

Z-Sequence: Photometric redshift predictions for galaxy clusters with sequential random k-nearest neighbours

Matthew C. Chan¹ and John P. Stott¹

E-mails: m.c.chan@lancaster.ac.uk and j.p.stott@lancaster.ac.uk

¹*Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We introduce Z-Sequence, a novel empirical model that utilises photometric measurements of observed galaxies within a specified search radius to estimate the photometric redshift of galaxy clusters. Z-Sequence itself is composed of a machine learning ensemble based on the k-nearest neighbours algorithm. We implement an automated feature selection strategy that iteratively determines appropriate combinations of filters and colours to minimise photometric redshift prediction error. We intend for Z-Sequence to be a standalone technique but it can be combined with cluster finders that do not intrinsically predict redshift, such as our own DEEP-CEE. In this proof-of-concept study we train, fine-tune and test Z-Sequence on publicly available cluster catalogues derived from the Sloan Digital Sky Survey. We determine the photometric redshift prediction error of Z-Sequence via the median value of $|\Delta z|/(1+z)$ (across a photometric redshift range of $0.05 \leq z \leq 0.6$) to be ~ 0.01 when applying a small search radius. The photometric redshift prediction error for test samples increases by 30-50 per cent when the search radius is enlarged, likely due to line-of-sight interloping galaxies. Eventually, we aim to apply Z-Sequence to upcoming imaging surveys such as the Legacy Survey of Space and Time to provide photometric redshift estimates for large samples of as yet undiscovered and distant clusters.

Key words: galaxies: clusters: general – methods: statistical – methods: data analysis – techniques: photometric – galaxies: distances and redshifts

1 INTRODUCTION

Galaxy clusters are the most massive gravitationally bound objects to have formed in the Universe, with deep potential wells that correspond to matter density peaks (Dressler 1984; Huss et al. 1999; Kravtsov & Borgani 2012). During the past few decades, the advent of modern imaging surveys has significantly contributed to the study of large scale structure and galaxy evolution across cosmic time. These surveys generate a huge abundance of data that encourages the need for automated algorithms (e.g. da Costa et al. 1998; Falco et al. 1999; York et al. 2000; Colless et al. 2001; Jones et al. 2009; Baldry et al. 2010; Eisenstein et al. 2011; Huchra et al. 2012; Blanton et al. 2017). From which, properties of clusters such as redshift, luminosity and richness can be estimated and used as probes for astrophysics and cosmology (e.g. Howlett et al. 2015; Planck Collaboration et al. 2016; de Haan et al. 2016; Ross et al. 2017; Gil-Marín et al. 2017; Beutler et al. 2017; Alam et al. 2017; Ata et al. 2018; Joudaki et al. 2018; Amendola et al. 2018; Abbott et al. 2018a).

The Legacy Survey of Space and Time (LSST)¹, Ivezić et al.

(2019) will be the state-of-the-art imaging survey for the next decade of astronomy. It will repeatedly image the entire southern hemisphere and is forecasted to generate up to twenty terabytes of data per night over a ten year period. Due to the quantity of data involved, the development of automated algorithms for LSST will be crucial to handle extensive data processing and analysis tasks. In addition, LSST will observe at deeper depths and wider sky coverage compared to previous surveys. This would increase the redshift range and lower the mass limit sensitivity of current cluster observations, such that thousands of new clusters are likely to be discovered.

There are presently two approaches used to determine galaxy redshifts, these are through spectroscopy and photometry (e.g. Walcher et al. 2011; Piattella 2018). However whilst the former is precise it is also time-consuming, expensive and difficult to perform for faint distant sources, which limits the number of observations with spectroscopic redshifts. Alternatively, photometric redshifts are fast to acquire and have been shown to be successful for faint distant sources (e.g. Ilbert et al. 2009). Conventional methods to

at the Vera Rubin Observatory operating with six broad-band filters: u , g , r , i , z and Y .

¹ LSST will be conducted using the 8.4-meter Simonyi Survey Telescope

estimate photometric redshift involve either empirical or template fitting algorithms. Empirical algorithms learn a target function of the underlying relationships between observed brightness, colour and spectroscopic redshift from a large training sample of galaxies (e.g. [Weinstein et al. 2004](#); [Lopes 2007](#); [Carrasco Kind & Brunner 2013](#); [Bilicki et al. 2018](#); [Pasquet et al. 2019](#)). Whilst, template fitting algorithms match observed fluxes to theoretical spectral energy distributions of different galaxy types at reference redshifts (e.g. [Bolzonella et al. 2000](#); [Babbedge et al. 2004](#); [Gorecki et al. 2014](#); [Fotopoulou & Paltani 2018](#)). Nevertheless, photometric redshifts tend to have larger measurement errors than spectroscopic redshifts since photometric filters operate with low wavelength resolution, which means that individual spectral features can not be utilised to determine redshift.

Photometric redshifts are often employed by imaging surveys to provide initial redshift estimates for many galaxies (e.g. [Sánchez et al. 2014](#); [Laigle et al. 2016](#); [Beck et al. 2016](#); [Tanaka et al. 2018](#)), of which sub-samples can be followed up with spectroscopic redshifts. Similarly, it is important to develop models that will provide researchers with accurate initial redshift estimates for large and deep samples of the cluster population. In terms of predictive power for the low to intermediate redshift regime, empirical algorithms with sufficient training samples will generally outperform template fitting algorithms because template fitting algorithms require more physical assumptions when constructing spectral energy distributions to reflect possible observations. Whereas for the high redshift regime, template fitting algorithms will typically outperform empirical algorithms since high redshift training samples are more difficult to obtain due to observing limitations ([Salvato et al. 2019](#)).

In order to estimate redshifts for clusters, it is first required to identify cluster members within a given search area. This can be conducted by utilising the red sequence ([Yee et al. 1999](#); [Gladders & Yee 2000](#)), which takes advantage of the fact that ‘red’ early-type galaxies are often found in clusters ([Dressler 1980](#)). From which, the red sequence is seen as a well-defined linear relationship in colour-magnitude space (CMS) that evolves with redshift ([Stott et al. 2009](#)). This sequence is sloped such that bright cluster members are redder than their fainter counterparts. In CMS, galaxy types can be differentiated based on their underlying stellar populations into a red sequence and blue cloud region ([Jin et al. 2014](#)). Generally, the red sequence contains predominately ‘red’ elliptical and lenticular galaxies, whilst the blue cloud contains mostly ‘blue’ spiral and ‘disk’-like galaxies. However, minority exceptions do exist such as ‘red’ spiral galaxies ([Wolf et al. 2009](#)) and ‘blue’ elliptical galaxies ([Schawinski et al. 2009](#)). From which, an empirical algorithm can estimate photometric redshift based on the observed red sequence (e.g. [Hsieh et al. 2005](#); [Rykoff et al. 2014](#)). This involves training an empirical algorithm to learn the redshifts from examples of known red sequences, such that the red sequence of an unknown cluster can be interpolated by the algorithm.

Additionally in order to break any colour-redshift degeneracies, where galaxies at different redshifts could have resembling colours, multi-dimensional CMS should be employed to reduce the reliance on specific colours. For example, a single colour that only utilised short wavelength optical filters would struggle to detect the red sequence of a high redshift cluster since the filters would be unable to observe the redshifted 4000Å break², which is a distinctive broad spectral feature seen in the continuum spectrum of elliptical

galaxies ([Dressler & Shectman 1987](#)). By utilising more colours, it is possible to straddle the 4000Å break to account for its transition at different redshifts ([Gladders & Yee 2000](#); [Stott et al. 2007](#)).

For this paper, we employ an automated feature selection strategy that selects appropriate combinations of filters and colours in multi-dimensional CMS. We intend for this feature selection process to be fully data-driven based on observed galaxy photometry data, such that the selected features are effective at minimising photometric redshift prediction error. This method also comes with multiple practical benefits. Firstly, it is able to work with incomplete filter sets, as it does not rely on any specific filter. Secondly, it does not depend on galaxy photometric redshift catalogues. Thirdly, this approach can be combined with cluster finders that do not naturally predict redshift, such as DEEP-CEE ([Chan & Stott 2019](#)), since Z-Sequence only requires input astronomical coordinates and a photometry catalogue to predict photometric redshift of clusters.

We structure this paper with the following layout. In §2 we outline our methodology where §2.1 describes our data pre-processing approach, §2.2 describes our feature selection strategy plus machine learning algorithm and §2.3 describes how we train our model. In §3 we present our results where §3.1 describes the feature selection and filter magnitude-cut analysis, §3.2 describes the hyper-parameter tuning and §3.3 plus §3.4 describes the tuned model performance on test sets. In §4 we review our findings where §4.1 discusses the effectiveness of the tuned model at making predictions and §4.2 discusses the practicality of the machine learning techniques used in this paper. Finally, in §5 we summarise this paper.

We assume the Λ CDM cosmological parameters $H_0 = 71 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.27$ and $\Omega_\Lambda = 0.73$.

2 METHODOLOGY

2.1 Preparation Of Photometric Datasets

We utilise candidate clusters detected in the Sloan Digital Sky Survey III (SDSS-III, [Eisenstein et al. 2011](#)) by the WHL12 ([Wen et al. 2012](#)) and redMaPPer ([Rykoff et al. 2014](#)) cluster catalogues as part of our training, validation and test sets under a supervised learning approach ([Kotsiantis et al. 2007](#)). WHL12 uses photometric redshifts of galaxies estimated by SDSS to identify overdense regions of galaxy clustering via a grouping algorithm, in which the cluster redshift was calculated from the median value of determined cluster members. Whilst redMaPPer search for the red sequence within CMS across the SDSS sky coverage. The observed red sequence profile of highly probable cluster members was then fit with a self-trained model of template red sequences to estimate cluster redshift. It should be noted that the full WHL12 cluster catalogue has a photometric redshift range of $0.05 \leq z \leq 0.7846$ and the full redMaPPer cluster catalogue has a photometric redshift range of $0.0811 \leq z \leq 0.5983$.

Initially, we apply two selection criterion to the WHL12 cluster catalogue to identify clusters that had photometric redshifts between $0.0 < z < 0.6$ and also contain more than twenty member galaxies. This provides us with an approximation of the distribution of clusters found at different redshifts. From which, we calculate a

² The 4000Å break is caused by the blanket absorption of photons at specific wavelengths from metals in the ionised atmospheres of old stellar populations ([Kauffmann et al. 2003](#)).

² The 4000Å break is caused by the blanket absorption of photons at spe-

mean photometric redshift of $z = 0.3127$ based on the selected clusters. We use this mean photometric redshift to determine an angular distance of 54.96 arcseconds, which corresponds with the average cluster core optical radius of ~ 250 kpc (Girardi et al. 1995). This angular distance also corresponds to a radius of approximately 100 kpc at $z = 0.1$ and 334 kpc at $z = 0.5$. We then cross-match the clusters from the full WHL12 and redMaPPer cluster catalogues that are within 54.96 arcseconds and also within a photometric redshift range of $\pm 0.04(1+z)$ as used by Wen et al. (2009)³. This ensures that we cleanly separate clusters to improve signal-to-noise in the dataset. The matching and non-matching clusters are then split into the following three datasets:

- **MWAR** - Cross-matched WHL12 and redMapper clusters.
- **WNMR** - WHL12 clusters with no cross-matched redMapper clusters.
- **RNMW** - redMapper clusters with no cross-matched WHL12 clusters.

Next, we reapply our initial two selection criterion to all the clusters in the MWAR, RNMW and WNMR datasets. This splits the clusters in each dataset into distinctive redshift and richness groupings, which can be used to examine how the Z-Sequence model performs on clusters that have these different properties. We set clusters that have properties within the selection criterion limits as the main training and test sets, whilst clusters that have properties outside the selection criterion limits are used as additional test sets. From which, the number of clusters within the selection criterion limits for the MWAR dataset is 8841 with a photometric redshift range of $0.0698 \leq z \leq 0.5986$, the WNMR dataset is 9723 with a photometric redshift range of $0.05 \leq z \leq 0.599$ and the RNMW dataset is 8646 with a photometric redshift range of $0.0811 \leq z \leq 0.5983$. In addition, the observed redshift distributions and positions of clusters from each dataset can be seen in Figures 1 and SA1 (available online).

We proceed to cross-match the astronomical coordinates of clusters in each dataset to galaxies found in the SDSS-III Data Release 9 photometric catalogue (SDSS-III DR9, Ahn et al. 2012) that are within the previously defined angular distance of 54.96 arcseconds. We select ‘primary’ observations⁴ of galaxies that have ‘clean’ photometry as determined by SDSS. This catalogue provides photometric measurements⁵ for the following filters and colours:

- **Filters:** $u, g, r, i, z,$
- **Colours:** $u-g, g-r, r-i, i-z, u-r, g-i, r-z, u-i, g-z, u-z,$

where we use these filters and colours as our input features in §2.2.

We assume that any of the SDSS identified galaxies which lie

³ Wen et al. (2009) suggests that a photometric redshift gap of $\pm 0.04(1+z)$ is a suitable indicator of true cluster richness, which corresponds to a rest frame velocity range of 24000 km s^{-1} to account for the uncertainty of the photometric redshifts.

