Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **450,** 305–316 (2015) doi:10.1093/mnras/stv599

Data augmentation for machine learning redshifts applied to Sloan Digital Sky Survey galaxies

Ben Hoyle,[1,2★] Markus Michael Rau,[1,3] Christopher Bonnett,[4] Stella Seitz[1,3] and Jochen Weller[1,2,3]

[1] *Universitaets-Sternwarte, Fakultaet fuer Physik, Ludwig-Maximilians Universitaet Muenchen, Scheinerstr. 1, D-81679 Muenchen, Germany*
[2] *Excellence Cluster Universe, Boltzmannstr. 2, D-85748 Garching, Germany*
[3] *Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, D-85748 Garching, Germany*
[4] *Institut de Fsica d'Altes Energies, Universitat Autonoma de Barcelona, E-08193 Bellaterra, Spain*

Accepted 2015 March 17. Received 2015 March 10; in original form 2015 January 27

ABSTRACT

We present analyses of data augmentation for machine learning redshift estimation. Data augmentation makes a training sample more closely resemble a test sample, if the two base samples differ, in order to improve measured statistics of the test sample. We perform two sets of analyses by selecting 800 000 (1.7 million) Sloan Digital Sky Survey Data Release 8 (Data Release 10) galaxies with spectroscopic redshifts. We construct a base training set by imposing an artificial *r*-band apparent magnitude cut to select only bright galaxies and then augment this base training set by using simulations and by applying the K-CORRECT package to artificially place training set galaxies at a higher redshift. We obtain redshift estimates for the remaining faint galaxy sample, which are not used during training. We find that data augmentation reduces the error on the recovered redshifts by 40 per cent in both sets of analyses, when compared to the difference in error between the ideal case and the non-augmented case. The outlier fraction is also reduced by at least 10 per cent and up to 80 per cent using data augmentation. We finally quantify how the recovered redshifts degrade as one probes to deeper magnitudes past the artificial magnitude limit of the bright training sample. We find that at all apparent magnitudes explored, the use of data augmentation with tree-based methods provide an estimate of the galaxy redshift with a low value of bias, although the error on the recovered redshifts increases as we probe to deeper magnitudes. These results have applications for surveys which have a spectroscopic training set which forms a biased sample of all photometric galaxies, for example if the spectroscopic detection magnitude limit is shallower than the photometric limit.

Key words: catalogues – surveys – galaxies: distances and redshifts.

1 INTRODUCTION

Photometric surveys can be maximally exploited for large-scale structure analyses once galaxies have been identified and their positions on the sky and in redshift space have been measured. Measuring accurate spectroscopic redshifts is costly and time intensive, and is typically only performed for a small subsample of all galaxies. In particular the spectroscopic sample is often a biased sample of the full photometric galaxy catalogue due to the limiting magnitude that a spectroscopic redshift for a galaxy can be measured, and the deeper limiting magnitude that a galaxy may be identified photometrically. This paper examines how the spectroscopic training set can be augmented (or complimented) to span an input feature space that more closely resembles that of the full photometric galaxy sample, to improve redshift estimates using machine learning.

Photometric redshifts can also be estimated by parametric techniques, for example from galaxy spectral energy distribution (SED) templates. Some templates encode our knowledge of stellar population models which result in predictions for the evolution of galaxy magnitudes and colours. The parametric encoding of the complex stellar physics coupled with the uncertainty of the parameters of the stellar population models, combine to produce redshift estimates which are little better than many non-parametric techniques (see e.g. Hildebrandt et al. 2010; Dahlen et al. 2013, for an overview of different techniques). Unlike non-parametric and machine learning techniques, the aforementioned template methods do not rely on training samples of galaxies, which must be assumed to be representative of the final sample of galaxies for which redshift estimates are required. Other template methods are generated completely from, or

★ E-mail: benhoyle1212@gmail.com

in combination with, empirical data, however, these templates both require tuning, and also rely upon representative training samples.

When an unbiased training sample is available, machine learning methods offer an alternative to template methods to estimate galaxy redshifts. The 'machine architecture' determines how to best manipulate the photometric galaxy input properties (or 'features') to produce a machine learning redshift. The machine attempts to learn the most effective manipulations to minimize the difference between the spectroscopic redshift and the machine learning redshift of the training sample.

The field of machine learning for photometric redshift analysis has been developing since Tagliaferri et al. (2003) used artificial neural networks (ANNs). A plethora of machine learning architectures, including tree-based methods, have been applied to the problem of point prediction redshift estimation (see e.g. Sánchez 2014, for a further list and routine comparisons), or to estimate the full redshift probability distribution function (Gerdes et al. 2010; Bonnett 2013; Carrasco Kind & Brunner 2013; Rau et al., 2015). Machine learning architectures have also had success in other fields of astronomy such as galaxy morphology identification, and star and quasar separation (see e.g. Lahav 1997; Yeche et al. 2009).

One may combine machine learning techniques with template-based methods, and with knowledge of semi-analytic models, by augmenting the training sample with information drawn from templates or simulations. Previous work in this area was initiated by Vanzella et al. (2004) who examined the use of synthetic SED templates (e.g. Fioc & Rocca-Volmerange 1997) to augment the very small galaxy training samples available to them at the time in order to measure galaxy redshifts. Using data augmentation, the authors reduced the redshift error from 0.18 to 0.11 for 227 galaxies selected from *Hubble Deep Field*-North/South (HDF-N/S; Cohen et al. 2000; Cristiani et al. 2000) using ANNs as the machine learning architecture. They did not extend their analysis to the Sloan Digital Sky Survey (SDSS) galaxies available at the time. More recently Wolf (2009) use a hybrid empirical and template-based χ^2 approach using ANNs to improve the estimate redshifts for SDSS selected quasars, but do not extrapolate analysis outside of the training set.

In this paper we extend this early analysis by using simulated galaxies drawn from the latest semi-analytic models with recent stellar population models, and by using standard template routines to augment a non-representative training sample of galaxies, to make it more closely resemble the 'test' sample of galaxies for which redshift estimates are required. We also show how much one may rely upon this augmentation as the original training samples become more unrepresentative of the test sample.

If the training sample covers the same input feature space as the test sample, but is biased with respect to number density, one may weigh the training set galaxies to more closely resemble the test set galaxies. This method has been applied to the full probability distribution function of the redshift distribution using a k-nearest neighbour (k-NN) weighting scheme (Lima et al. 2008), and to individual galaxies (Cunha et al. 2009). This has also been applied by Sánchez (2014) using the 'covariate shift' method. We do not use the covariate shift method, or other weighting schemes in this work, because the training and test sets are defined to be unrepresentative of each other, and instead rely on data augmentation to make the samples more similar.

This paper is organized as follows: in Section 2 we describe the data sample and the data augmentation process; we present the machine learning methodology, and the analysis and results using the augmented data in Section 3; discuss in Section 4 and conclude in Section 5.

2 DATA AND AUGMENTED DATA

In this study we use a mixture of observational data drawn from two SDSS data releases, combined with simulations and data which are augmented from the observational training data.

2.1 Observational data set

The observational data in this study are drawn from SDSS Data Release 8 (Aihara et al. 2011) and the SDSS Data Release 10 (Ahn et al. 2014). The SDSS I–III uses a 4-m telescope at Apache Point Observatory in New Mexico and has CCD wide field photometry in five bands (u, g, r, i, z; Smith et al. 2002; Gunn et al. 2006), and an expansive spectroscopic follow-up program (Eisenstein et al. 2011) covering π steradian of the northern sky. The SDSS collaboration has obtained two million galaxy spectra using dual fibre-fed spectrographs. An automated photometric pipeline performed object classification to a magnitude of $r \approx 22$ and measures photometric properties of more than 100 million galaxies. The complete data sample, and many derived catalogues such as the photometric properties, are publicly available through the CasJobs server.[1]

The SDSS is well suited to the analyses presented in this paper due to the enormous number of photometrically selected galaxies with spectroscopic redshifts to use as training, cross-validation and test samples. We select galaxies from CasJobs with both spectroscopic redshifts and photometric properties using the query shown in Section A1 of Appendix A.

The MYSQL query extracts model magnitudes from which we construct all possible colours combinations. We only examine model magnitudes and colours in this work so that we can trivially combine the observed data with simulations and augmented data. Recent work has shown an improvement to the machine learning redshift measurement by using more photometric properties as input features (Hoyle et al. 2015).

We further select galaxies that have the internal SDSS photometric galaxy classification $type = 3$, have spectroscopic redshifts above 0.05, spectroscopic redshift errors less than 0.1, r-band magnitudes above 18 and apparent magnitude errors below 0.4. This reduces the sample size to 802 590 galaxies for DR8 and 1710 822 galaxies in DR10. The main differences between these data samples are that the DR10 sample probes to higher redshifts and deeper r-band magnitudes, see Fig. 1.

