}$		
21					g_{psf}/i_{exp}			$\sigma_{g_{model}}$	g_{petro}/r_{petro}	z_{model}/z_{psf}
22										$i_{psf} - i_{petro}$

Notes. The 20th branch, indicated with the $*$ symbol, is the best performing subset with respect to the experiments using the RF. The *ratios* and *photometric ratios* are indicated, respectively, with vertical lines and dots. The *differences* are marked with horizontal lines and the *errors* with north west lines. The color code for the features is the same as shown in the chord diagram in Fig. A.1.

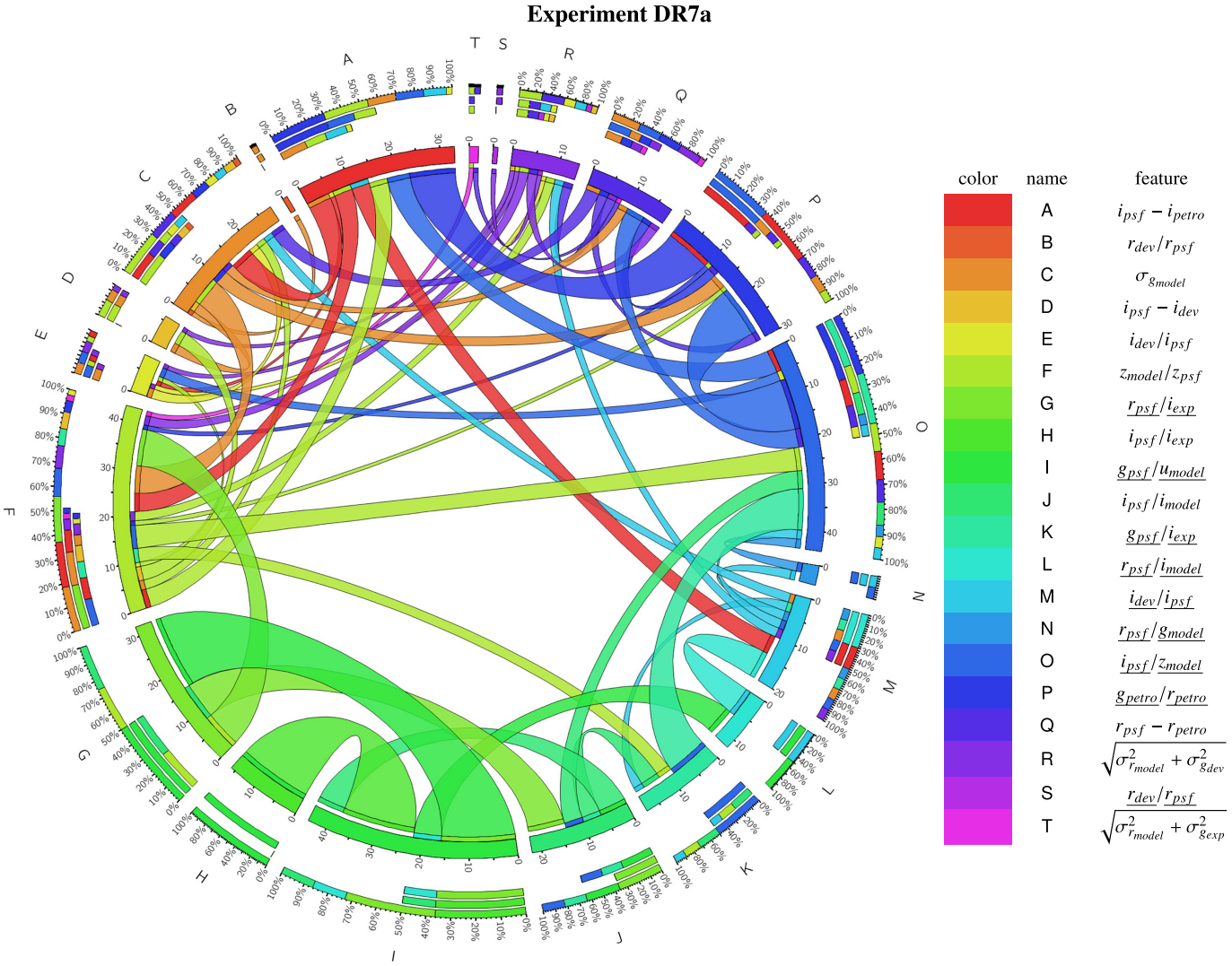
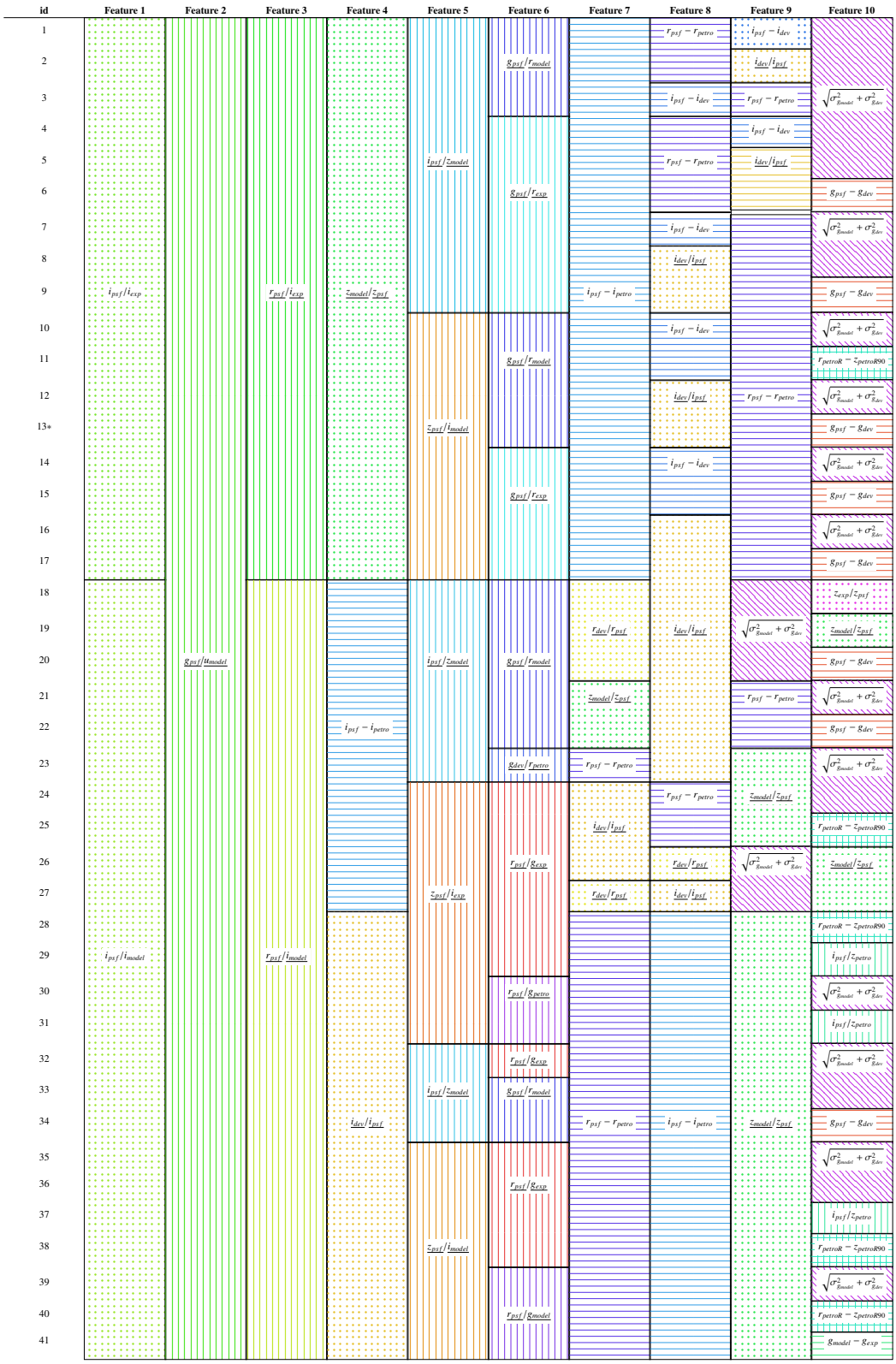


Fig. A.1. Chord diagram for the experiment DR7a. Every feature is associated to a specific colour, and starting from the first features (H, J) it is possible to follow all the possible paths of the tree, depicting the different feature subsets.

Table A.2. Detailed feature branches obtained from the feature selection for the experiment DR7b.