⁴ The term ‘primary’ refers to the best imaging observation recorded for a survey object if it was seen multiple times during an observing run in an SDSS plate, whilst other observations of the object are called ‘secondary’. A more in-depth explanation can be found on <http://www.sdss3.org/dr9/help/glossary.php>

⁵ SDSS ‘modelMag’ measurements are used for filter magnitudes and colours of galaxies. This approach ensures the same aperture is used for all filters and the resultant magnitudes are calculated based off the best-fit model parameters observed in the r-band. For further details see <http://www.sdss3.org/dr9/algorithms/magnitudes.php>

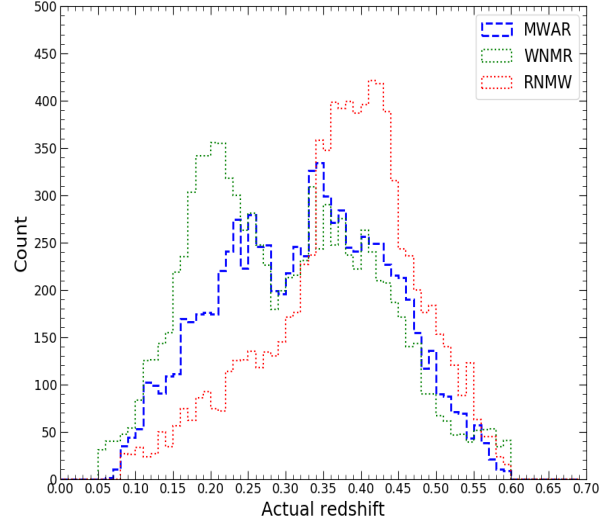


Figure 1. Frequency histogram of the ‘actual’ redshift distributions of clusters, where photometric redshifts of clusters in the MWAR (blue dashed line) and WNMR (green dotted line) datasets are originally estimated by WHL12. Whilst the photometric redshifts of clusters in the RNMW (red dotted line) dataset are originally estimated by redMaPPer.

along the line-of-sight and within 54.96 arcseconds of the input astronomical coordinates are part of the same cluster, from which we assign each individual galaxy a cluster ID number for cross-referencing. To reduce the number of interloped galaxies, we empirically set multiple search radii of approximately 50, 100 and 150 kpc at the mean photometric redshift of $z = 0.3127$, which corresponds to angular distances of 10, 21 and 32 arcseconds respectively. The number of interlopers will also depend on the position accuracy of the input cluster coordinates relative to the true cluster centroid. The reason we employ multiple search radii was to ensure that if the smallest search radius did not find a galaxy in the SDSS-III DR9 photometric catalogue, then the search radius would increase until a galaxy was found. This also provides a test for the effectiveness of the algorithm when given different views of the cluster core. It should be noted that this results in multiple forms of the training/validation/test sets that contain additional galaxies in clusters found within each search radius.

We assign the MWAR dataset as the training/validation sets and WNMR/RNMW datasets as test sets. The redshift distributions of the clusters in these datasets can be seen in Figure SA2 (available online) for each search radius. We chose the MWAR dataset as the training set since we expect that these clusters would be more likely to host a populated core, where the red sequence would be well-defined (Kodama et al. 1998; Gladders et al. 1998; de Propris et al. 1999; Lidman et al. 2008; Mei et al. 2009; Newman et al. 2014; Strazzullo et al. 2016) in comparison to clusters in the WNMR/RNMW datasets, given the nature of the methods of WHL12 and redMaPPer. We want our model to learn and utilise ‘red sequence’-like features found within high dimensional CMS to effectively predict photometric redshifts across a broad redshift range.

Finally, we investigate how varying the brightness for filter magnitude-cuts (see Table 1) could improve the accuracy of photometric redshift estimates, as this will remove galaxies from the less well-defined faint end of the red sequence that have relatively large filter magnitude errors and filter magnitude values fainter than a

Filter	LM [mag]	LM-0.5 [mag]	LM-1.0 [mag]	LM-1.5 [mag]	LM-2.0 [mag]	LM-2.5 [mag]
<i>u</i>	21.6	21.1	20.6	20.1	19.6	19.1
<i>g</i>	22.2	21.7	21.2	20.7	20.2	19.7
<i>r</i>	22.2	21.7	21.2	20.7	20.2	19.7
<i>i</i>	21.3	20.8	20.3	19.8	19.3	18.8
<i>z</i>	20.7	20.2	19.7	19.2	18.7	18.2

Table 1. This table contains the SDSS limiting magnitude (LM) values of each filter with specified magnitude-cuts. The LM values are determined from 95 per cent completeness studies of point sources⁶. The filter magnitude values shown are converted from the SDSS *ugriz* magnitude system (Lupton et al. 1999) to AB magnitude system (Oke & Gunn 1983). It should be noted that the SDSS *ugriz* magnitude system is very similar to the AB magnitude system but not exact (Doi et al. 2010), such that $u_{AB} = u_{SDSS} - 0.04$ and $z_{AB} = z_{SDSS} + 0.02$ (Abazajian et al. 2004).

specified limiting magnitude⁶ value. In addition, we also compare the performance of using filter magnitude-cuts to a control group dataset that had no filter magnitude-cuts applied.

2.2 Model Techniques

2.2.1 Feature Selection Process

It should be noted that we have a total of 32,768 possible combinations for the input features (see the filters and colours described in §2.1) that could be tested. Due to the computational cost involved to examine all these combinations, we decide to employ an automated feature selection technique known as Sequential Forward Selection (SFS, Guyon & Elisseeff 2003) to determine appropriate filters and colours. This technique is a ‘greedy’ iterative strategy that builds a subset of features via a bottom-up selection approach starting from an empty feature subset. Each iteration evaluates the performance of feature combinations, where SFS selects and stores the feature that best satisfies an objective function⁷ into the empty feature subset. From which, we employ a multi-objective function that checks if the following conditions are satisfied in each iteration of SFS:

- (i) The formula below calculates the photometric redshift prediction error:

$$E_z = \frac{|P_i - A_i|}{(1 + A_i)}, \quad (1)$$

where E_z is the photometric redshift prediction error for each tested cluster, P_i is the estimated photometric redshift for each cluster and A_i is the ‘actual’⁸ photometric redshift for each cluster. Figure SA3 (available online) shows a direct comparison of

⁶ Limiting magnitudes for the SDSS telescope are found by repeated observations of a patch of sky to obtain a magnitude value that provides 95 per cent completeness of point sources (York et al. 2000; Strauss et al. 2002; Ivezić et al. 2004). See SDSS imaging camera scope at <http://www.sdss3.org/dr9/scope.php> for magnitude limits of each filter.

⁷ An objective function is a general term used to describe a function of defined conditions that is minimised or maximised to find the optimal solution for the given objective (Goodfellow et al. 2016).

⁸ This depended on which dataset was used as the photometric redshifts of clusters in the MWAR and WNMR datasets were from the WHL12 cluster catalogue whilst photometric redshifts of clusters in the RNMW dataset were from the redMaPPer cluster catalogue.

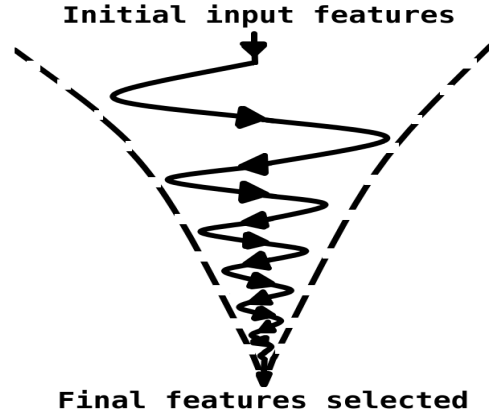


Figure 2. A simplified perspective of the SFS strategy. The solid line with black arrows indicate the path taken by SFS to select features and the dashed lines represent the boundaries of feature space. It can be seen that as SFS progresses the feature space shrinks due to the reduced number of possible outcomes, where SFS continues until it converges on a set of features. This diagram was inspired by Gutierrez-Osuna (2011).

the photometric redshifts for cross-matched clusters from the WHL12 and redMaPPer cluster catalogues, where both catalogues appear to be in good agreement.

The median of photometric redshift prediction errors produced during an iteration must be lower than the median of photometric redshift prediction errors from the previous iteration to continue SFS iterations.

- (ii) Filter magnitude-cuts are used to remove galaxies fainter than a specified magnitude threshold for each photometry filter to improve the signal-to-noise of the datasets. This can result in clusters with no galaxies remaining. We determine a percentage of clusters retained by counting the number of clusters that have galaxies remaining, after filter magnitude-cuts are applied, from the initial total in a dataset. From which, we set a threshold for the percentage of clusters retained in the MWAR dataset must be equal or greater than 95 per cent⁹ to continue SFS iterations.

In Figure 2 we observe that the SFS strategy is a computationally efficient approach as it searches through a reduced number of possible combinations, where all selected features are not included for reconsideration in subsequent SFS iterations. The process continues until the objective function is no longer satisfied with the remainder of the input features. We also compare the performance of these features to a control group of features that are not selected with SFS, where the control group features are *g*, *r*, *i*, *g-r*, *r-i*, *g-i*. We assume that the control group features would perform well since these filters and colours would likely display ‘red sequence’-like features over a wide range of redshifts in CMS (Stott et al. 2009; Rykoff et al. 2014) accounting for the shifting of the 4000 Å break (Hamilton 1985).

⁹ A tolerable percentage of data purposely excluded from the dataset should be low, otherwise systematic biases and sample misrepresentation induced by the missing data could be introduced into our analysis (Kang 2013).

2.2.2 Machine Learning Algorithm

We adopt the sequential random k-nearest neighbours (SRKNN, Park & Kim 2015) algorithm as the foundation of our model. The SRKNN algorithm is an ensemble (Dietterich 2000) that aggregates multiple k-nearest neighbours (KNN, Fix 1951; Cover & Hart 1967) models into one global model (see Figure 3). The KNN algorithm is classed as a non-parametric learning method (Webb 2010) in the field of machine learning that can be used for non-linear regression tasks. This means that the algorithm has no learnable parameters to train (e.g. weights in a neural network algorithm, McCulloch & Pitts 1943). Predictions for the KNN algorithm are produced by averaging the labelled values of the nearest neighbour training data points to the input data points, where we use the Euclidean distance metric¹⁰ to compute distances. The main characteristics of the SRKNN algorithm involve bootstrap with replacement (Efron 1979; Efron & Tibshirani 1986) of the training set and random initialisation of input features to train each internal KNN model. These traits can improve the overall accuracy of predictions as a greater variety of features would be considered for each internal KNN model.

The SRKNN algorithm has three main hyper-parameter settings that should be optimised before deployment. These hyper-parameter settings are listed as follows:

- The number of internal KNN models (also equivalent to number of bootstrap resamples used).
- The number of randomly initialised input features.
- The number of nearest neighbours.

Park & Kim (2015) suggests that the performance of the SRKNN algorithm depends on the values assigned for each hyper-parameter setting, where the optimal values vary for different datasets. In §3.2 we examine and tune each hyper-parameter setting with the MWAR validation set.

2.3 Outline Of Model Training

Here, we describe the steps used to train and test our model for each search radius. The key points are summarised as follows:

1. Candidate clusters from the WHL12 and redMaPPer cluster catalogues were split into training, validation and test sets. The MWAR dataset was designated as the training/validation set (80:20 per cent split ratio), whilst the RNMW/WNMR datasets were used as test sets. Photometric measurements of observed galaxies in the clusters was obtained from the SDSS-III DR9 photometric catalogue and full-sky dust reddening maps (Schlegel et al. 1998; Schlafly & Finkbeiner 2011) were also used to account for galactic extinction.
2. All the filters and colours described in §2.1 are assigned as input features to a single KNN algorithm for feature selection and filter magnitude-cut analysis. If a filter was used as part of an input feature, then the corresponding filter magnitude-cut was applied to exclude galaxies that had poor photometric measurements in that filter. The mean and standard deviation

¹⁰ It is known that distance comparisons in Euclidean space can become less effective with increasing dimensionality as the distance ratios become more uniform (Aggarwal et al. 2002). This means that other distance metrics such as cosine, Chi-squared, Manhattan and Minkowski (Hu et al. 2016) could also be considered.

were also calculated for each feature in the MWAR training set to perform feature scaling¹¹. From which, all input datasets to our model will require feature scaling with the same mean and standard deviation values determined for the MWAR training set.

3. Thirty repetitions of ten-fold cross validation (Stone 1974) were computed with SFS for a individual KNN algorithm, where a single nearest neighbour was used¹². This process was important for multiple reasons. Firstly, to analyse the stability of the KNN algorithm from minor changes to the training set. Secondly, to examine the relative frequency of features selected by SFS. Thirdly, to evaluate how filter magnitude-cuts affect the accuracy of photometric redshift predictions. Lastly, to provide a basis for comparing an individual algorithm with an ensemble algorithm.
4. The optimal filter magnitude-cuts determined for a single KNN algorithm were utilised for the SRKNN algorithm via transfer learning (Torrey & Shavlik 2010). From which, the training data for the internal KNN models of the SRKNN algorithm were built with bootstrap resamples, where bootstrap with replacement of the MWAR training set was used. Any clusters that were not used for bootstrapping of an internal KNN model were instead used for feature selection training of that internal KNN model with SFS. This ensured that all available training data was utilised.
5. The hyper-parameter settings of the SRKNN algorithm were tuned via a grid search strategy (Bergstra & Bengio 2012) using hold-out validation (Reitermanova 2010) of the MWAR validation set. This also examined how each of the hyper-parameter settings affected the model performance and generalisation.
6. Evaluation of the tuned model performance was obtained with the WNMR/RNMW test sets, which were all unseen clusters. Uncertainties for the photometric redshift estimate of each cluster were approximated with empirical bootstrap confidence intervals. Additionally, the tuned model was run on clusters with low richness¹³ and clusters at high redshift¹⁴ to assess the response of the tuned model on clusters with unseen properties.

3 RESULTS

3.1 Feature Selection and Filter Magnitude-Cut Analysis

Following the procedure described in §2.3, we first examine the stability of photometric redshift predictions for a single KNN algorithm. As seen in Figure 4, we observe that for brighter filter magnitude-cuts the number of selected features by SFS are more contrast, such that the resultant feature subsets for fainter filter magnitude-cuts are more strongly influenced by the observations

¹¹ All photometric measurements of features are standardised with zero-mean centering and unit variance, which is necessary for the comparison of Euclidean distance measurements (Raschka 2014).

¹² A single nearest neighbour minimises algorithmic biases which in turn maximises the variance of predictions (Friedman et al. 2001).