2.2 Training, cross-validation and test samples

In this paper we explore the effect of assigning redshift estimates for galaxies that fall outside of the feature space spanned by the original training set and cross-validation sets. This is to mimic the situation that the results of a system trained on a spectroscopic data set are applied to a (potentially deeper) photometric data set. Therefore, we intentionally create a test set of galaxies that are fainter in the r band than those galaxies used for training and cross-validation.

This is performed by applying an r-band apparent magnitude cut of 18.5 (20.5) to distinguish bright training and cross-validation sample galaxies from faint test sample galaxies for SDSS DR8 (DR10), which results in a faint galaxy sample of size 124 844

[1] skyserver.sdss3.org/CasJobs

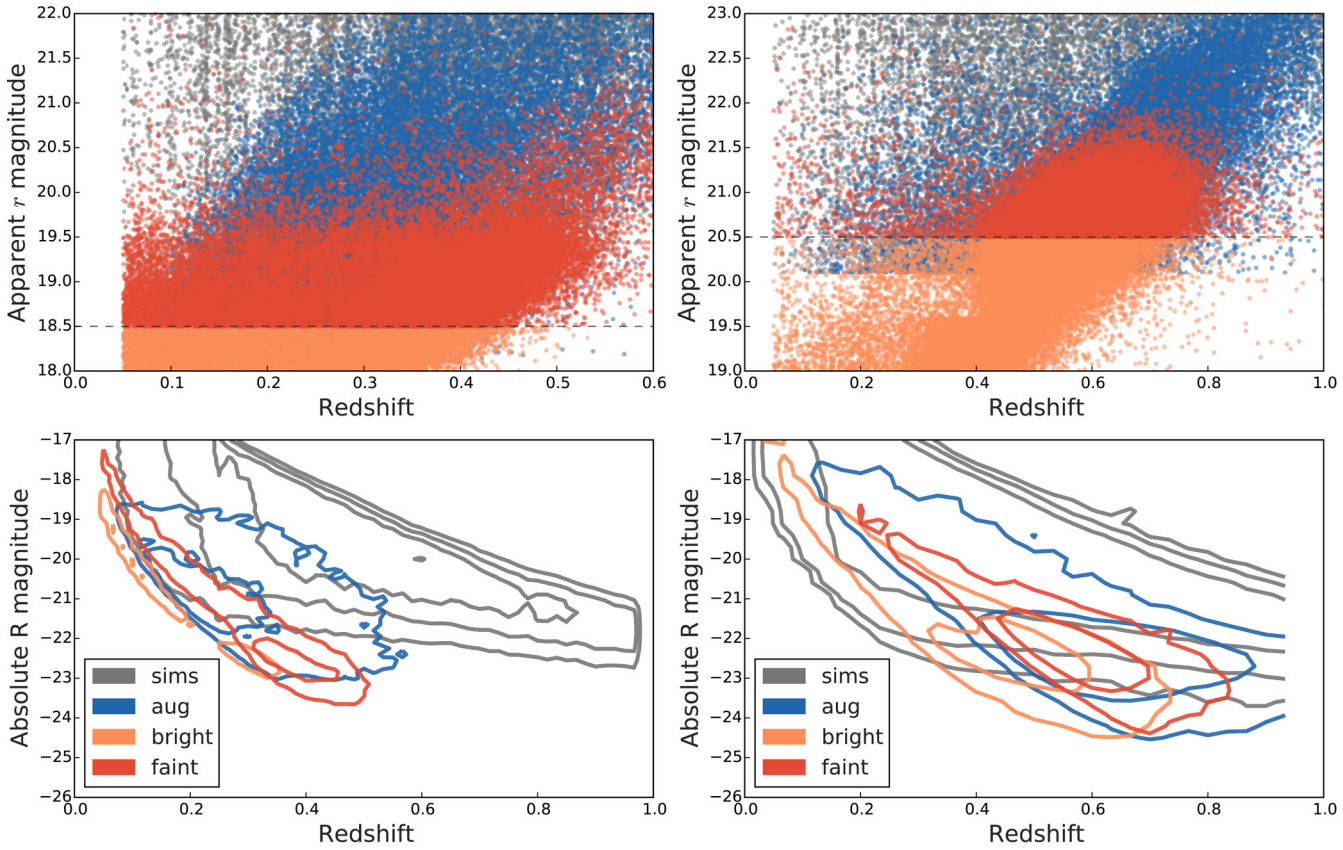

Figure 1. The distribution of magnitude and redshift for a random selection of each of the SDSS DR8 (left-hand panels) and DR10 (right-hand panels) data samples. In the upper panels, the dashed line marks the separation between the faint test sample and the bright training sample. We modify the bright samples (in each distinct set of analyses) to generate the augmented data sample using SDSS K-CORRECT. We further augmented the training set using the simulated galaxies produced using the latest semi-analytic models. The top panels show a random selection of data, and the bottom panels show the absolute magnitudes density contours for each sample. In the legend 'sims' denotes the augmented data using simulations, and 'aug' corresponds to the K-CORRECT data augmentation of the bright galaxy sample.

(304 455). We then perform data augmentation to modify and enhance the training and cross-validation samples so that they more closely resemble the test sample. This process is described in more detail below.

We also ask the following question: how well would we have done, if we were to have had training and cross-validation data that are drawn from the same sample of galaxies as the test data? We refer to this benchmark analyses as the 'ideal' case. To answer this question we build a separate training, cross-validation and test sample from the faint data sample, each of size one-third of the full faint samples in both sets of analyses. We expect that the recovered redshift errors using the augmented data sets will be somewhere between the values using this ideal case, and the value obtained by having no training data that overlaps in feature space with the test data.

We follow standard machine learning methodology, such that the training sample is used to train the machine learning system for a given machine learning hyperparameter set. The cross-validation sample is then used to select the best values for the hyperparameters of the learned system. Once the best set of hyperparameters has been decided upon, the test sample is used to measure the true ability of the learned machine to generalize to a new data set. Unless otherwise specified, the test sample in this analysis is always the ideal case test sample of faint galaxies, and is of size 33 per cent of the full faint sample in both sets of analyses. This choice of test sample size allows a fair comparison between the different combinations of data sets used for training and the benchmark ideal case.

2.3 Data augmentation using simulations

The bright data in both the DR8 and DR10 analyses are augmented using galaxies extracted from light cones (Henriques et al. 2012) created from the Millennium Simulation (Springel et al. 2005). In detail we extract redshifts and SDSS estimated magnitudes of galaxies from the table *Henriques2014a* which uses the latest semi-analytic models (see Henriques et al. 2014, for more details) and separately incorporate the stellar populations models of Bruzual & Charlot (2003) and Maraston (2005). We extract semi-analytic galaxies from both sets of stellar populations models, using the query presented in Section A2 of Appendix A.

The output of the query produces a data set of size 4859 249. We select subsamples of simulated galaxies by applying a lower limit r-band apparent magnitude selection of 18.1 (20.1) to mimic the faint galaxy sample of DR8 (DR10) for which we are interested in recovering redshift estimates. We note that this limit is slightly brighter than that of the faint galaxy test sample. This choice allows the machine learning framework to draw on a slightly brighter sample while training, which may aid the redshift assignment. We examine the effect of more carefully selecting training data in Section 3.2.1. The final sample sizes are 346 116 (1205 192) to

be used in the DR8 (DR10) analysis of which we form a training set of size 66 per cent and a cross-validation set of size 33 per cent.

The semi-analytic models have been tuned to match the abundance fractions of red and blue galaxies as a function of mass at redshift $z = 0$. However the detailed colour distribution of galaxies and their corresponding colour evolution is described by the complex stellar population evolution physics combined with the predicted star formation and metallicity histories. We note that over the redshift range of interest here, the models have not been fine tuned to replicate the magnitudes and colours of observed galaxies and therefore provide an independent estimate of the galaxy distribution.

We do not require a test sample for the augmented data set because we are not interested in obtaining a redshift estimate for the augmented data at test time. We are using the augmented data sets to help in the training and cross-validation stages. We show that the addition of the augmented data in these stages will result in an improvement in the redshift estimate of the test set of faint galaxies. For brevity we refer to this simulated augmented data sample as 'simulations' in what follows.

2.4 Data augmentation using K-CORRECT

We use the SDSS K-CORRECT package (Blanton & Roweis 2007) to augment the bright training sample. K-CORRECT is able to estimate the apparent magnitudes that a galaxy of a given magnitude and redshift would have, if it were at a different redshift.