Notes. The 13th branch, indicated with the $*$ symbol, is the best performing subset with respect to the experiments using the RF. The *ratios* and *photometric ratios* are indicated, respectively, with vertical lines and dots. The *differences* are marked with horizontal lines and the *errors* are with north west lines. Finally, the only feature composed by radius is indicated with a grid. The color code for the features is the as same shown in the chord diagram in Fig. A.2.

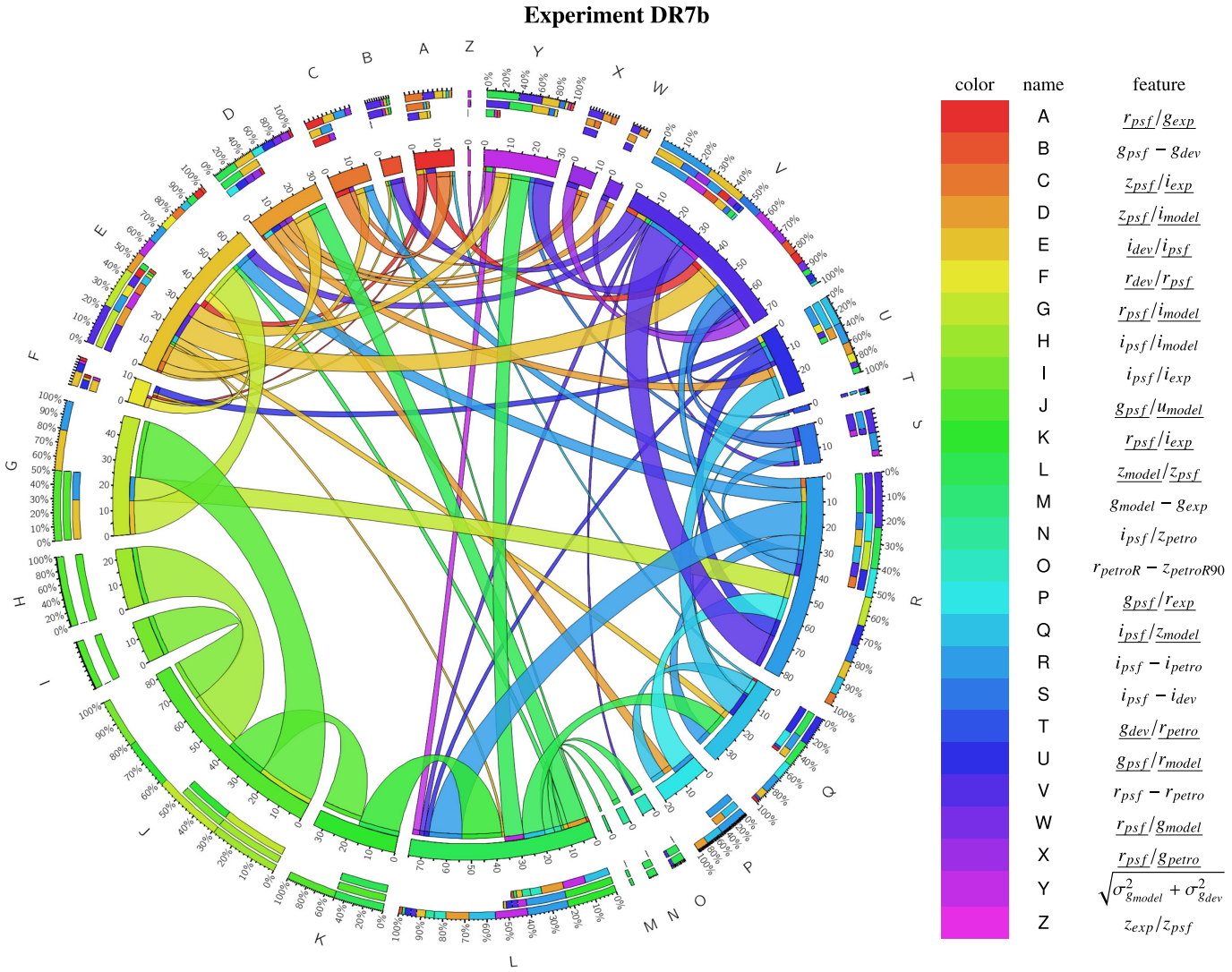


Fig. A.2. Chord diagram for the experiment DR7b. Every feature is associated to a specific colour, and starting from the first features (H, I) it is possible to follow all the possible paths of the tree, depicting the different feature subsets.