¹³ We define a cluster with low richness as a cluster that has a richness of twenty or fewer member galaxies.

¹⁴ We define a cluster at high redshift as a cluster that has a photometric redshift equal or greater than 0.6, which is the upper limit of our training set.

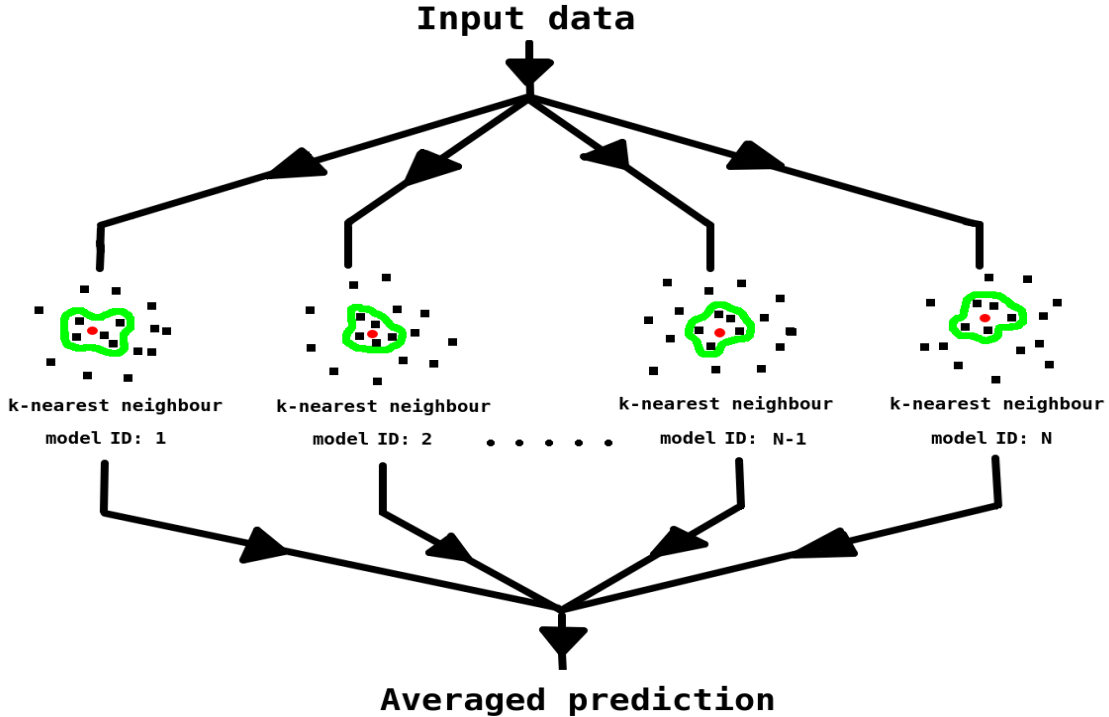


Figure 3. A schematic diagram of the SRKNN algorithm. The solid lines with black arrows indicates the flow of input data to an ‘N’ number of internal KNN models. In this example diagram, we use a red circle in each internal KNN model to represent an input test data point, black squares represent training data points and the green outline show the nearest neighbour training data points from the input test data point. From which, the median of training label values for the corresponding nearest neighbour training data points is used as a prediction for an internal KNN model, where the global model prediction is approximated with the median of predictions across all internal KNN models.

in the MWAR training set itself. However, as seen from the corresponding photometric redshift prediction errors, we find that this did not significantly alter the stability of predictions. We also compare the performance of SFS selected features with the control group features (see §§2.2.1), which had not been SFS selected. We repeat the same procedure used to analyse the SFS selected features for the control group features as well. From which, in Figure S1 (available online) we find that the control group features tend to have larger photometric redshift prediction errors in comparison to the SFS selected features for each search radius.

By repeatedly applying ten-fold cross validation to the MWAR training set we could also examine the relative frequency of features selected by SFS. This was done by calculating the relative frequency of features observed in the best performing feature subsets across all thirty repeats. As seen from Table 2, we find that some of the features are frequently selected whilst other features are rarely chosen, such that certain features are more likely to be picked by SFS if they are present in the input features.

Next, we determine the optimal filter magnitude-cut for each search radius by identifying filter magnitude-cut values that returned the lowest photometric redshift prediction error and retained at least 95 per cent of clusters. In Figures 4 and S1 (available online), we find that the LM filter magnitude-cut is the optimal filter magnitude-cut for the 10 and 21 arcseconds search radii whilst the LM-0.5 filter magnitude-cut is the optimal filter magnitude-cut for the 32 arcseconds search radius. We also compare whether applying filter magnitude-cuts improves the predictive performance of the model. In Figures 4 and S1 (available online) we find that a dataset, NC, with no filter magnitude-cuts applied to it, is not the

optimal filter magnitude-cut for any search radius whilst datasets with filter magnitude-cuts applied often had lower photometric redshift prediction errors.

We also assess how magnitude-cuts of the filters themselves affect the percentage of clusters retained in the MWAR training set, where the optimal filter magnitude-cut for each search radius was applied. From Figure 5 we find that all filters, except for the u filter, satisfied the 95 per cent cluster retention threshold at each search radius. In addition, we observe in Table 2 that the u filter did not appear in any final feature subset. From which, we decide that all input features which did not involve the u filter would be used as the new input features for the SRKNN algorithm to reduce the computational cost of evaluating redundant features during feature selection training. One would expect the u filter to be a poor predictor of redshift beyond very low redshift as it will probe further into the UV with increased redshift.

3.2 Hyper-Parameter Tuning Analysis Of The SRKNN Algorithm

We combine the optimal filter magnitude-cuts learned in §3.1 with a grid search strategy to fine-tune the SRKNN algorithm, which is known as inductive transfer learning (Vilalta et al. 2010; Segev & El-Yaniv 2016). We assume that the knowledge learned for the KNN algorithm is appropriate for the SRKNN algorithm, since the SRKNN algorithm is an extension of the KNN algorithm. From which, we ran the grid search on all combinations of hyper-parameter settings with a specified range of values to evaluate how each hyper-parameter setting affects model generalisation and pre-

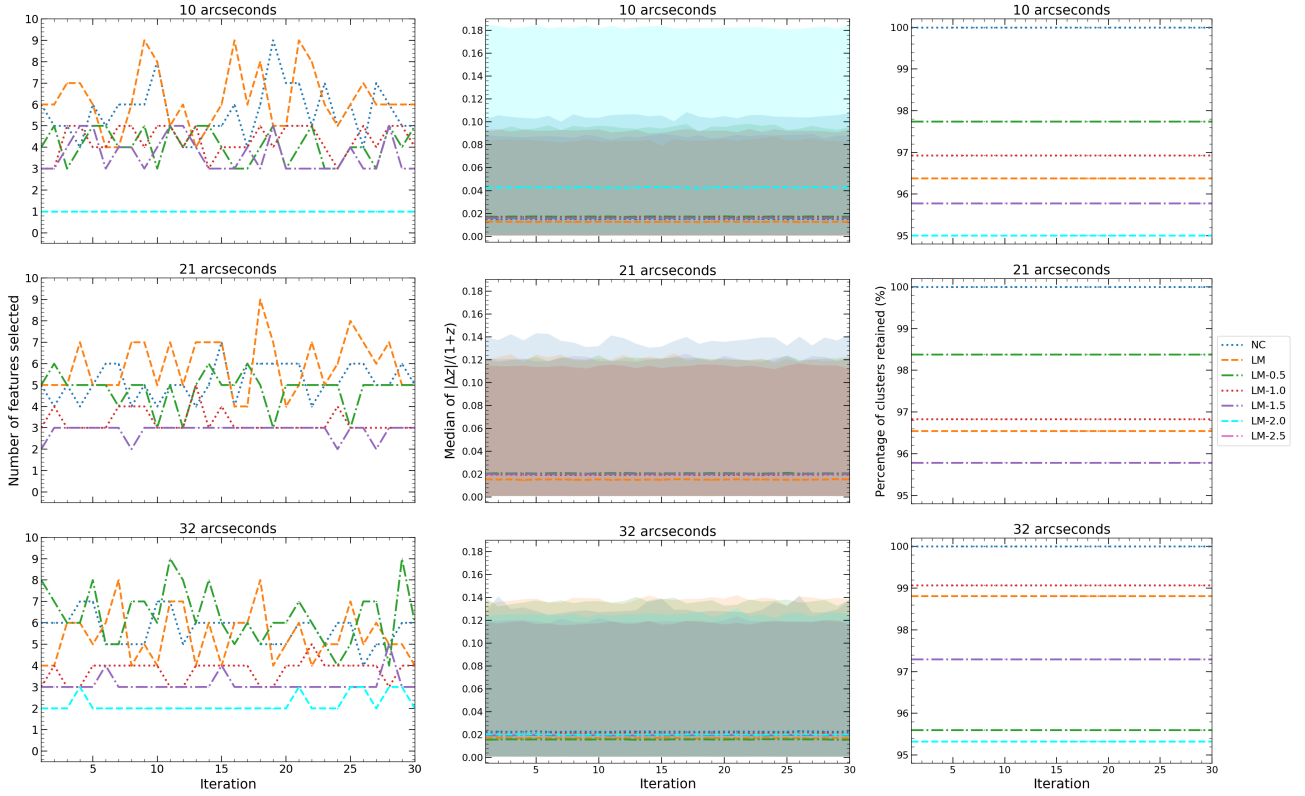


Figure 4. Plots displaying the results from applying filter magnitude-cuts to the MWAR training set using a single KNN algorithm with SFS selected features for each search radii (10 arcseconds on the top row, 21 arcseconds on the middle row and 32 arcseconds on the bottom row). ‘NC’ represents a dataset with no filter magnitude-cuts applied and ‘LM’ represents the MWAR dataset with SFS selected features where filter magnitude-cuts are applied to the limiting magnitude of SDSS. In addition, ‘LM’ is the faintest filter magnitude-cut whilst ‘LM-2.5’ is the brightest filter magnitude-cut. Left column: Number of features selected for the best performing feature subset in ten-fold cross validation across thirty repeats. Middle column: Median of photometric redshift prediction errors ($|\Delta z|/(1+z)$) across all tested clusters for the best performing feature subset in ten-fold cross validation across thirty repeats, where the shaded regions represent 95 per cent confidence intervals. Right column: Percentage of test clusters retained after filter magnitude-cuts are applied with the best performing feature subset in ten-fold cross validation across thirty repeats. It should also be noted that if the percentage of clusters retained, after filter magnitude-cuts are applied, do not satisfy the 95 per cent cluster retainment threshold we would not display the corresponding results in the other columns.

Search Radius [arcseconds]	Optimal Filter Magnitude-Cut [mag]	SFS Selected Features	Relative Frequency Of SFS Selected Features (per cent)
10	LM	$r-i, g-z, r-z, g, g-i, z, r, i-z, g-r, i$	100, 100, 90, 83, 67, 53, 47, 40, 30, 13
21	LM	$z, r-i, g-i, g-z, r, g, g-r, i, r-z$	87, 80, 80, 70, 63, 60, 60, 47, 47
32	LM-0.5	$g-z, r-i, g-i, g-r, g, i-z, z, r, i, r-z$	93, 83, 83, 77, 70, 60, 50, 47, 43, 27

Table 2. A table displaying the relative frequency of features selected by SFS across thirty repeats of ten-fold cross validation on the MWAR training set with a single KNN algorithm at the optimal filter magnitude-cut for each search radius. The selected features are listed in the same order as the corresponding relative frequency. It can be seen that the z filter, rather than a colour, has the highest relative frequency amongst the features at the 21 arcseconds search radius for a single KNN algorithm but the relative frequency diminishes when the z filter is instead used in an ensemble (see §3.2).

dictive performance. The following hyper-parameter setting values are used in the grid search:

- The number of internal KNN models - 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000.
- The number of initialised random features - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.
- The number of nearest neighbours - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25.

We utilise validation curves (VanderPlas 2016) to analyse the response from different hyper-parameter setting combinations of

the SRKNN algorithm. This involves fixing each hyper-parameter setting as a constant with respect to the other hyper-parameter settings to compute the median of photometric redshift prediction errors across all tested clusters with that fixed hyper-parameter setting. We focus on minimising the photometric redshift prediction error on the MWAR validation set rather than the MWAR training set. Since the MWAR training set had already been seen by the model, the results from the MWAR training set would be biased whilst the MWAR validation set remains unseen by the model. However, running the model on both the MWAR training and validation sets is still beneficial to check the generalisation of the

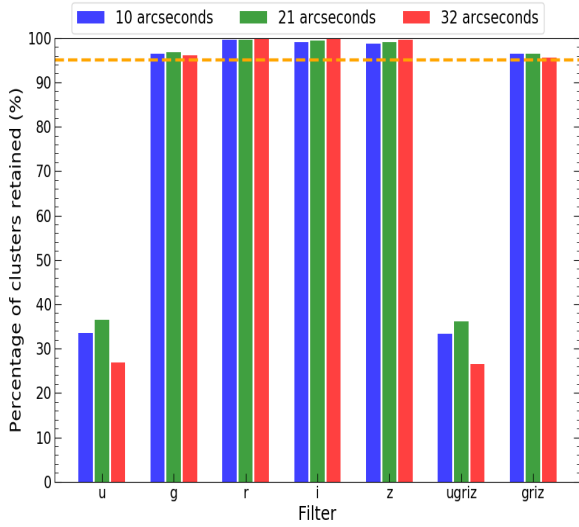


Figure 5. Percentage of clusters retained in the MWAR training set after applying the optimal filter magnitude-cuts for each search radius to the *u*, *g*, *r*, *i*, *z*, *ugriz* and *griz* filters. The orange dashed line highlights the 95 per cent cluster retainment threshold.

hyper-parameter settings, as the model can overfit and underfit when applied on its own training data.