K-CORRECT performs a χ^2 analysis by comparing the input galaxy magnitudes and redshift with different synthetic galaxy spectra to identify the best template to use as a base for which to approximate the evolution of the galaxy. The synthetic galaxy spectra are drawn from the Kewley et al. (2001) and Bruzual & Charlot (2003) stellar population models. In detail the templates explore a range of parameter space corresponding to different star formation histories with varying amounts of stellar metallicities and ages, and models for galactic dust extinction. The χ^2 minimization is performed on weighted combinations of (up to five) SEDs, and the best-fitting combinations are identified from sets of observation data, including earlier SDSS data releases. We refer the reader to Blanton & Roweis (2007) and howdy.physics.nyu.edu/index.php/Kcorrect for more details.

We perform this data augmentation by first randomly selecting (with replacement) galaxies from the bright training sample. The values of the apparent magnitudes of the selected galaxies are Gaussian resampled with a scatter of 1 per cent of the measured SDSS magnitude errors. The K-CORRECT package is deterministic and therefore this additional magnitude scatter allows the generation of similar, but distinct, augmented training examples. We then assign a new redshift to the galaxy by sampling from a Gaussian of width 0.2 centred at the spectroscopic redshift. The choice of redshift resampling ensures that the resampled redshifts are not much larger than that measurable by the SDSS. We do not expect the results of this analysis to differ widely with other choices of redshift resampling. We finally pass the spectroscopic redshift, the slightly resampled magnitudes and the new redshift to K-CORRECT. K-CORRECT computes an estimate of the apparent magnitudes that the galaxy would have if it were at the new redshift.

We apply the same apparent magnitude and redshift selection as in Section 2.3. The final sample sizes of K-CORRECT augmented data are 435 172 (532 710) for the DR8 (DR10) analysis of which we again form a training set of size 66 per cent and a cross-validation set of size 33 per cent. We again note that we do not require a test sample for the augmented data. For brevity we refer to the K-CORRECT augmented data sample as 'augmented' data in what follows.

2.5 Visualizing the data samples

In Fig. 1 we present the *r*-band magnitude and redshift distributions of the different data samples. The left-hand (right-hand) panels correspond to the DR8 (DR10) analysis. We show the bright and faint observed galaxy samples, and the apparent magnitude limit separating these samples by the dashed line. We also show the simulated galaxy samples in grey and the augmented data in blue. For clarity we have only plotted a randomly selected subsample of the full data sets. The data points in the top panels show the redshift distribution against apparent magnitude, and bottom panels show density contours of the redshift distribution against absolute magnitude, as further determined by SDSS K-CORRECT.

We note that both the K-CORRECT augmented data and the simulated data occupy the same apparent magnitude space as the faint galaxy sample. It is due to this overlap that these augmented data sets are able to improve the redshift estimates of the faint test samples.

3 ANALYSIS AND RESULTS

We first briefly introduce the machine learning architecture employed in this work and then demonstrate how the addition of augmented data can improve the estimated machine learning redshifts for the DR8 and DR10 sets of analyses. We then show how the machine learning redshifts degrade as we push the analyses deeper past the apparent magnitude limits of the training sets. This is equivalent to extrapolating further into the unknown.

3.1 Tree-based methods

The SCIKIT-LEARN (Pedregosa et al. 2011) package written in PYTHON has a very efficient implementation of decision trees for regression (Breiman et al. 1984). The tree-based machine learning architecture recursively partitions the input feature dimensions into an increasing number of bins. Each bin is chosen to minimize the scatter of the output feature, which for these purposes is the spectroscopic redshift. This results in data with very similar spectroscopic redshifts being within the same, or possibly nearby bins.

The power of tree-based methods is enhanced by combining many trees. One technique to do this is called adaptive boosting or AdaBoost (Drucker 1997; Freund & Schapire 1997), which adds trees sequentially to generate an ensemble of trees. In the following we will refer to this sample as a forest, but this term should not be confused with the term 'random forest' which instead builds trees simultaneously. AdaBoost weighs each new tree by its ability to predict redshifts correctly, and decides how trees are grown such that redshift estimates are improved for the data with poorly estimated redshifts.

We choose to vary the following sets of hyperparameters, for a single decision tree: the number of data on each leaf node; for AdaBoost: the loss function and the number of trees; while training: the training data set (e.g. the bright sample or augmented data sample) and the size of the randomly selected training sample. For more details about combining trees with AdaBoost we refer the reader to Hastie, Tibshirani & Friedman (2001).[2]

[2] statweb.stanford.edu/tibs/ElemStatLearn

Table 1. The values of the median μ and dispersion σ_{68}, and the outlier rate of the redshift scaled residuals calculated using the test set with 41 000 faint SDSS DR8 galaxies. The top row has no data augmentation, the last row presents the ideal case and the rows in between use different augmented data sets for training and cross-validation.

Training	Cross-validation	μ	σ_{68}	Out. rate
Bright	Bright	0.0065	0.0312	2.8 per cent
Simulations	Simulations	0.0179	0.0388	5.65 per cent
Bright+augmented+simulations	Bright	0.0007	0.028	1.71 per cent
Bright+augmented+simulations	Simulations	0.0002	0.028	1.8 per cent
Bright+augmented	Augmented	−0.0001	0.0273	1.9 per cent
Bright+augmented+simulations	Bright+augmented+simulations	0.0002	0.0279	1.77 per cent
Bright+augmented+simulations	Augmented	0.0008	0.0281	1.74 per cent
Faint	Faint	0.0	0.024	1.51 per cent

Table 2. The same as Table 1 but for the analyses using 100 000 faint galaxies from SDSS DR10 as the test sample.

Training	Cross-validation	μ	σ_{68}	Out. rate
Bright	Bright	0.0062	0.0349	1.76 per cent
Simulations	Simulations	0.011	0.0554	2.88 per cent
Bright+augmented+simulations	Bright	−0.0017	0.0342	1.69 per cent
Bright+augmented+simulations	Simulations	−0.0016	0.0341	1.74 per cent
Bright+augmented	Augmented	−0.0038	0.0339	1.68 per cent
Bright+augmented+simulations	Bright+augmented+simulations	−0.0025	0.0338	1.67 per cent
Bright+augmented+simulations	Augmented	−0.002	0.0335	1.73 per cent
Faint	Faint	−0.0026	0.0315	1.48 per cent

We note that using an exponential loss function with AdaBoost has been shown to behave poorly in the presence of classification noise (see e.g. Dietterich 2000). In this work we explore the three different loss functions available within the SCIKIT-LEARN implementation, and we provide the final choice of hyper-parameters values in Section 3.2.

3.2 The effect of augmenting the data set

We perform separate analyses for the training and cross-validation bright data set and augmented data sets. We furthermore analyse different combinations of training and combined cross-validation samples, e.g. combinations of simulations and bright data, or simulations and K-CORRECT augmented data. We perform two independent sets of analyses, first on the SDSS DR8 catalogue, with bright training and faint test samples defined by $r = 18.5$, and then on the SDSS DR10 samples, with bright training and faint test samples defined by $r = 20.5$.

During the analyses we generate 200 distinct forests for each combination of training and cross-validation samples. For each forest we randomly choose the hyperparameters, and draw a random sample of random size from the training sample for training. Once the forest has been trained we input the cross-validation sample to obtain a machine learning redshift z, and use this to determine the redshift scaled residuals, $\Delta_{z'} = (z - z_{\rm spec})/(1 + z_{\rm spec})$. We calculate the value σ_{68} from $\Delta_{z'}$ which is the value of the dispersion that encloses 68 per cent of the $\Delta_{z'}$, and is analogous to the standard deviation for Gaussian statistics. We calculate the value σ_{68} using the cross-validation set and select the forest with the smallest value as the winning forest.

At test time the faint test data is passed through the winning forest to obtain a redshift estimate. We next calculate $\Delta_{z'}$ and σ_{68} as before, and additionally determine the outlier rate, defined as the percentage of data with $|\Delta_{z'}| > 0.15$ (following e.g. Hildebrandt et al. 2010), and the median μ of the distribution of Δ'_z. We use test samples of size 33 per cent of the size of the faint galaxy sample size, corresponding to 41 000 for the faint SDSS DR8 galaxies and 100 000 for the faint SDSS DR10 galaxies.

We present the results of the analysis performed on the SDSS DR8 sample in Table 1. The top row shows the results of training and cross-validation on only the bright data. The quoted values are calculated on the test sample of faint galaxies, which we reiterate are never used in the training and cross-validation stages. This is equivalent to extrapolating the redshift estimate into area of input feature space which is unexplored by the training data. The extrapolation of analysis from a training set to an unrepresentative test set is poor machine learning etiquette, which is akin to extrapolating a result into the unknown. We include this analysis simply as a benchmark.