Appendix B: Data

The SDSS object IDs and coordinates of the extracted quasars for the three catalogues are available as supplementary information, as ASCII files.
dr7a.csv contains the SDSS object IDs and coordinates of the quasars for experiment DR7a.
dr7b.csv contains the SDSS object IDs and coordinates of the quasars for experiment DR7b.
dr7+9.csv contains the SDSS object IDs and coordinates of the quasars for experiment DR7+9.

Appendix C: Code

The code of the DCMDN model is available on the ASCL[6].

[6] http://www.ascl.net/ascl:1709.006

Appendix D: SDSS QSO query

In the following, the statements used to query the SDSS database are provided.

D.1. Experiment DR7

```
SELECT
    s.specObjID, p.objid, p.ra, p.dec, s.targetObjID, s.z, s.zErr,
    p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i, p.psfMag_z,
    p.psfMagErr_u, p.psfMagErr_g, p.psfMagErr_r, p.psfMagErr_i, p.psfMagErr_z,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
    p.modelMagErr_u, p.modelMagErr_g, p.modelMagErr_r, p.modelMagErr_i, p.modelMagErr_z,
    p.devMag_u, p.devMag_g, p.devMag_r, p.devMag_i, p.devMag_z,
    p.devMagErr_u, p.devMagErr_g, p.devMagErr_r, p.devMagErr_i, p.devMagErr_z,
    p.expMag_u, p.expMag_g, p.expMag_r, p.expMag_i, p.expMag_z,
    p.expMagErr_u, p.expMagErr_g, p.expMagErr_r, p.expMagErr_i, p.expMagErr_z,
    p.petroMag_u, p.petroMag_g, p.petroMag_r, p.petroMag_i, p.petroMag_z,
    p.petroMagErr_u, p.petroMagErr_g, p.petroMagErr_r, p.petroMagErr_i, p.petroMagErr_z,
    p.extinction_u, p.extinction_g, p.extinction_r, p.extinction_i, p.extinction_z,
    p.devRad_u, p.devRad_g, p.devRad_r, p.devRad_i, p.devRad_z,
    p.expRad_u, p.expRad_g, p.expRad_r, p.expRad_i, p.expRad_z,
    p.petroRad_u, p.petroRad_g, p.petroRad_r, p.petroRad_i, p.petroRad_z,
    p.petroR90_u, p.petroR90_g, p.petroR90_r, p.petroR90_i, p.petroR90_z,
    p.petroR50_u, p.petroR50_g, p.petroR50_r, p.petroR50_i, p.petroR50_z,
    p.devAB_u, p.devAB_g, p.devAB_r, p.devAB_i, p.devAB_z,
    p.expAB_u, p.expAB_g, p.expAB_r, p.expAB_i, p.expAB_z i

FROM
   SpecPhoto as s, PhotoObjAll as p

WHERE
    p.mode = 1 AND p.SpecObjID = s.SpecObjID AND
    dbo.fPhotoFlags('PEAKCENTER') != 0 AND
    dbo.fPhotoFlags('NOTCHECKED') != 0 AND
    dbo.fPhotoFlags('DEBLEND_NOPEAK') != 0 AND
    dbo.fPhotoFlags('PSF_FLUX_INTERP') != 0 AND
    dbo.fPhotoFlags('BAD_COUNTS_ERROR') != 0 AND
    dbo.fPhotoFlags('INTERP_CENTER') != 0 AND
    p.objid=s.objid and (specClass = 3 OR specClass = 4) AND
    s.psfMag_i > 14.5 AND (s.psfMag_i - s.extinction_i) < 21.3 AND
    s.psfMagErr_i < 0.2
```