Firstly, we evaluate the model performance based on the number of nearest neighbours for each search radius. In Figure 6, we find that for a small number of nearest neighbours the model has high predictive variance as we observe a large difference between the training and validation errors. Although, we notice that the overall photometric redshift prediction error decreases as the number of nearest neighbours increases for the MWAR validation set, whereas the overall photometric redshift prediction error increases as the number of nearest neighbours increases for the MWAR training set. It can be seen that the number of nearest neighbours is a very important hyper-parameter setting to tune since the model performance varies a lot depending on the value used. From which, we determine the optimal values for the number of nearest neighbours of each search radius to be 19 for 10 arcseconds, 19 for 21 arcseconds and 16 for 32 arcseconds. It should be noted that the number of nearest neighbours value with the lowest photometric redshift prediction error was actually 25 for each search radius. We purposely avoid selecting this value since the number of nearest neighbours value has a large impact on the model performance, such that selecting the hyper-parameter value with the lowest photometric prediction error could likely overfit the model on the MWAR validation set itself. Instead, we prefer to choose more conservative values for the optimal number of nearest neighbours to balance model generalisation and performance.

Secondly, we examine the model performance based on the number of initialised random features for each search radius. In Figure 7, we find that for both the MWAR training and validation sets, the change in the photometric redshift prediction errors quickly decreases for a small number of initialised random features but then slowly decreases when a medium to large number of initialised random features was used. From which, we observe that the overall redshift prediction error decreases as the number of initialised random features increases. This implies that the number of initialised random features is also an important hyper-parameter setting to tune, since the model performance on the MWAR train-

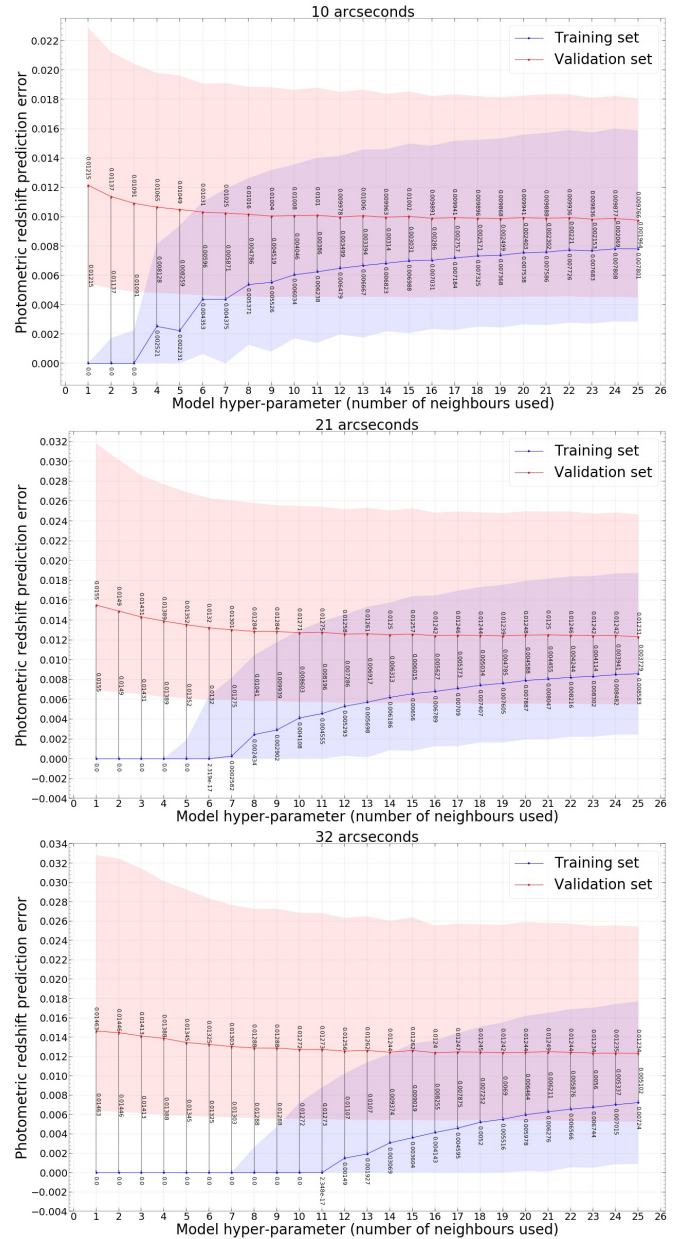


Figure 6. Validation curves from tuning the number of nearest neighbours hyper-parameter setting, where the photometric redshift prediction errors of the MWAR training (blue) and validation (red) sets are shown for each search radii (10 arcseconds on the top row, 21 arcseconds in the middle row and 32 arcseconds on the bottom row). The individual points display the median of photometric redshift prediction errors across all tested clusters and the shaded regions represent the 25th and 75th percentiles of the photometric redshift prediction errors for a fixed number of nearest neighbours with respect to the other hyper-parameter settings of the SRKNN algorithm. We also label the difference between the individual points of the training and validation errors.

ing and validation sets is somewhat reliant on the value selected. We determine the optimal values for the number of initialised random features of each search radius to be 9 for 10 arcseconds, 8 for 21 arcseconds and 7 for 32 arcseconds. Although, it can be seen that having no initialised random features (using all features for the input features) at times had lower photometric redshift prediction errors. However, this could also worsen model generalisation

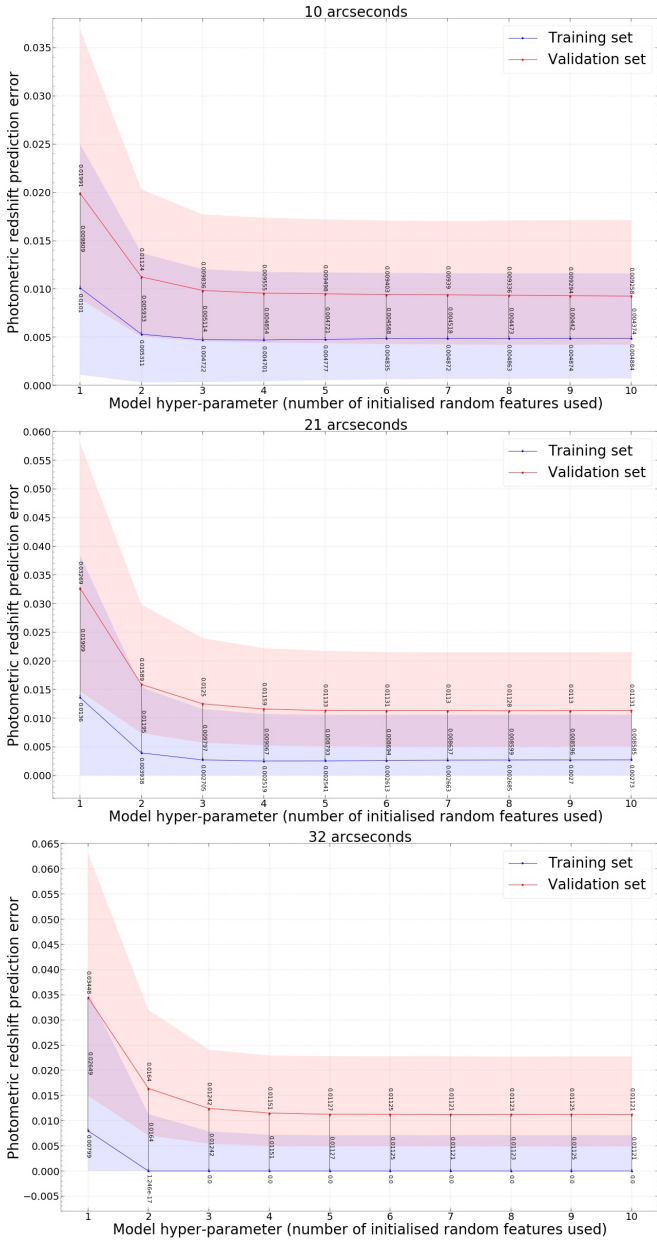


Figure 7. This figure is equivalent to Figure 6 except we tune the number of initialised random features hyper-parameter setting.

since strongly correlated features would not be restricted during SFS. Therefore, we again decide to select more conservative values for the optimal number of initialised random features.

Thirdly, we assess the model performance and behaviour based on the number of bootstrap resamples used for each search radius. Figure 8 shows that for the MWAR training and validation sets the change in the photometric redshift prediction error steeply decreases when a very small number of bootstrap resamples used but then remains flat as the number of bootstrap resamples increases. This tells us that the number of bootstrap resamples used is not a particularly important hyper-parameter setting to tune as the impact on the model performance for the MWAR training and validation sets is minimal. [Efron & Tibshirani \(1994\)](#) suggests that using fifty to two hundred bootstrap resamples is sufficient to calculate standard errors whereas bootstrap confidence interval esti-

mates require at least one order of magnitude higher computational cost. From which, we decide that using one thousand bootstrap resamples for each search radius would be enough to benefit from bootstrap confidence intervals. We also consider that since SFS would have selected different features for each bootstrap sample, we would not expect all internal KNN models to return predictions after filter magnitude-cuts are applied. Figure 9 displays the percentage of clusters returned with full, partial and no bootstrap resamples returned for estimating photometric redshift at each search radius. We find that employing a large number of bootstrap resamples reduces the percentage of clusters returned with no bootstrap resamples. Whilst for clusters with a full set of bootstrap resamples returned, the percentage of clusters returned initially drops but then remains flat as the number of bootstrap resamples increases. Whereas for clusters with partial bootstrap resamples returned, the percentage of clusters returned gradually increases as the number of bootstrap resamples increases. For this work, we prefer to minimise the percentage of clusters returned with no bootstrap resamples, since we want as many clusters as possible to have photometric redshift estimates. In Figure 10 we calculate the relative frequency of features selected by SFS with respect to the number of bootstrap resamples used at each search radius. It can be seen that as the number of bootstrap resamples increases, the spread of the relative frequency amongst the features decreases. From which, we also observe that the features with the highest relative frequency appear to be colours whilst features with the lowest relative frequency are filters. The model has learned that colours are more significant than filters for estimating photometric redshifts of clusters.

3.3 Model Performance Analysis With Test Sets

We use the WNMW/RNMW test sets to assess the performance of the SRKNN algorithm with the optimal hyper-parameters learned in §3.2 for each search radius. As described earlier in §2.1, the test sets contain clusters from the WHL12 and redMaPPer cluster catalogues with no corresponding cross-match. A summarised version of the test results can be found in Table 3.

In Figures 11, 12, 13, 14, 15 and 16 we compare the known photometric redshifts with the predicted photometric redshifts for clusters in the WNMW/RNMW test sets that had full bootstrap resamples returned by the tuned model. We find that as the search radius increases the median of photometric redshift prediction errors across all tested clusters in both test sets increases as well possibly due to line-of-sight interloping galaxies. From which, in Figures SA4, SA5, SA6, SA7, SA8 and SA9 (available online) we also examine the spatial distribution of several clusters with relatively large photometric redshift prediction errors. We repeatedly observe that if line-of-sight interloping galaxies are present within the search radii of clusters, the resultant model predictions have relatively large photometric redshift prediction errors. Moreover in Figures 11, 12, 13, 14, 15 and 16 it can be seen that the width of the 95 per cent confidence intervals around predictions decreases as the search radius increases, as shown by wider intervals. This means there is lower precision of the predicted photometric redshift value. Despite this, we find that the tuned model seems to perform well at all redshifts since the majority of cases have relatively low photometric redshift prediction errors for each search radius. Although, we notice that an increasing number of cases have relatively large photometric redshift prediction errors near to the redshift training boundaries of the MWAR training set as the search radius increases. Furthermore, we also examine the performance of the tuned model on clusters in the WNMW/RNMW test sets with only partial boot-

Test Set	Search Radius [arcseconds]	Optimal Filter Magnitude-Cut [mag]	# Clusters (total)	# Clusters (radius)	# Clusters (tested)	\widetilde{E}_z
WNMR	10	LM	9723	8844	8442	0.0106
WNMR	21	LM	9723	9564	9057	0.013
WNMR	32	LM-0.5	9723	9691	9057	0.014
RNMW	10	LM	8646	8131	7319	0.0123
RNMW	21	LM	8646	8577	7870	0.0156
RNMW	32	LM-0.5	8646	8635	7416	0.0181

Table 3. A table displaying the median of photometric redshift prediction errors (\widetilde{E}_z , where $E_z = |\Delta z|/(1+z)$) across all tested clusters for each test set, search radius and optimal filter magnitude-cut. We also show the total number of clusters in the original full dataset (total), the number of clusters that have galaxies within the specified search radius (radius) and the number of clusters that have galaxies within the specified search radius after filter magnitude-cuts (tested). The values in this table summarise the test results in Figures 11, 12, 13, 14, 15 and 16.