The final row of Table 1 corresponds to a standard machine learning experiment. Here we train, cross-validate and test on random samples drawn solely from the faint galaxies. This is also included as a benchmark, and shows how well one might do in an ideal machine learning experiment, but this is not the main objective of this paper. All the other rows show combinations of training and cross-validations sets. The results of an identical analysis using SDSS DR10 sample is presented in Table 2. We have also explored many other combinations such as training on the simulations, and cross-validating on the combined data, augmented data and simulations; however, the other combinations never perform substantially better or worse than the augmentation results listed in Tables 1 and 2.

We see that most of the results from the data augmentation are between that of the best possible case (faint, faint) and the worse case (bright, bright), apart from the analyses using both the simulations training sample and the simulations cross-validation sample. However, while the values in this analysis are the poorest, we should

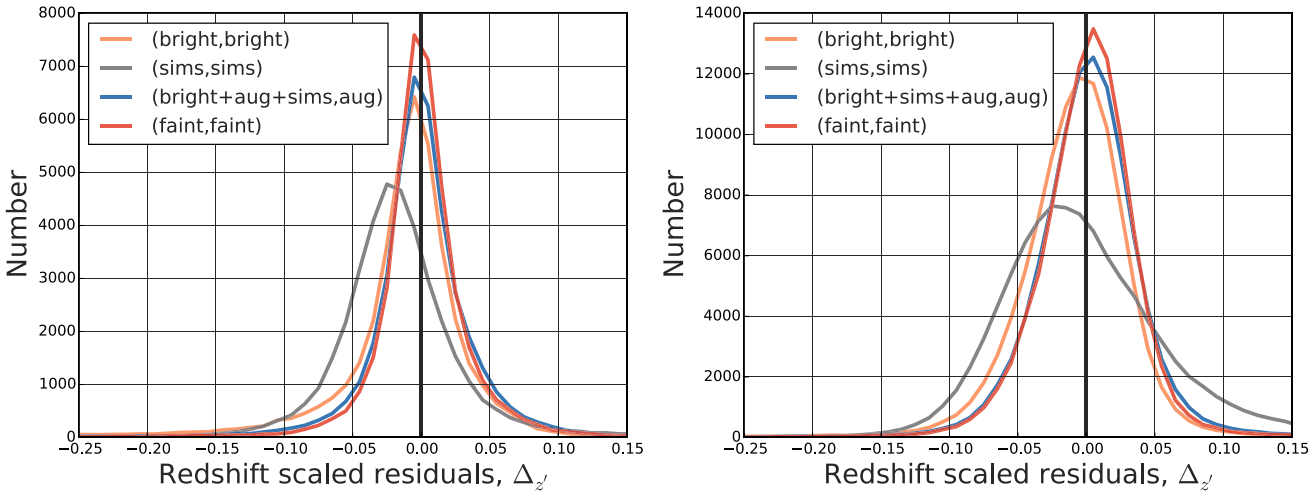

Figure 2. The distribution of test sample redshift residuals scaled by $1/(1+z)$ for different combinations of training and cross-validation samples used in the machine learning process (see legend). In the legend 'sims' denotes the simulations augmented data, and 'aug' corresponds to the K-CORRECT data augmentation of the bright galaxy sample. For clarity only a few combinations are shown. The left-hand panel corresponds to the DR8 analysis, and the right-hand panel corresponds to DR10 analysis. The lines show the results using the test sample which presents an unbiased estimate of the true redshift error because it has not been used in the training or cross-validation processes.

note that the simulations assume nothing about the data in the redshift ranges of interest. They are using stellar population physics with observational anchors at $z = 0$ and 2. In itself this is still a remarkable result. We could have ignored all observed galaxy data, and K-CORRECT augmented data and still obtain a redshift error of 0.039 for SDSS DR8 analysis (0.055 for the SDSS DR10), and an outlier rate <5 per cent (<3 per cent).

Examining the cases in both sets of analyses of training on combinations of bright data, augmented data and simulations, and using the augmented data as the cross-validation set, we find that these values improve the redshift error by 10 per cent for SDSS DR8 (4 per cent for SDSS DR10) compared to the bright data alone. We find that the outlier rate in these cases improves by 40 per cent for the SDSS DR8 analysis (and a very modest 2 per cent for the SDSS DR10 analysis).

While these absolute values are of interest, it is perhaps more interesting to identify the level of improvement that we are able to achieve on the test sample with respect to the two benchmark cases. The benchmark cases correspond to having no augmented data (bright, bright) and the ideal machine learning case (faint, faint). We determine the relative ratio of improvement for each measured statistic with respect to these benchmarks using

$$I = \frac{(\text{bright, bright}) - (\text{T, CV})}{(\text{bright, bright}) - (\text{faint, faint})}, \quad (1)$$

where (T, CV) are the different training and cross-validation samples. For the case highlighted above (bright+augmented + simulations, augmented) we find that the relative ratio of improvement for σ_{68} is 41 per cent for both the SDSS DR8 and DR10 analyses. The relative ratio of improvement for the outlier rate is 83 per cent for the SDSS DR8 sample and 11 per cent for the SDSS DR10 sample.

This means using data augmentation we are able to improve the redshift estimates from the worst case (bright samples) and recover up to 41 per cent (for σ_{68}) of the possible improvement that we may hope to achieve if we had the ideal case.

This analyses shows the power of using both augmented data and simulated data, when estimating the redshift of galaxies in the cases when one has a training sample which is not representative of the test sample. For completeness we state the forest hyperparameters of the best-fitting system using the data, simulation, augmented training and augmented cross-validation case. For DR8 this is MINLEAF:24,NUMTREES:19, NUMTRAINEXAMPLES:400320,LOSS:EXPONENTIAL and for DR10 this is MINLEAF:3,NUMTREES:56, NUMTRAINEXAMPLES:659765,LOSS:SQUARE.

We note that the smallest redshift errors in both sets of analyses, occur when the training and cross-validation samples are truly representative of the test sample, i.e. in the ideal case. We do not find that the addition of augmented data to the ideal case improves the recovered redshift estimates.

In Fig. 2 we show the distribution of $\Delta_{z'}$ for a selection of the analyses listed in Tables 1 and 2. We note that all distributions are both more peaked and have longer tails than a Gaussian distribution, which motivates our choice of σ_{68}. Also note the offset of the peak in the distribution when using the simulations as training and cross-validation sets, this can also been seen in the results Tables 1 and 2.

In Fig. 3 we show the machine learning redshift of the faint sample of test SDSS DR10 galaxies against the spectroscopic redshift, for different combinations of training and cross-validation samples, as shown in the legend of each panel. We find that the scatter is the smallest in the ideal case (faint, faint) of training and cross-validating on the faint galaxies, the simulations has the largest scatter, and the augmented data case has a scatter between the no augmentation case (bright, bright) and the ideal case. We show the redshift range $0.4 < z < 0.8$ which contains most of the faint galaxies.

Finally, we compare these results obtained using data augmentation applied to the DR8 (DR10) faint samples, with the photometric redshifts available from within SDSS CasJobs (Budavári et al. 2000; Csabai et al. 2007; Abazajian et al. 2009, using a hybrid k-NN and template approach) for the same galaxies. Using the SDSS photometric redshifts we find the values $\sigma_{68} = 0.026$ (0.037) and an outlier rate of 2.30 per cent (2.70 per cent) for the DR8 (DR10) analyses. For DR10 the measured values of σ_{68} are improved using data augmentation but for DR8 the values of σ_{68} are similar. In both DR8 and DR10 we find that the outlier fraction is reduced using