D.2. Experiment DR7b

```
SELECT
    s.specObjID, p.objid, p.ra, p.dec, s.targetObjID, s.z, s.zErr,
    p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i, p.psfMag_z,
    p.psfMagErr_u, p.psfMagErr_g, p.psfMagErr_r, p.psfMagErr_i, p.psfMagErr_z,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
    p.modelMagErr_u, p.modelMagErr_g, p.modelMagErr_r, p.modelMagErr_i, p.modelMagErr_z,
    p.devMag_u, p.devMag_g, p.devMag_r, p.devMag_i, p.devMag_z,
    p.devMagErr_u, p.devMagErr_g, p.devMagErr_r, p.devMagErr_i, p.devMagErr_z,
    p.expMag_u, p.expMag_g, p.expMag_r, p.expMag_i, p.expMag_z,
    p.expMagErr_u, p.expMagErr_g, p.expMagErr_r, p.expMagErr_i, p.expMagErr_z,
    p.petroMag_u, p.petroMag_g, p.petroMag_r, p.petroMag_i, p.petroMag_z,
    p.petroMagErr_u, p.petroMagErr_g, p.petroMagErr_r, p.petroMagErr_i, p.petroMagErr_z,
    p.extinction_u, p.extinction_g, p.extinction_r, p.extinction_i, p.extinction_z,
    p.devRad_u, p.devRad_g, p.devRad_r, p.devRad_i, p.devRad_z,
    p.expRad_u, p.expRad_g, p.expRad_r, p.expRad_i, p.expRad_z,
    p.petroRad_u, p.petroRad_g, p.petroRad_r, p.petroRad_i, p.petroRad_z,
    p.petroR90_u, p.petroR90_g, p.petroR90_r, p.petroR90_i, p.petroR90_z,
    p.petroR50_u, p.petroR50_g, p.petroR50_r, p.petroR50_i, p.petroR50_z,
    p.devAB_u, p.devAB_g, p.devAB_r, p.devAB_i, p.devAB_z,
    p.expAB_u, p.expAB_g, p.expAB_r, p.expAB_i, p.expAB_z
```

```
    into mydb.qso_dr7_noflags from SpecPhoto as s, PhotoObjAll as p

WHERE
    p.SpecObjID = s.SpecObjID AND
    p.objid=s.objid and (specClass = 3 OR specClass = 4)
```

D.3. Experiment DR7+9

```
SELECT
   m.objid, m.ra AS ra1, m.dec AS dec1,
   n.objid, n.distance,
   p.ra AS ra2, p.dec AS dec2,
   p.objid, p.ra, p.dec, p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i,
   p.psfMag_z,p.psfMagErr_u, p.psfMagErr_g, p.psfMagErr_r, p.psfMagErr_i,
   p.psfMagErr_z,p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
   p.modelMagErr_u, p.modelMagErr_g, p.modelMagErr_r, p.modelMagErr_i,
   p.modelMagErr_z,p.devMag_u, p.devMag_g, p.devMag_r, p.devMag_i, p.devMag_z,
   p.devMagErr_u, p.devMagErr_g, p.devMagErr_r, p.devMagErr_i, p.devMagErr_z,
   p.expMag_u, p.expMag_g, p.expMag_r, p.expMag_i, p.expMag_z,p.expMagErr_u, p.expMagErr_g,
   p.expMagErr_r, p.expMagErr_i, p.expMagErr_z,p.petroMag_u, p.petroMag_g, p.petroMag_r,
   p.petroMag_i, p.petroMag_z,p.petroMagErr_u, p.petroMagErr_g, p.petroMagErr_r,
   p.petroMagErr_i, p.petroMagErr_z,p.extinction_u, p.extinction_g, p.extinction_r,
   p.extinction_i, p.extinction_z,p.devRad_u, p.devRad_g, p.devRad_r, p.devRad_i,
   p.devRad_z,p.expRad_u, p.expRad_g, p.expRad_r, p.expRad_i, p.expRad_z,p.petroRad_u,
   p.petroRad_g, p.petroRad_r, p.petroRad_i, p.petroRad_z,p.petroR90_u, p.petroR90_g,
   p.petroR90_r, p.petroR90_i, p.petroR90_z,p.petroR50_u, p.petroR50_g, p.petroR50_r,
   p.petroR50_i, p.petroR50_z,p.devAB_u, p.devAB_g, p.devAB_r, p.devAB_i, p.devAB_z,p.expAB_u,
   p.expAB_g, p.expAB_r, p.expAB_i, p.expAB_z
   into mydb.quasar_dr7_dr9_allphoto from MyDB.dr7_dr9_quasar AS m

CROSS APPLY dbo.fGetNearestObjEq( m.ra, m.dec, 0.5) AS n
JOIN PhotoObj AS p ON n.objid=p.objid
```