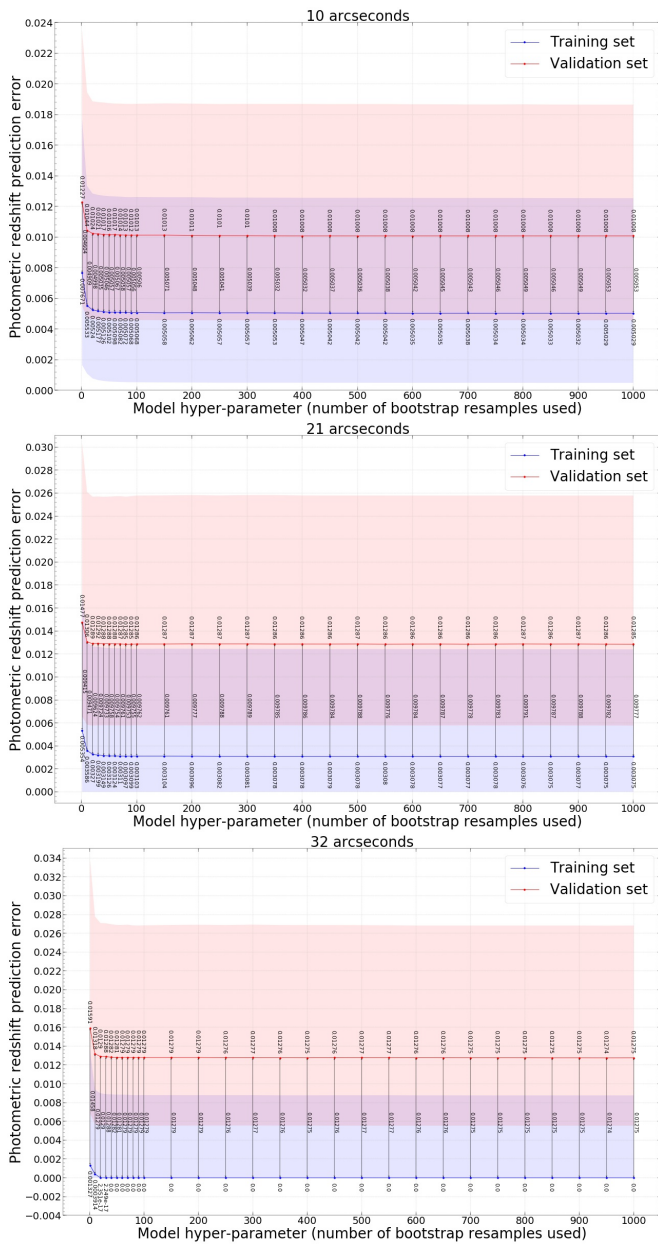


Figure 8. This figure is equivalent to Figure 6 except we tune the number of bootstrap resamples hyper-parameter setting.

strap resamples returned for each search radius. From Figures S2, S3, S4, S5, S6 and S7 (available online) we find that in almost all cases the photometric redshift prediction error is poorly constrained when partial bootstrap resamples are used.

In Figures 17 and 18 we determine the number of galaxies used in photometric redshift predictions of clusters from the WNMR/RNMW test sets that had full bootstrap resamples returned by the tuned model for each search radius. This examines how the tuned model performs with respect to different numbers of galaxies. It can be seen that as the search radius increases the number of galaxies used in photometric redshift predictions increases too. From which, we find that the median of photometric redshift prediction errors across all tested clusters is similar regardless of the number of galaxies used by the tuned model. Although, we notice that clusters with larger numbers of galaxies used for photometric redshift predictions are frequently seen between low and intermediate redshifts with relatively low photometric redshift prediction errors. Whereas clusters at considerably lower and higher redshifts rarely have large numbers of galaxies used for photometric redshift predictions and also have relatively large photometric redshift prediction errors.

In Figures 19 and 20 we examine the redshift distribution of clusters from the WNMR/RNMW test sets with no bootstrap resamples returned by the tuned model for each search radius. We observe that the redshift distributions are predominantly skewed towards higher redshifts. This could be due to the galaxies in clusters at higher redshifts having poorer photometric measurements in comparison to the galaxies in clusters at lower redshifts. Although, it should be noted that the redshift distribution for the RNMW dataset itself is also heavily skewed towards higher redshifts.

3.4 Further Model Testing

We also test the tuned model on additional clusters that reside in unseen parameter space, such as clusters with low richness and clusters at redshift equal or greater than 0.6. This was to analyse the generalisation of the tuned model, by running it on clusters with properties that it had not been trained for, which are also likely to be encountered in surveys. We apply the same analysis procedure as performed in §3.3 and provide the full results in the online supplementary material. For this section we will only describe the response of the tuned model with respect to different cluster properties.

In Figures S8, S9 and S10 (available online) we ran the tuned model on clusters with low richness, which have a richness of twenty or fewer member galaxies such that they did not qualify

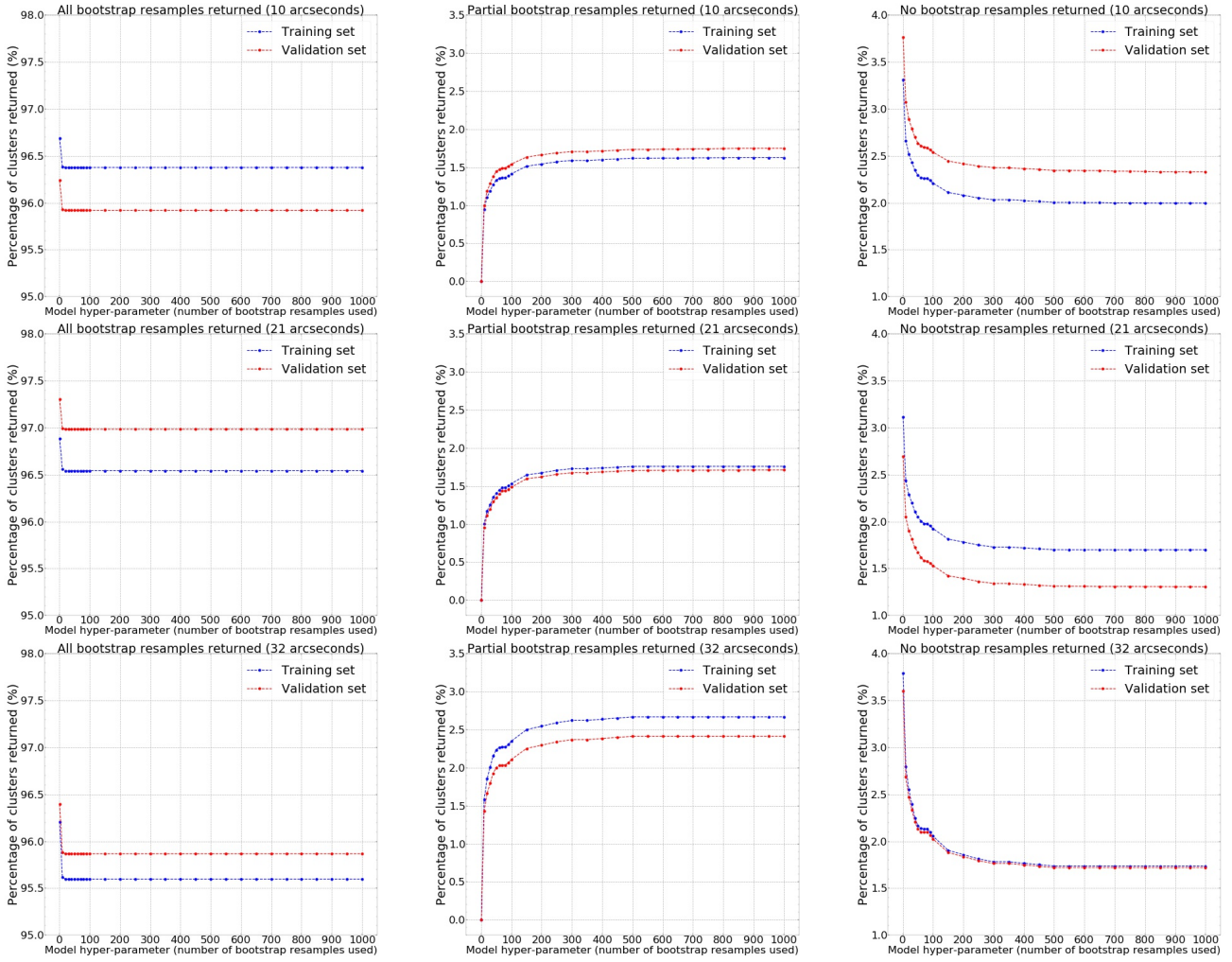


Figure 9. Validation curves from tuning the number of bootstrap resamples hyper-parameter setting, where the percentage of clusters returned with full, partial and no bootstrap resamples are from the MWAR training (blue) and validation (red) sets at each search radii (10 arcseconds on the top row, 21 arcseconds in the middle row and 32 arcseconds on the bottom row). The individual points display the percentage of clusters returned across a fixed number of bootstrap resamples with respect to the other hyper-parameter settings of the SRKNN algorithm.

for the MWAR dataset, to obtain photometric redshift predictions that have full bootstrap resamples returned at each search radius. We find that the number of cases with relatively large photometric redshift prediction errors increases as the search radius increases, particularly at higher redshifts. However, we also notice that the median of photometric redshift prediction errors for each search radius remains relatively low when compared to the median of photometric redshift prediction errors for the WNMR/RNMW test sets. Moreover, we observe that the precision of the 95 per cent confidence intervals becomes worse towards the redshift training boundaries when the search radius increases.

In Figures S16, S17 and S18 (available online) we ran the tuned model on clusters at high redshift, which have a redshift beyond the redshift training boundaries such that they did not qualify for the WNMR dataset, to obtain photometric redshift predictions that have full bootstrap resamples returned at each search radius. We immediately notice that the overall accuracy of photometric redshift predictions is low when compared to the other test sets, as the tuned model constantly underestimates the photometric redshifts regardless of the search radius used. We also observe that the precision of the 95 per cent confidence intervals around predictions

is poorly constrained, such that it would be difficult to distinguish clusters at high redshift from poorly constrained clusters at intermediate redshift.

In Figures S24, S25 and S26 (available online) we ran the tuned model on clusters at high redshift with low richness, which have a richness of twenty or fewer member galaxies and a redshift beyond the redshift training boundaries such that they did not qualify for the WNMR dataset, to obtain photometric redshift predictions that have full bootstrap resamples returned at each search radius. Similar to the results in Figures S16, S17 and S18 for clusters at high redshift, we find that the overall accuracy of photometric redshift predictions is also low, as the tuned model constantly underestimates the photometric redshifts. In addition, the 95 per cent confidence intervals around predictions are also poorly constrained regardless of the search radius used.

In Figures S32, S33 and S34 (available online) we ran the tuned model on clusters with low richness, which have a richness of twenty or fewer member galaxies such that they did not qualify for the WNMR dataset, to obtain photometric redshift predictions that have full bootstrap resamples returned at each search radius. Similar to the results in Figures S8, S9 and S10 for clusters with

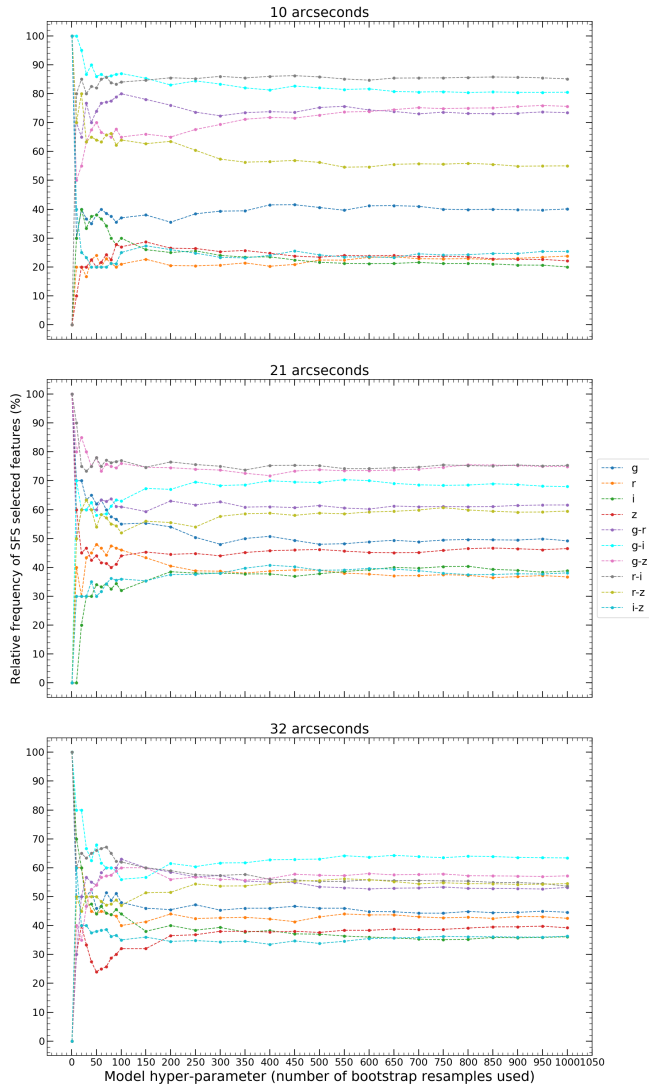


Figure 10. Validation curves from tuning the number of bootstrap resamples hyper-parameter setting, where the relative frequency of features selected by SFS with the MWAR training set is shown for each search radii (10 arcseconds on the top row, 21 arcseconds in the middle row and 32 arcseconds on the bottom row). The individual points display the relative frequency of features selected by SFS across a fixed number of bootstrap resamples with respect to the other hyper-parameter settings of the SRKNN algorithm.

low richness, we find that the overall accuracy of the photometric redshift predictions is high, as only a minority of cases have relatively large photometric redshift prediction errors. Although, we also observe that the precision of the 95 per cent confidence intervals becomes worse towards the redshift training boundaries when the search radius increases.