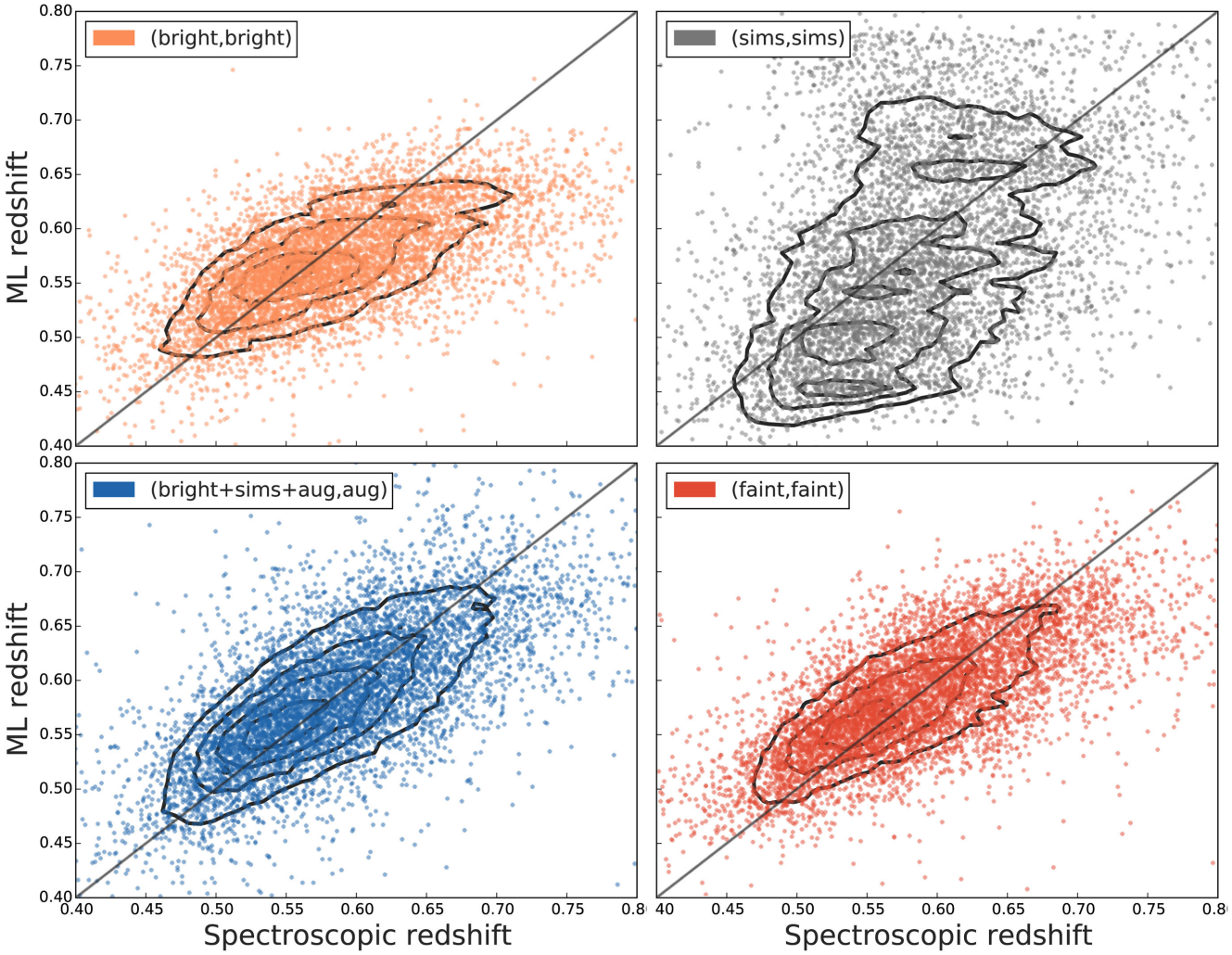

Figure 3. The machine learning redshift of the faint sample of SDSS DR10 test galaxies against the spectroscopic redshift, for different combinations of training and cross-validation samples as shown in the legend of each panel. See Fig. 2 for keyword definitions. The test galaxies are not used during training, and represent the ability of the learned system to estimate redshifts for the test galaxies which fall outside of the original (bright) training set.

the data augmentation procedure and forests. Therefore we conclude that, remarkably, we find that assuming no knowledge of real galaxies, but using data augmentation, actually improves the redshift estimates compared with the standard SDSS machine learning photometric redshifts which does train on real galaxies. These results are probably more due to the machine learning architecture used than the data augmentation process, see e.g. fig. 4 of Hoyle et al. (2015).

3.2.1 Nearest neighbour selection of augmented data

We further explore the effect of more carefully selecting all of the augmented data sets such that their input features (i.e. the magnitudes and colours) are similar to the faint test sample. The samples are chosen by selecting the three nearest neighbours in each of the training and cross-validation augmented data sets, to each of the faint galaxies in the test set. The k-NN nearest neighbour search algorithm (see e.g. Altman 1992) is used to perform the matching. The above analysis is then repeated and the results are analysed. We find that this method of carefully selecting the training and cross-validation samples does not noticeably affect the final results.

3.3 The effect of probing to deeper magnitudes

We next explore how the recovered redshifts degrade as we probe to increasingly deeper *r*-band magnitudes, past the artificially imposed magnitude limit of the training sample. This is analogous to extrapolating deeper into the unpopulated *r*-band magnitude and corresponding colour dimensions, while relying more heavily on the data augmentation to provide estimates for how galaxies in these parts of input feature space scale with redshift.

For this section we again perform two independent analyses using SDSS DR8 and DR10 data. However we now use the entire faint galaxy sample as the test sample. This corresponds to 124 000 galaxies for SDSS DR8 and 300 000 galaxies for SDSS DR10. We can use the full test sample because the faint samples are not used during the training and cross-validation, and we will not compare results with the ideal case. In both sets of analyses we pass the faint test data through the best forest found in the previous section determined using the bright data, augmented data and simulations as a training sample, and the augmented data as a cross-validation sample.

We group the test data into bins of apparent magnitude *r*, and construct $\Delta_{z'}(r)$ by comparing the machine learning redshift of the

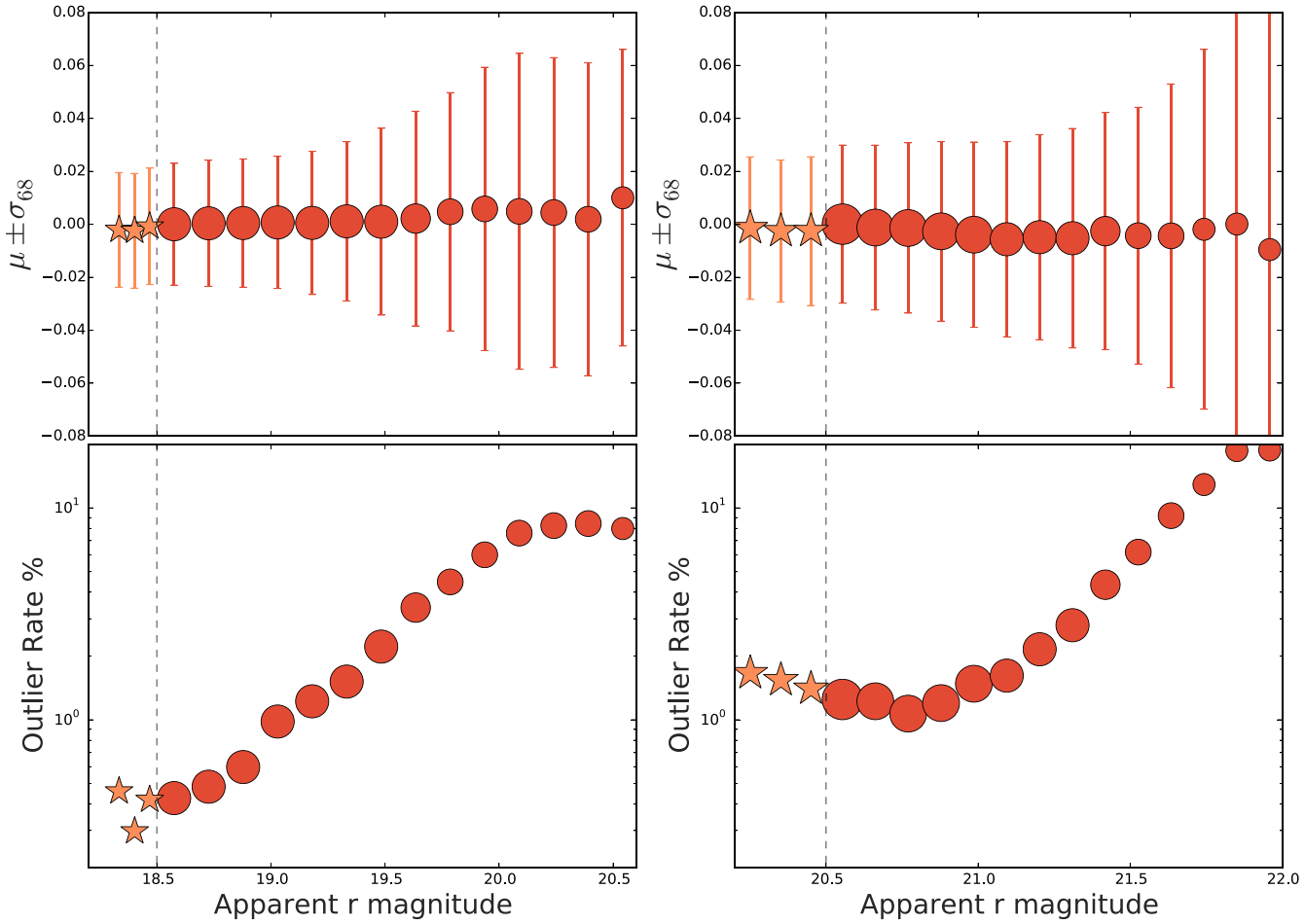

Figure 4. The effect on the recovered redshifts as measured on the faint galaxy sample as we probe to increasingly deeper r-band magnitudes, past the artificially imposed magnitude limit of training sample (dashed line). The left-hand (right-hand) panels show the SDSS DR8 (DR10) analysis. Top panel shows the median value μ, and the dispersion measured by σ_{68} of the redshift scaled residual distributions $\Delta_{z'}$. The lower panels show the outlier rate defined as $|\Delta_{z'}| > 0.15$. The starred data points to the left of the vertical dashed line are measured from the bright data, without data augmentation. The data points to the right of the vertical dashed line are measured from the faint data using data augmentation. The area of the symbol is proportional to the square root of the number of test set galaxies in the magnitude bin.

galaxies in each bin with the spectroscopic redshift. We measure μ, σ_{68} and the outlier rates from $\Delta_{z'}(r)$, and present the results in Fig. 4. The SDSS DR8 analysis is again shown in the left-hand panels, and the DR10 analysis in the right-hand panels. The area of the plotting symbol is proportional to the square root of the number of data in each magnitude bin.