Lastly, we also evaluate the effectiveness from increasing the search radius on the performance of photometric redshift predictions and the number of clusters with full bootstrap resamples returned. For example, if a cluster did not have a photometric redshift estimate with full bootstrap resamples returned within a 10 arcseconds search radius, we would try using a 21 arcseconds search radius instead. From which, if a 21 arcseconds search radius was not sufficient, we would then try using a 32 arcseconds search instead. In Figures S40, S43, S46, S49, S52 and S55 we find that

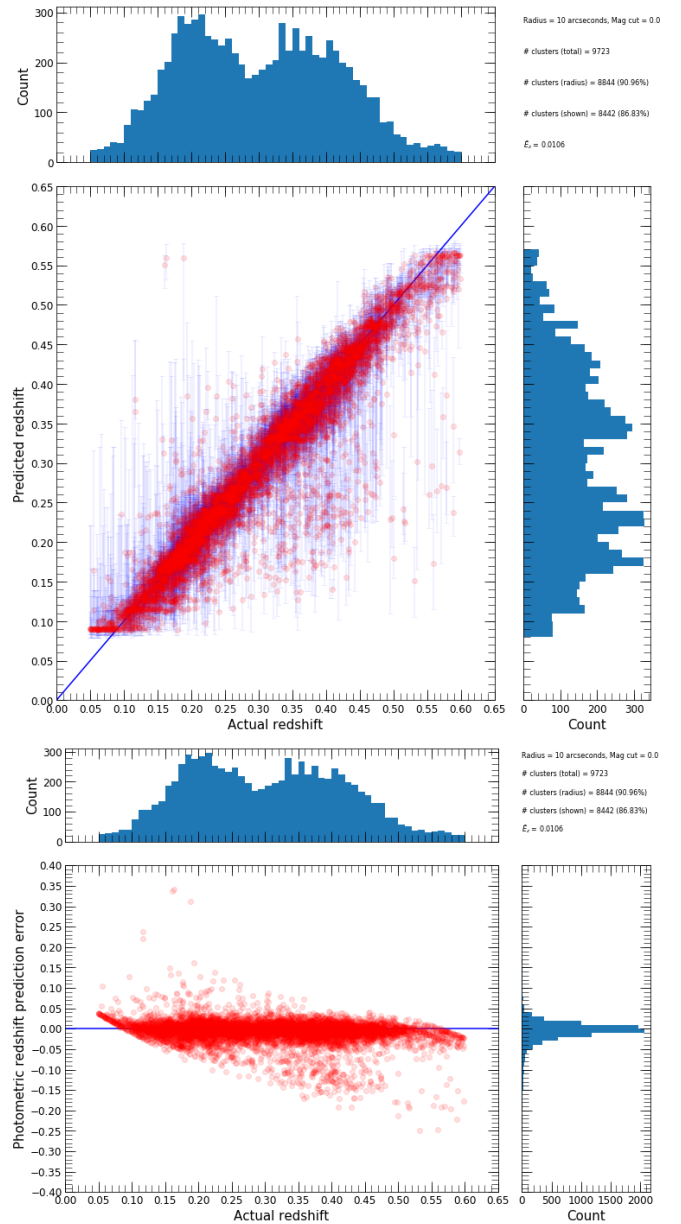


Figure 11. Plots displaying the performance of photometric redshift predictions of clusters for the WNMR test set that had full bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus ‘actual’ photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus ‘actual’ redshift of tested clusters with frequency histograms of the distributions. Other: ‘# clusters (total)’ represents the total number of clusters in the WNMR dataset, ‘# clusters (radius)’ represents the number of clusters in the WNMR test set that have observed galaxies within a 10 arcseconds search radius, ‘# clusters (shown)’ represents the number of clusters in the WNMR test set that have observed galaxies within a 10 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

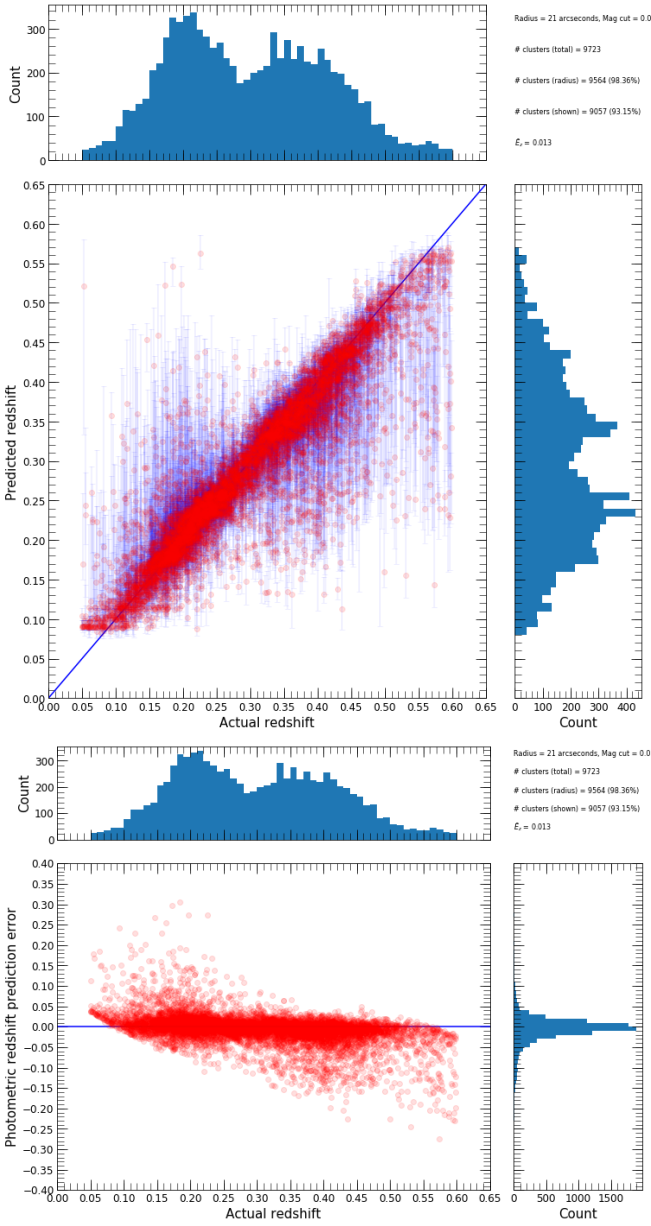


Figure 12. This figure is equivalent to Figure 11 except we examine the performance of photometric redshift predictions of clusters within a 21-arcseconds search radius.

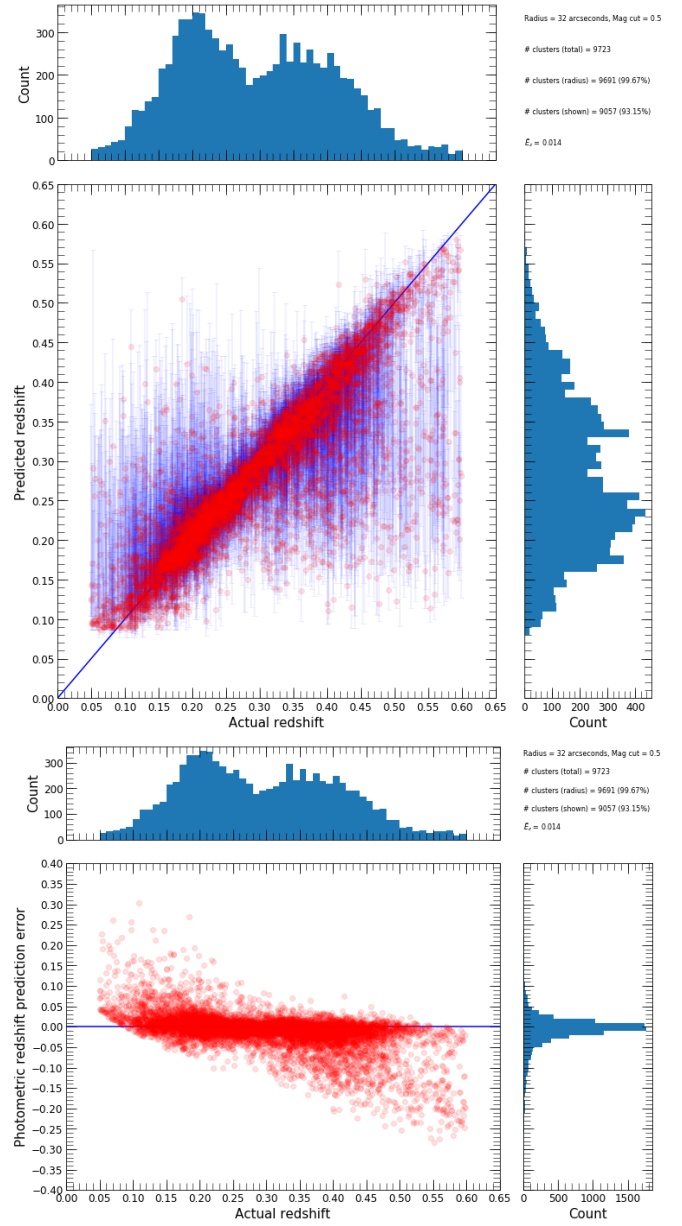


Figure 13. This figure is equivalent to Figure 11 except we examine the performance of photometric redshift predictions of clusters within a 32-arcseconds search radius.

as the search radius increases the overall accuracy of photometric redshift estimates decreases. Although, this can still be beneficial rather than having clusters with no photometric redshift estimates at all. We also observe that as the search radius increases the number of photometric redshift estimates with full bootstrap resamples returned decreases as well. These trends can be seen repeating for all of the test sets.

4 DISCUSSION

4.1 Effectiveness Of Z-Sequence For Photometric Redshift Predictions

In §3.3 we employ samples from the WHL12 and redMaPPer cluster catalogues to examine the performance of the tuned model. From Figures 11, 12 and 13 it can be seen that majority of clusters in the WNMNR test set are observed at low to intermediate redshifts, whereas from Figures 14, 15 and 16 it can be seen that majority of clusters in the RNMW test set are observed at intermediate redshift. This tells us that the methods used to estimate photometric redshifts in WHL12 and redMaPPer can significantly influence the resultant redshift distributions. Although, we find that the tuned model does not have much difficulty in working with either of these redshift distributions, as the overall performance of photometric redshift

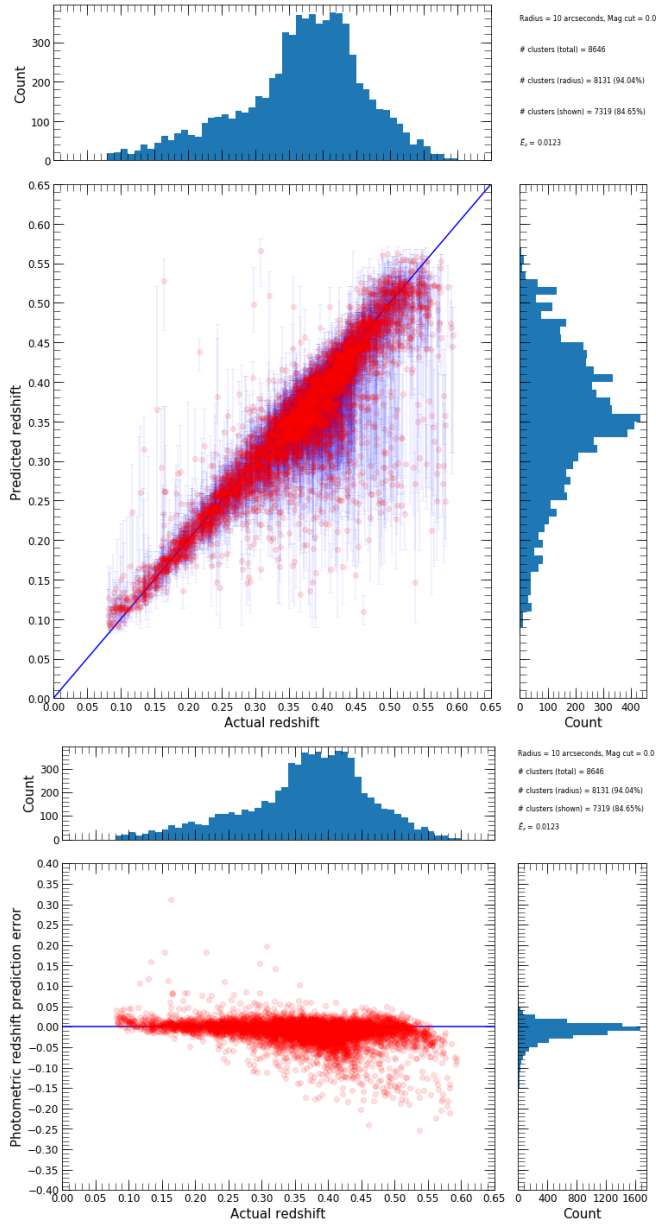


Figure 14. This figure is equivalent to Figure 11 except we examine the performance of photometric redshift predictions of clusters within a 10 arcseconds search radius for the RNMW test set.

prediction errors for both test sets are similar. From which, we can infer that Z-Sequence can be effectively utilised across a wide range of redshifts if the appropriate training data is available.

For this paper, we assign the photometric redshifts of the WHL12 and redMaPPer cluster catalogues as ‘actual’ redshifts to examine the model performance on a large sample of clusters. Since we aim to minimise data wastage, it is important to try to utilise all available clusters even though not all clusters will have spectroscopic redshifts. We are aware that the ‘actual’ photometric redshifts for clusters in WHL12 and redMaPPer have a scatter of ~ 0.01 from spectroscopic redshifts. This is similar to the scatter in our photometric redshift prediction errors of ~ 0.01 from the ‘actual’ photometric redshifts, which suggests that our model is as accurate as it can be based on the data used for training and testing. We expect that our photometric redshift prediction error would

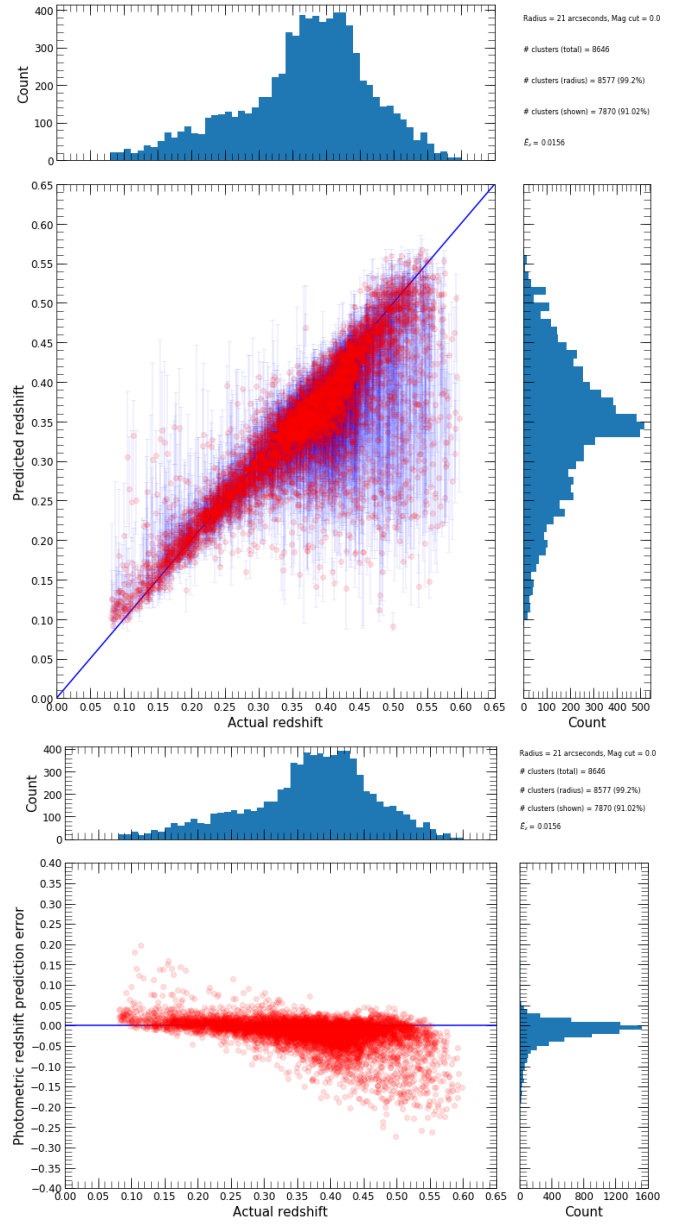


Figure 15. This figure is equivalent to Figure 11 except we examine the performance of photometric redshift predictions of clusters within a 21 arcseconds search radius for the RNMW test set.

decrease if we trained on a large, entirely spectroscopic sample instead as the scatter associated with the photometric redshifts in the WHL12 and redMaPPer catalogues will be removed. In addition, it should be noted that the flaring seen in Figures 14, 15 and 16 lowers the predicted redshift values between “actual” redshifts of $0.35 \geq z \geq 0.45$ for the RNMW test set. This is due to the flaring originating from redMaPPer itself and not from our algorithm, as it also occurs in Figure 7 of Rykoff et al. (2014).

In §3.4 we test the tuned model on clusters with unseen properties. We find that the tuned model performs well on clusters in similar parameter space to the MWAR training set and it also performs well on clusters of all richnesses within the redshift training boundaries. However, the tuned model performs poorly on clusters beyond the redshift training boundaries. This tells us that the performance of the tuned model is more dependent on the redshift of

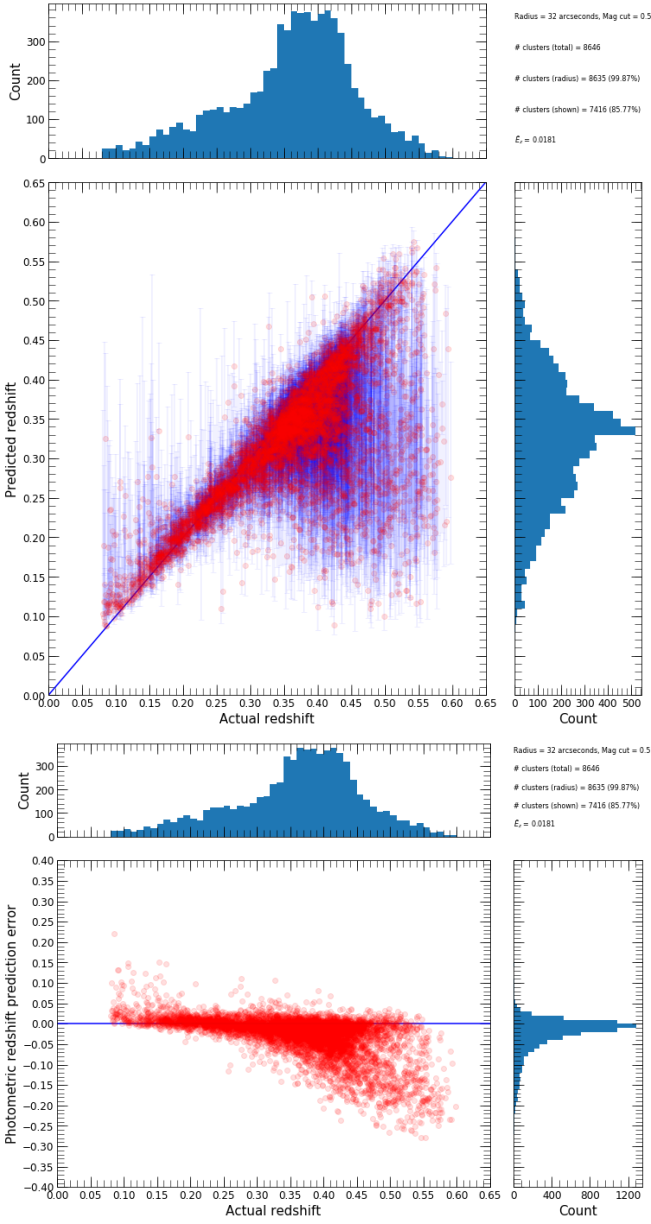


Figure 16. This figure is equivalent to Figure 11 except we examine the performance of photometric redshift predictions of clusters within a 32 arcseconds search radius for the RNMW test set.

the cluster than the richness of the cluster. The tuned model is only effective on clusters at the redshift range it was trained for since we are limited to the redshift range of the majority of clusters available in SDSS. In addition, we observe an apparent feature seen at the lower and upper boundaries for predicted photometric redshifts in Figures 11, 12 and 13. We believe the cause of the apparent feature is due the nature of the machine learning algorithm itself. This is because the k-nearest neighbours algorithm calculates its prediction from the labels of the nearest neighbour examples in the training set when given an input data point, where the photometric redshift limits of the MWAR dataset is $0.0698 \leq z \leq 0.5986$ whilst the WNMR dataset is $0.05 \leq z \leq 0.599$. This means that all photometric redshift predictions are bounded within the photometric redshift training range, such that clusters with ‘actual’ redshifts outside the boundaries could end up as part of the apparent feature. This ex-

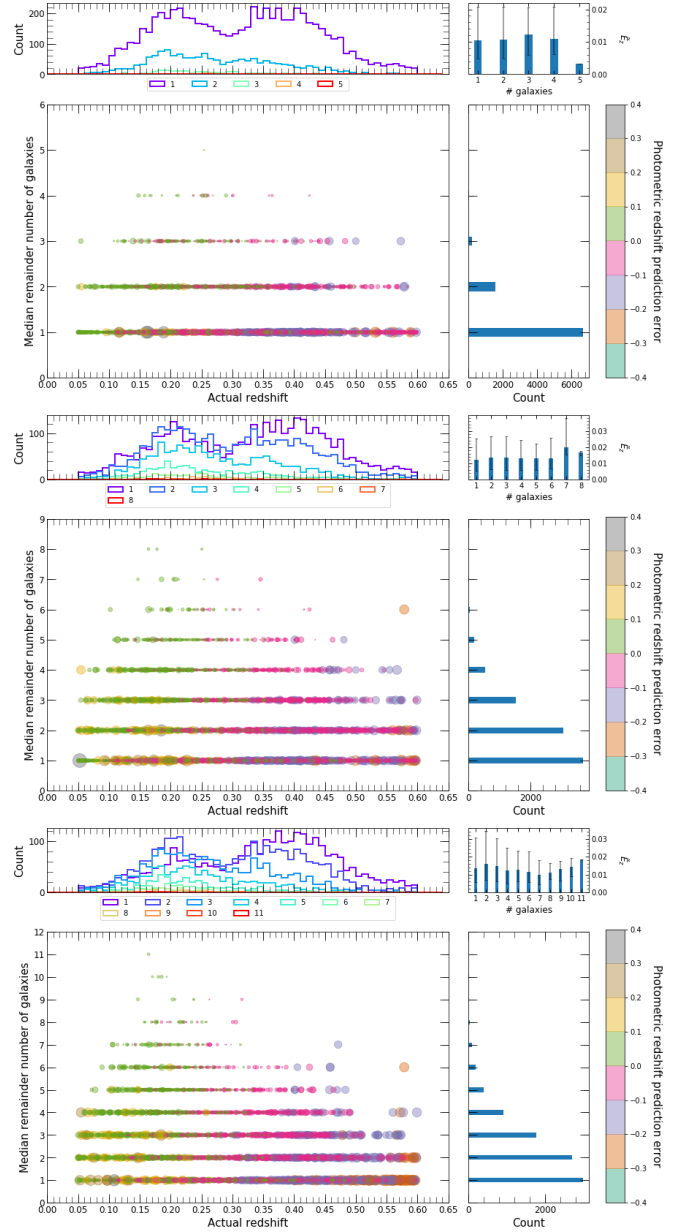


Figure 17. Plots displaying the number of galaxies used in photometric redshift predictions versus ‘actual’ redshift of tested clusters for the WNMR test set, where predictions had full bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

plains why we do not observe the apparent feature in Figures 14, 15 and 16 as the photometric redshift limits of the RNMW dataset is $0.0811 \leq z \leq 0.5983$. As a further demonstration of the success of our algorithm, we note that the WNMR and RNMW test sets consist of clusters found in one catalogue and not the other. This could mean that these clusters are more difficult to detect and therefore potentially harder to assign a redshift value via other photometric redshift prediction methods, whereas our algorithm can estimate redshifts for the majority of these clusters. It should also be noted

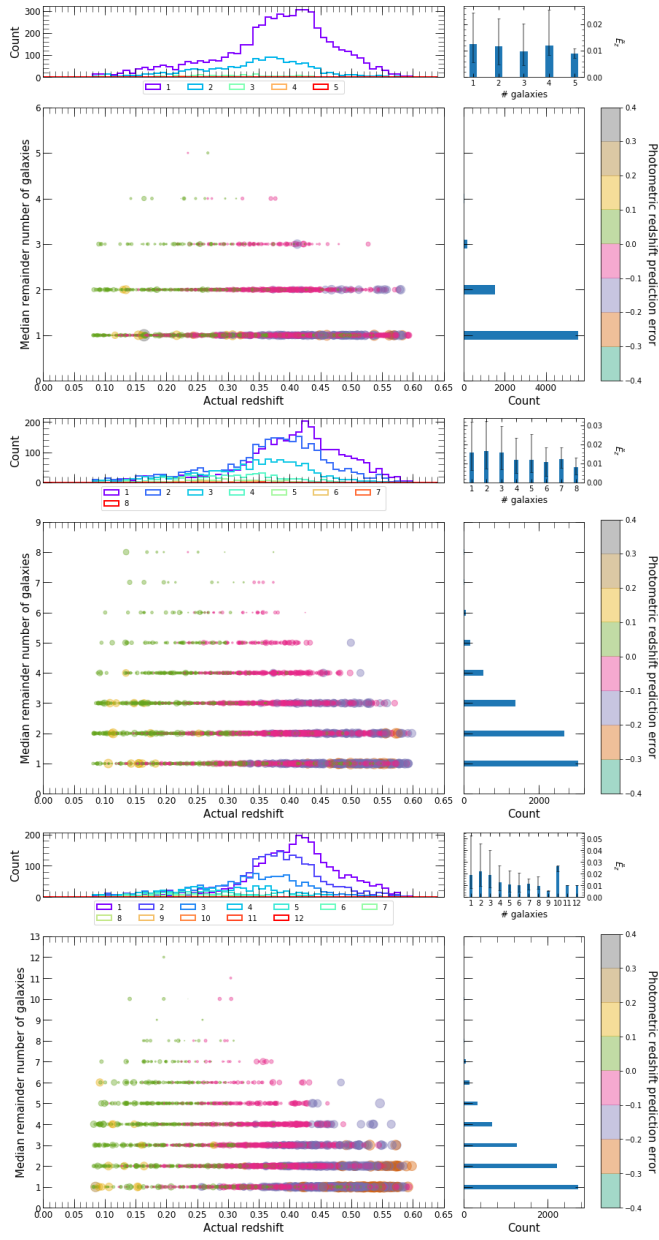


Figure 18. This figure is equivalent to Figure 17 except we examine the number of galaxies used in photometric redshift predictions for the RNMW test set.

that the observed magnitude errors for all SDSS filters increases with redshift, as seen in Figure SA10 (available online). This means it would be difficult for any empirical algorithm to make accurate photometric estimates in the high redshift regime. However, we expect our model would be successful at estimating photometric redshifts for high redshift clusters if trained on imaging surveys such as LSST or Euclid, which will have greater photometric depths to increase the redshift limits of cluster detection when compared with SDSS.

We notice in Table 3 that the median value of $|\Delta z|/(1+z)$ increases for the WNMR and RNMW test sets by 32 per cent and 47 per cent respectively when the search radius is enlarged from 10 arcseconds to 32 arcseconds. This can also be seen in §3.4 where the number of cases with accurate photometric redshift es-

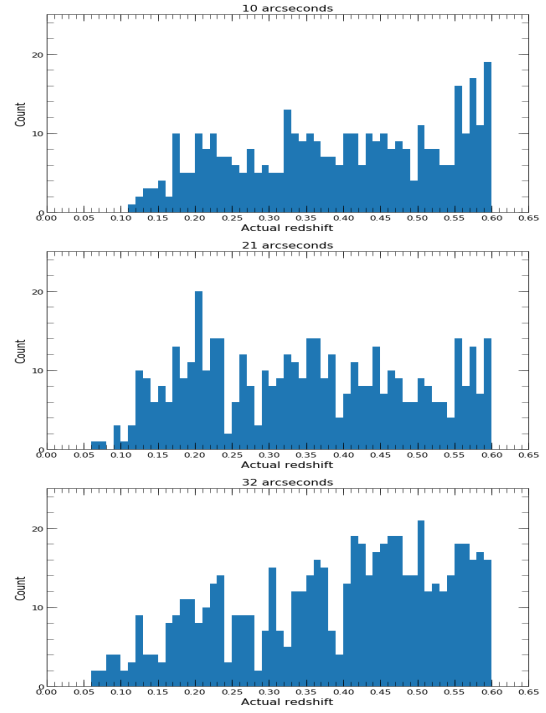


Figure 19. Frequency histograms displaying the ‘actual’ redshift distributions of clusters from the WNMR test set that had no bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds

search radius.