To make the results of these analyses more applicable to other data sets, we construct benchmark samples from the bright galaxy data sets. The benchmark sample corresponds to bright galaxies near the artificially imposed boundary. For the DR8 analysis we construct a benchmark sample from the bright galaxies within the magnitude range $18.2 < r < 18.5$ which consists of 225 000 galaxies, and for the DR10 analysis the benchmark is constructed from galaxies within the magnitude range $20.3 < r < 20.5$ and consists of 175 000 galaxies. We construct random samples of training, cross-validation and test samples of size (50, 25 and 25 per cent). We perform a standard machine learning redshift analysis on the benchmark samples, similar to that described in Section 3.2 but without using the data augmentation. The results of the benchmark analysis are shown in the panels of Fig. 4 and correspond to the starred data points to the left of the vertical dashed line, which marks the divide between the bright and faint galaxy samples.

The top panels in Fig. 4 show how the values of the median μ and σ_{68} change as we probe to deeper r-band magnitudes past the magnitude limit of the bright sample. As we might expect, we find that the recovered values of σ_{68} degrade the deeper we probe past the magnitude limit of the training sample. We find that using data augmentation results in a well controlled, small valued bias for the redshift estimates even as we probe more than 1.5 mag deeper past the artificial magnitude limit. This is true for both sets of analyses.

The bottom panels of Fig. 4 show that the outlier rates also increase as we move to deeper magnitudes. This effect is again seen in both sets of analyses. Comparing the starred points of the benchmark sample to the left of the dashed line with the data augmentation analyses to the right of the dashed line, we find that the outlier rate for the DR8 (DR10) analysis is very similar directly across the magnitude limit, but steadily degrade to within a factor of 5 (2) at a depth of 1 mag past the limit.

We remark on the success of data augmentation to estimate redshifts of galaxies which are fainter than the original training set. We find that for all magnitude bins examined here the size of σ_{68} is within a factor of $\approx$2 of the boundary data set, when one attempts to measure galaxy redshifts 1 mag deeper than the limit of the training sample. The errors degrade further if one attempts to

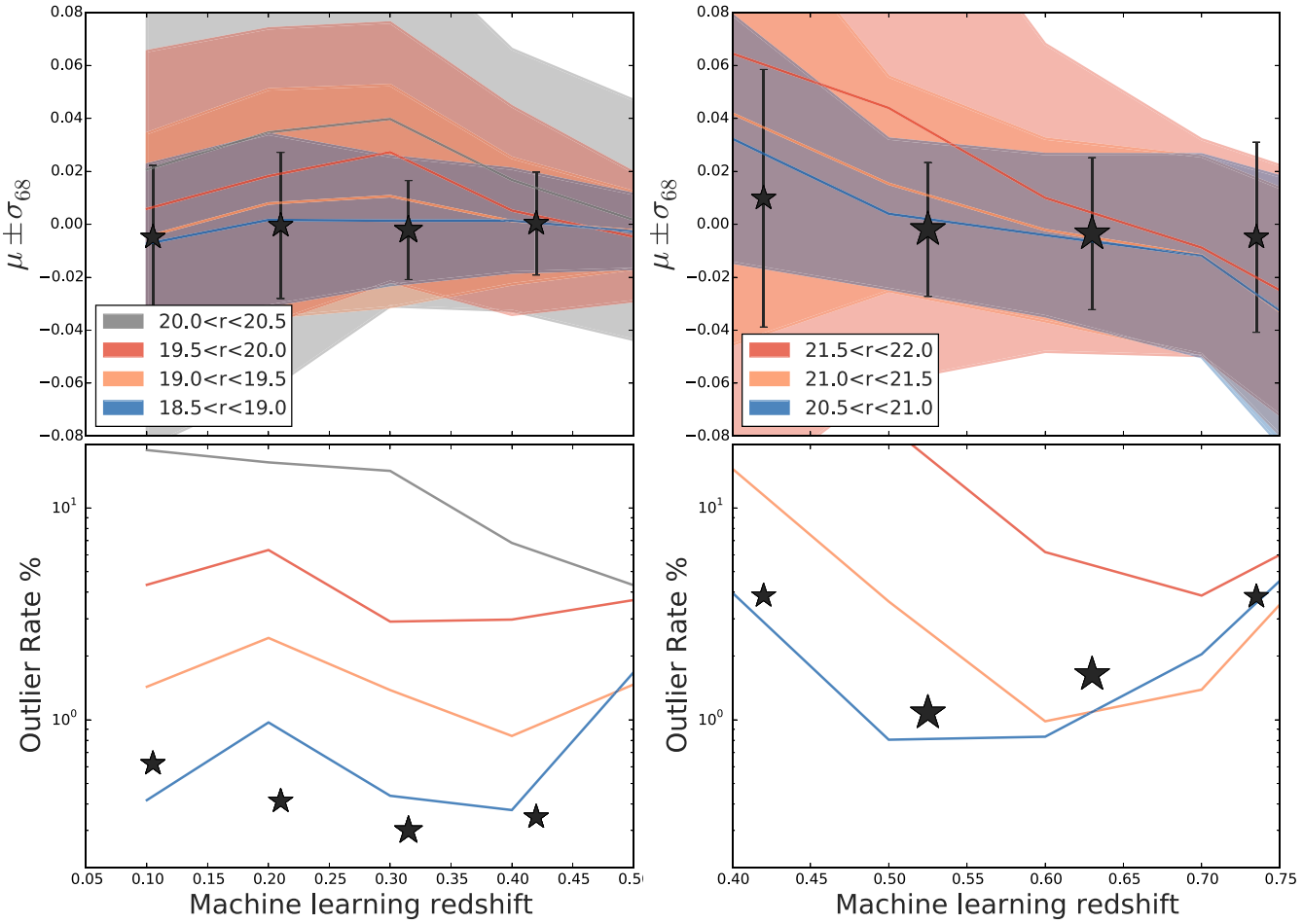

Figure 5. The effect on the recovered redshifts as measured on the faint galaxy sample as a function of machine learning redshift and the apparent r-band magnitude. The starred data points show the results of the same benchmark sample as Fig. 4 subdivided by redshift not apparent magnitude and are measured from the bright data, without data augmentation. The left-hand (right-hand) panels show the SDSS DR8 (DR10) analysis. Top panels show the median value μ, and the shaded region shows the dispersion measured by σ_{68} of the redshift scaled residual distributions $\Delta_{z'}$. The lower panels show the outlier rate defined as $|\Delta_{z'}| > 0.15$.

recover redshift estimates 1.5 and 2 mag deeper than the training and cross-validation samples. These results suggest that we can use data augmentation to explore past the magnitude limit of a training galaxy sample, if we are willing to accept a degradation in the size of the recovered redshift error. However, we are likely to produce redshift estimates with low bias.

3.4 The effect of probing to deeper magnitudes as a function of redshift

We next examine the effect of the recovered machine learning redshift as a function of both apparent r-band magnitude and the estimated machine learning redshift, z. We calculate $\Delta_{z'}(z, r)$ in redshift slices of width 0.1 and apparent magnitude bins of width 0.5 past the artificially imposed magnitude limit, and show these results in Fig. 5. The chosen magnitude bin ranges are shown in the legend, and the left-hand panels show the results for the SDSS DR8 analysis and the right-hand panels show the results for SDSS DR10. We again calculate benchmark values of $\Delta_{z'}(z, r)$ using bright data which is just brighter than the apparent magnitude limit, see Section 3.3, but now subdivided this sample by machine learning redshift not apparent r-band magnitude. We show these benchmark values as a function of redshift using the starred data points in Fig. 5.

Fig. 5 indicates that the measured statistics at each redshift of the benchmark sample, most closely resemble those of the faint galaxy sample which are closest to the artificial limiting magnitude. In particular the data augmentation applied to galaxies with a r-band magnitude up to 0.5 mag deeper than the benchmark sample is very well controlled. The median, spread of the data σ_{68} and outlier fraction of $\Delta_{z'}$ differ little from the starred data points. Examining the results of data in the magnitude bin $0.5 < r < 1.0$ mag deeper than the benchmark sample, we find that all measured statistics of $\Delta_{z'}$ degrade. The outlier fraction of this sample is increased by a factor of a few for the SDSS DR10 analysis to a factor of 10 for the SDSS DR8 analysis.

Using data augmentation to extrapolate to deeper apparent magnitudes shows that both the errors and bias increase and the outlier fraction increases by more than an order of magnitude compared to the bench mark sample.

4 DISCUSSION

We next discuss and interpret these results in terms of photometric depth and SED templates, and with respect to other works.

Examining Fig. 3 we find that using K-CORRECT templates to augment the bright data improves the redshift estimates of the faint

test sample. This suggests that the galaxy population which we create using K-CORRECT is a reasonable approximation to that of the true underlying galaxy population. This also implies that the change in the galaxy populations are smooth across the redshift ranges and apparent magnitude depths explored. A similar result is described in Brimioulle et al. (2013, see their fig. 3) using shallow Canada–France–Hawaii Telescope Legacy Survey-Wide (CFHTLS-W) photometric data, together with deep photometric CFHTLS-Deep (CFHTLS-D) data and spectroscopic samples. The templates were chosen to optimally match photometric redshift estimates to spectroscopic ones for CFHTLS-W galaxies with magnitudes $i < 22$. If the same templates were then applied to a fainter CFHTLS-W sample with $22 < i < 24$, the redshift accuracy deteriorated. However when (for overlapping fields) the CFHTLS-W photometry was replaced with that of the deeper CFHTLS-D survey, the redshift accuracy improved and reached the same value as for the bright, $i < 22$ CFHTLS-W sample. The authors conclude that the limiting factor in measuring photometric redshifts for faint galaxies is the signal-to-noise ratio (or depth) of the photometry and not the template set used.

The peak of the redshift-scaled residual histogram (Fig. 2) for the simulations is slightly offset from the line $x = 0$ and from the other distributions. One can think of the simulations as being a realization of the template methods for galaxies at different redshifts, and with a range of physical properties. The semi-analytic models use SED templates which encode stellar physics and our knowledge about galaxy evolution to produce realizations of these models as simulated galaxies. This slight redshift offset noted above has been seen by others when estimating photometric redshifts using templates methods applied to SDSS galaxies (Greisel et al. 2013).

We note that semi-analytic models are not observed data: they use models to extrapolate observed galaxy properties between redshifts, and therefore may not encode real observational effects e.g. from the evolution of multiband colours and number densities. Another reason for a colour mismatch in the semi-analytic models can arise if recipes which relate the star formation and feedback processes to merging of haloes are not correct. Furthermore, the simple stellar population models often assume a single epoch of star formation. However, the fact that data augmentation using simulations provides reasonable redshift estimates suggest that the current models do provide a good approximation to these effects.

Finally, we note that none of the data augmentation methods is better than using nature itself; the (faint, faint) ideal case produces the best of the machine learning redshift estimates examined here. This reaffirms that we can still improve our stellar populations models, and templates in order to more closely mimic galaxies observed in nature (see also Tojeiro et al. 2011; Greisel et al. 2013).

5 CONCLUSIONS

Photometric surveys can be maximally exploited for large-scale structure analysis once galaxies have been identified and their positions on the sky and in redshift space measured. Very accurate spectroscopic redshifts are only measured on a small and often biased subset of the full photometric galaxy sample due to the integration times required to obtain a reliable measurement, and the spectroscopic magnitude limit of the survey typically being shallower than the photometric detection magnitude limit.

This implies that if one attempts to estimate photometric redshifts of all galaxies using a machine learning architecture, one may be applying the results of a spectroscopic training sample which is not fully representative of the final photometric test sample. For the SDSS (York et al. 2000) this corresponds to a limit in the apparent r-band magnitude.

An alternative to machine learning methods is template methods, which can also estimate the photometric redshift of galaxies of all magnitudes, including those which are deeper than a spectroscopic training sample, however, such a sample need not even exist. The templates encode our physical knowledge of galaxy colour evolution through stellar population models.

Remaining within the machine learning framework, one may use the knowledge gained from stellar population models to simulate galaxy properties using semi-analytic models, and augment (or compliment) the original spectroscopic training sample using the simulated galaxies. The reason to augment the training and cross-validation data is to ensure that they populate the same input feature parameter space as the final test sample. Even though this augmentation is a 'best guess' of how the galaxies should appear, the process uses testable physics (if the templates are based on stellar population models), and it is still preferable to not having any training or cross-validation data in the input feature parameter space occupied by the test data.

Another approach to augment observed data is by using the public SDSS K-CORRECT package (Blanton & Roweis 2007). One may even use spectra obtained by other sources, and estimate their magnitudes as if they were to have been observed in the photometric survey in question. This last approach is being actively explored by the authors (see also Vanzella et al. 2004). If only the feature space number density of training galaxies is biased compared to the test galaxies, one may use galaxy reweighting schemes (e.g. Lima et al. 2008; Cunha et al. 2009) or the covariate shift method (Sánchez 2014).

In this paper we present a study of the effect on the recovered machine learning redshift applied to a non-representative sample of test galaxies which are selected to be fainter in the r band than the training sample of galaxies. We perform two sets of analyses, drawing on 800 000 galaxies from the SDSS DR8 and 1.7 million galaxies from SDSS DR10. The main difference between these data samples is that the DR10 sample probes higher redshifts and deeper r-band magnitudes. We apply a r-band apparent magnitude limit of 18.5 (20.5) for the DR8 (DR10) galaxies to identify bright training and bright cross-validation samples, and faint test samples.

We augment the bright galaxy training and cross-validation data in both sets of analyses using the latest semi-analytic models (Henriques et al. 2014) which separately incorporate two different stellar populations models (Bruzual & Charlot 2003; Maraston 2005) applied to the Millennium Simulation (Springel et al. 2005). We further augmented the bright galaxy training and cross-validation samples using the SDSS K-CORRECT package, which estimates the apparent magnitude of an observed galaxy if it were to exist at a different redshift.

By choosing to perform data augmentation with existing simulations and with K-CORRECT we are restricted to the available input features. These features are magnitudes and colours. Recent work has shown that the use of additional features can lead to improvements in the recovered redshift estimates (Hoyle et al. 2015). It would be interesting to determine if the construction of additional input features such as radii and galaxy shapes could also lead to a further improvement in machine learning redshift estimates through the use of data augmentation.

We explore combinations of training and cross-validation samples, for example by using the bright data, K-CORRECT augmented data and simulated data as a training sample, and the K-CORRECT augmented data as a cross-validation sample. For each

combination we tune the hyperparameters of the machine learning architecture. Finally, we pass the faint test data into the machine to estimate a machine learning photometric redshift z and calculate $\Delta_{z'} = (z - z_{\rm spec})/(1 + z_{\rm spec})$. We determine the value σ_{68} which contains 68 per cent of the data about the median value μ of $\Delta_{z'}$, and also calculate the outlier rate defined as the fraction of data with $|\Delta_{z'}| > 0.15$.

The machine learning architecture is chosen to be decision tress for regression, of which many trees are grown with the algorithm AdaBoost (Breiman et al. 1984; Drucker 1997; Freund & Schapire 1997) which we refer to as a forest. We define two benchmark samples corresponding to the worst case (no data augmentation) and the ideal case (the training, cross-validation and test sample are all drawn from the faint sample). We present our results with respect to these two benchmark cases.

We find that the use of the augmented data sets improves the error σ_{68}, on machine learning redshift estimates by 41 per cent in both DR8 and DR10 sets of analyses, when compared to the ideal case. This means that using data augmentation we are able to improve the redshift estimates from the worst case (bright samples) and recover up to 41 per cent of the possible improvement that we may hope to achieve if we had the ideal case (faint samples).

We find the outlier rate is also improved by 80 per cent for the SDSS DR8 analysis and 10 per cent for the SDSS DR10 analysis. It is satisfying to note that using only the simulated galaxies as training and cross-validation samples we still recover a reasonable photometric error of $\approx 1.7 \times \sigma_{68}^{\rm ideal}$, and an outlier rate of only a factor of 2 (4) higher than the ideal case when applied to the SDSS DR8 (DR10) faint test sample. This shows how accurately semi-analytic models can replicate the magnitudes, colours and redshifts of observed galaxies.

We also compare these results with the photometric redshifts available from within SDSS CasJobs (Budavári et al. 2000; Csabai et al. 2007; Abazajian et al. 2009) for the same galaxies. The SDSS photometric redshifts are trained on real galaxies, while data augmentation trains only on augmented galaxies. For DR10 the measured values of σ_{68} are improved by 10 per cent using data augmentation. However for DR8 the values of σ_{68} remain very similar. In both DR8 and DR10 we find that the outlier fraction is reduced by $\approx$30 per cent using the data augmentation procedure and forests.

We next explore the change in the recovered redshift errors and outlier rate, as we probe to increasingly deeper r-band magnitudes, past the artificially imposed magnitude limit of the bright training samples. This makes the training and test samples more dissimilar. We find that as one pushes deeper past the r-band magnitude limit, both the error σ_{68} and the outlier rate slowly degrade. We note that the median values of the $\Delta_{z'}$ are always close to 0, and therefore the combination of data augmentation and tree-based methods produce only a small value of bias in the estimate of the true galaxy redshift, even when applied to an unrepresentative test sample of galaxies. We do however note that the level of bias reported here (using the SDSS data set) is larger than the science requirements of the Dark Energy Survey (The Dark Energy Survey Collaboration 2005). Further investigation is underway by the authors to understand if data augmentation can achieve the specified level of precision.

Applying these analyses to galaxies with an r-band magnitude up to 0.5 mag deeper than the artificially imposed magnitude limit of the benchmark training samples result in very well controlled statistics for all machine learning redshifts ranges explored. The median, spread of the data σ_{68} and outlier fraction of $\Delta_{z'}$ differ little from the benchmark samples. Examining the results of data in the magnitude bin $0.5 < r < 1.0$ mag deeper than the benchmark sample, we find that all measured statistics of $\Delta_{z'}$ degrade. The outlier fraction of this sample is increased by a factor of a few for the SDSS DR10 analysis to a factor of 10 for the SDSS DR8 analysis.

Although one should not extrapolate these results to new or different data sets, we do expect data augmentation to also improve other similar analysis. In this paper we have concentrated on the improvement to the point estimate of the redshift while using data augmentation. The full shape of the redshift distribution function, or the conditional probability distribution function for individual galaxies is also important quantities which we will examine in future work (Rau et al., 2015).

Finally, we conclude that the use of data augmentation presents an accessible way to include our knowledge of the Universe, and in particular the magnitude evolution of galaxies, into the machine learning training and cross-validation samples. This is particularly important if the training and cross-validations samples are drawn from non-representative samples of the final test data. We have shown that this method works well for SDSS galaxies using two sets of analyses. Similar analysis could be performed on other data sets and surveys, in particular for surveys with existing dedicated simulations.

ACKNOWLEDGEMENTS

The authors thank the anonymous referee for comments and suggestions which have improved the paper. BH would like to thank Kerstin Paech for useful discussions. SS and MMR are supported by the Transregional Collaborative Research Centre TRR 33 – The Dark Universe and the DFG cluster of excellence 'Origin and Structure of the Universe'. Funding for this project (CB) was partially provided by the Spanish Ministerio de Economa y Com- petitividad (MINECO) under projects FPA2013-47986, and Centro de Excelencia Severo Ochoa SEV-2012-0234. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the US 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 Millennium Simulation data bases used in this paper and the web application providing online access to them were constructed as part of the activities of the German Astrophysical Virtual Observatory.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Ahn C. P. et al., 2014, ApJS, 211, 17
Aihara H. et al., 2011, ApJS, 193, 29
Altman N. S., 1992, Am. Stat., 46, 175
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Bonnett C., 2013, preprint (arXiv:e-prints)
Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees. Wadsworth International Group, Belmont, CA
Brimioulle F., Seitz S., Lerchster M., Bender R., Snigula J., 2013, MNRAS, 432, 1046
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Budavári T., Szalay A. S., Connolly A. J., Csabai I., Dickinson M., 2000, AJ, 120, 1588
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Cohen J. G., Hogg D. W., Blandford R., Cowie L. L., Hu E., Songaila A., Shopbell P., Richberg K., 2000, ApJ, 538, 29

Cristiani S. et al., 2000, A&A, 359, 489
Csabai I., Dobos L., Trencséni M., Herczegh G., Józsa P., Purger N., Budavári T., Szalay A. S., 2007, Astron. Nachr., 328, 852
Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, MNRAS, 396, 2379
Dahlen T. et al., 2013, ApJ, 775, 93
Dietterich T. G., 2000, Machine Learning, 40, 139
Drucker H., 1997, in Fisher D. H., ed., Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97. Morgan Kaufmann Publishers Inc., San Francisco, p. 107
Eisenstein D. J. et al., 2011, AJ, 142, 72
Fioc M., Rocca-Volmerange B., 1997, A&A, 326, 950
Freund Y., Schapire R. E., 1997, J. Comput. Syst. Sci., 55, 119
Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823
Greisel N., Seitz S., Drory N., Bender R., Saglia R. P., Snigula J., 2013, ApJ, 768, 117
Gunn J. E. et al., 2006, AJ, 131, 2332
Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York
Henriques B. M. B., White S. D. M., Lemson G., Thomas P. A., Guo Q., Marleau G.-D., Overzier R. A., 2012, MNRAS, 421, 2904
Henriques B., White S., Thomas P., Angulo R., Guo Q., Lemson G., Springel V., Overzier R., 2014, preprint (arXiv:e-prints)
Hildebrandt H. et al., 2010, A&A, 523, A31
Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015, MNRAS, 449, 1275
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001, ApJ, 556, 121
Lahav O., 1997, in Di Gesu V., Duff M. J. B., Heck A., Maccarone M. C., Scarsi L., Zimmerman H. U., eds, Data Analysis in Astronomy. World Scientific Press, Singapore, p. 43
Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118
Maraston C., 2005, MNRAS, 362, 799
Pedregosa F. et al., 2011, J. Machine Learning Res., 12, 2825
Rau M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, preprint (arXiv:1503.08215)
Sánchez C. et al., 2014, MNRAS, 445, 1482
Smith J. A. et al., 2002, AJ, 123, 2121
Springel V. et al., 2005, Nature, 435, 629
Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, in Apolloni B., Marinaro M., Tagliaferri R., eds, Lecture Notes in Computer Science, Vol. 2859, Neural Nets. Springer-Verlag, Berlin, p. 226
The Dark Energy Survey Collaboration 2005, preprint (astro-ph/e-prints)
Tojeiro R., Percival W. J., Heavens A. F., Jimenez R., 2011, MNRAS, 413, 434
Vanzella E. et al., 2004, A&A, 423, 761
Wolf C., 2009, MNRAS, 397, 520
Yeche C. et al., 2009, preprint (arXiv:e-prints)
York D. G. et al., 2000, AJ, 120, 1579

APPENDIX A: MYSQL QUERIES

We obtain observation and simulated data from the SDSS and the Millennium Simulation using the following MYSQL queries.

A1 SDSS

To obtain SDSS data the following MYSQL query is run in the DR8 and then separately in the DR10 schema:

```
SELECT s.specObjID, s.objid, s.ra,s.dec,
s.z AS specz, s.zerr AS specz_err,
s.dered_u,s.dered_g,s.dered_g,s.dered_i,
s.dered_z,s.modelMagErr_u,s.modelMagErr_g,
s.modelMagErr_r,s.modelMagErr_i,s.modelMagErr_z,
s.type as specType, q.type as photpType
INTO mydb.specPhotoTable FROM SpecPhotoAll AS s
JOIN photoObjAll AS q
ON s.objid=q.objid AND q.cModelMag_u>0
AND q.dered_g>0 AND q.dered_g>0
AND q.dered_z>0 AND q.dered_i>0
```

A2 Millennium Simulation

The simulated galaxies are obtained using the following query in the Millennium Simulation MYSQL interface:

```
select galaxyId, ra, dec, z_app as spec_z,
SDSS_u as u, SDSS_g as g, SDSS_r as r,
SDSS_i as i, SDSS_z as z from
Henriques2014a.cones.MRscPlanck1_SPM_0NUM
where z_app >0.05 and z_app <1.0
```

This query is run by replacing the string `SPM` with `M05` or `BC03` to explore the different stellar population models, and the string `NUM` is replaced by the integers running from `1` to `3`.

This paper has been typeset from a TEX/LATEX file prepared by the author.