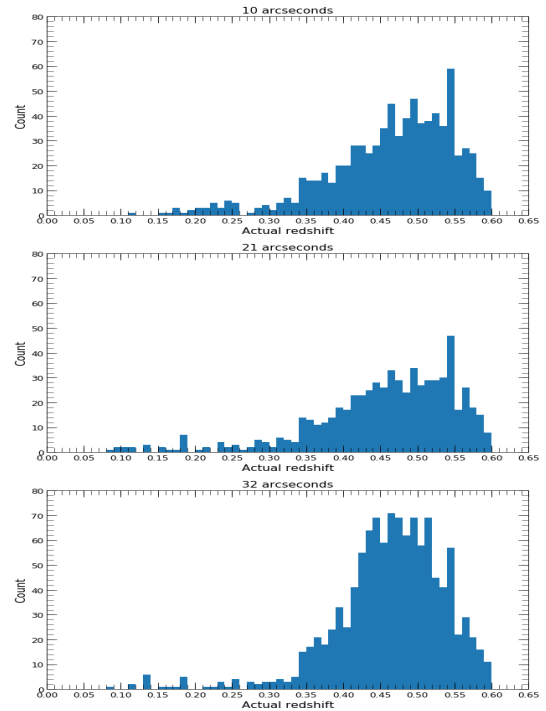


Figure 20. This figure is equivalent to Figure 19 except we examine the ‘actual’ redshift distributions of clusters that had no bootstrap resamples returned for the RNMW test set.

imates decreases as the search radius increases, as a larger search radius is more likely to include interlopers. From Figures SA6, SA7 and SA9 (available online), we find that interlopers are evident in contaminating estimates with relatively large photometric redshift prediction errors if they appear in the test set. Whilst Figures SA4, SA5 and SA8 (available online) indicate that interlopers are also somewhat present within the training set itself, as we find that some model predictions for clusters with no obvious interlopers in the test set still have relatively large photometric redshift prediction errors. Subsequently, we aim to further improve the accuracy of the Z-Sequence model in future work by developing new strategies to constrain interlopers, such as with unsupervised machine learning techniques that identifies the presence of line-of-sight interloping galaxies and multiple projected line-of-sight clusters. This new method can be employed as an additional pre-processing tool to accompany the Z-Sequence model. From which, we could increase the size of the search radius once the obvious interlopers are removed and examine whether the photometric redshift prediction accuracy significantly improves if more cluster members are included. In addition, Figures SA6 and SA9 (available online) show that filter magnitude-cuts are also partially responsible for estimates with relatively large photometric redshift prediction errors, as we find that all of the galaxy members in some cluster cores are removed from model predictions due to poor photometry measurements. Furthermore, we notice in Figure SA7 (available online) that the 95 per cent confidence interval for the photometric redshift estimate involving the interloper becomes considerably wider in comparison to the photometric redshift estimates without the interloper. This shows that the bootstrap confidence intervals could indicate whether interlopers are involved in the model prediction. Although, it should be noted that Figures 17 and 18 show that the majority of the model predictions seem to employ relatively few galaxies for each search radii, such that it would be difficult to constrain interlopers in most instances. Moreover, by comparing the number of clusters that have photometric redshift estimates with full bootstrap resamples returned exclusively within each of the 10, 21, 32 arcseconds search radii (see Figures S40, S43, S46, S49, S52 and S55 [available online]), we discover that the majority of cases are actually within the 10 arcseconds search radius whereas only a minority of cases require an increase in the search radius to 21 and 32 arcseconds. This suggests that if we were to retrain the model on different surveys, we could consider not needing to employ multiple large search radii as the computational cost for training the model could outweigh the benefits gained.

It is worth noting that our approach results in photometric redshift predictions with full, partial and no bootstrap resamples returned. This is primarily due to the use of filter magnitude-cuts in each internal KNN model, which excludes galaxies with poorer photometric measurements from the cluster before any predictions are made. Although, we observe in §§3.1 that applying filter magnitude-cuts can improve the overall accuracy of photometric redshift estimates. From which, we find that photometric redshift predictions with full bootstrap resamples returned are fairly accurate, as seen in §§3.3. However, it can also be seen that photometric redshift predictions with partial bootstrap resamples returned have low accuracy. This could be caused by the remaining bootstrap resamples not utilising strong predictive features. Subsequently, we advise that future photometric redshift estimates with partial bootstrap resamples returned should be flagged and used cautiously.

4.2 Practicality Of The Machine Learning Techniques Used In This Paper

For this paper, we are aware that the KNN algorithm can suffer from a dimensionality effect known as the ‘curse of dimensionality’ (Hastie et al. 2009). This can cause training samples to be disproportionately represented and sparsely distributed in high dimensional feature space, especially when the number of input features is greater than the number of training samples. As a consequence, this restricts the performance of machine learning algorithms due to the high complexity learning involved. There are several approaches that can be used to limit the impact of this dimensionality effect, which include feature selection techniques (e.g. Sequential Feature Selection [Guyon & Elisseeff 2003], Chi-Squared Test [Pearson 1900], Fisher Score [Duda et al. 2001]) and feature extraction techniques (e.g. Principal Component Analysis [Pearson 1901; Hotelling 1933], Independent Component Analysis [Comon 1994; Hyvärinen & Oja 2000], Partial Least Squares Regression [Wold 1983; Wold et al. 2001]). These techniques promote useful features and ignore redundant features to subsequently constrain the dimensionality of the feature space. For a classification scenario, Raudys & Jain (1991) suggests that if the number of input features is not too large, such as between five to ten, then at least between fifty to one hundred corresponding training samples are required per class to minimise the ‘curse of dimensionality’. In our case, we ensure that the MWAR training set has a sufficient number of observations in the majority of redshift bins, as seen in Figure SA2 (available online). In addition, we prefer to use a feature selection method that employs features which maximise prediction accuracy rather than a feature extraction method that projects statistically significant features into a reduced feature space.

The most commonly used sequential feature selection strategies are Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). These methods are designed to be computationally efficient by searching through fewer combinations of feature space to provide a quasi-optimal solution rather than a global optimal solution. As described earlier in §§2.2.1, SFS iteratively adds features to an empty feature subset in a forward manner whilst SBS iteratively eliminates features from a full feature subset in a backward manner (Aha & Bankert 1996). This means that SBS will examine more high dimensional combinations of features compared with SFS, which could increase prediction accuracy but at a much higher computational cost. Nonetheless, we decide that SBS is not compatible for this work since the 95 per cent cluster retention threshold would be immediately bypassed if all features are used at the same time, as seen in Figure 5. Although, we could consider SBS as an alternative feature selection strategy in imaging surveys that have greater filter sensitivity than SDSS. We also compare the performance between SFS and manual feature selection. From comparing Figures 4 and S1 (available online), we find that SFS selected features consistently perform better than the manually selected features for the KNN algorithm. This means that SFS is more precise than manual feature selection at taking into account minor details in the datasets. From which, we decide that SFS has better synergy for working with bootstrap resamples in the SRKNN algorithm. It should be noted that we also randomly initialise the input features to the SRKNN algorithm as an additional starting step to SFS to reduce the impact from strong collinear features (see Figures SA11 and SA12 [available online]) during the feature selection process. Furthermore, in Figure 10 it can be seen that using a large number of bootstrap resamples for the SRKNN algorithm improves the stability for the relative frequency of SFS selected features. This

is in contrast to using an individual algorithm such as the KNN algorithm (see Table 2) or using just one bootstrap resample in the SRKNN algorithm. This tells us that the SRKNN algorithm with a large number of bootstrap resamples is able to cope with minor changes to the training set, which would otherwise result in completely different features being used by the model.

The bias-variance tradeoff describes how generalised a supervised machine learning algorithm is at learning a target function (Briscoe & Feldman 2011). If an algorithm is highly dependent on the training dataset during learning, it will perform poorly on new data. This results in many predictions with high variance and low bias. On the other hand, if an algorithm makes a lot of assumptions from the training dataset during learning, it will reduce the predictive power of the algorithm. This results in many predictions with high bias and low variance. For example, the bias-variance tradeoff for the KNN algorithm varies depending on the number of nearest neighbours used, where using low values for the number of nearest neighbours can induce overfitting whilst using high values for the number of nearest neighbours can induce underfitting (Valencia-Zapata et al. 2017; Neal 2019). For the SRKNN algorithm, we examine a wide range of number of nearest neighbours from 1 to 25 but this range could be extended with increased computation in future work to explore a larger number of nearest neighbours. In §3.2 we had chosen a value for the number of nearest neighbours that shows no obvious indications of overfitting or underfitting. It is also known that ensemble algorithms can intrinsically reduce the overall variance of predictions for a model by averaging estimates from multiple models that individually have high variance predictions (e.g. Böhlmann 2012). This effect can be observed in the random forest (RF, Breiman 2001) algorithm, which is an ensemble that averages the estimates from multiple decision trees (DT, Breiman et al. 1984; Quinlan 1986).

The main difference between the SRKNN and RF algorithms is the choice of internal model, such that each ensemble is better suited for different applications. The KNN algorithm utilises instance-based learning (Aha et al. 1991), which means it has no learnable parameters. Whilst the DT algorithm utilises partition-based learning (Strobl et al. 2009), which means it learns optimal splitting parameters. It should be noted that the KNN algorithm can support a similar partition strategy to the DT algorithm by utilising K-Dimensional Tree (Bentley 1975) or Ball Tree search (Omhundro 1989). Generally, the KNN algorithm provides higher flexibility for evaluating complex patterns whereas the DT algorithm has greater interpretability for understanding underlying decisions (Mohanapriya & Jayabalan 2018). In Figures SA13, SA14, SA15, SA16, SA17 and SA18 (available online) we use the t-Distributed Stochastic Neighbour Embedding (t-SNE, van der Maaten & Hinton 2008) algorithm to visualise how the feature space of the MWAR training set appears in two-dimensional space with and without feature scaling applied for each search radius. We observe that galaxies with similar photometric redshifts are somewhat clustered to form smooth transitions from low to intermediate redshifts when feature scaling is applied. Moreover, we also observe that galaxies with similar photometric redshifts are considerably dispersed across feature space when feature scaling is not applied. Nevertheless, the structure of these feature spaces would be difficult for the DT algorithm to apply partitions, whilst the KNN algorithm is better suited to work with these smooth transitions, regardless of whether feature scaling is applied. From which, the SRKNN algorithm would also be more applicable at handling photometry data to estimate photometric redshifts than the RF algorithm.

There are numerous hyper-parameter setting optimisation

strategies available for machine learning algorithms that are suited for different situations. The most commonly used strategies are grid search, random search (Bergstra & Bengio 2012) and Bayesian optimisation (Wu et al. 2019). These strategies require the user to define a range of hyper-parameter setting values that will be explored. The simplest approach is grid search, which evaluates all combinations of hyper-parameter settings but this approach can incur high computational cost. Whereas random search can be computationally cheaper, as it iteratively examines random combinations of hyper-parameter settings to compute an approximate solution. For machine learning algorithms with relatively few hyper-parameter settings, such as linear regression, grid search is more preferable to determine optimal hyper-parameter settings. However, as the number of hyper-parameter settings increases, it becomes computationally favourable to apply random search. Alternatively, if the number of hyper-parameter settings is relatively large, such as neural networks, then it is applicable to employ Bayesian optimisation. This uses Bayes theorem (Bayes & Price 1763; Joyce 2019) to generate probability estimates of the optimal hyper-parameter settings, which involves incorporating prior assumptions of the hyper-parameter settings and iteratively updating a probabilistic distribution of the search space. This means that Bayesian optimisation can minimise the number of hyper-parameter setting combinations that need to be tested. Although in this work, we decide that it is appropriate to utilise grid search to determine the optimal hyper-parameter settings, since the number of hyper-parameter settings for the SRKNN algorithm is relatively low.

We are also aware that the accuracy of photometric redshift estimates has a dependency on the accuracy of the cluster finder used to locate the cluster. For this work, we treat all input data points in CMS with uniform distance weighting. This means that all input data points are not influenced by the distance to the training set data points. However, this may reduce the accuracy of photometric redshift estimates in regions of the sky that have many line-of-sight interloping galaxies since the cluster finder would be unable to clearly define the cluster core, where the red sequence is most well-defined. To limit the dependency on the cluster finder, we could consider simple non-uniform weighting strategies for the SRKNN algorithm such as inverse distance weighting (Dudani 1976). This computes weights based on the distance of the input data points to the training set data points, where the significance of the training set data points decreases as the distance increases. The reason we do not utilise this approach is due to the fact that it is also highly susceptible to noise in the training set. Although, in future work we could consider inverse distance weighting as an alternative, if we can further constrain line-of-sight interloping galaxies within the training set. In addition, the reason we do not utilise photometric redshift estimates of individual galaxies determined by SDSS itself is due to the fact that our method allows us to operate in situations where no photometric redshifts of individual galaxies are available.

In k-fold cross validation the dataset is partitioned into ‘k’ number of folds, whilst in hold-out validation the dataset is split into distinct sets. For k-fold cross validation five or ten ‘k’ folds is commonly employed, whereas for hold-out validation a seventy/thirty or eighty/twenty percentage split of the dataset is typically applied. Each approach is suited for different circumstances to balance between computational cost and bias-variance sample misrepresentation tradeoff (Raschka 2018). This means that k-fold cross validation benefits from a low variance evaluation at a high computational cost. Whereas hold-out validation produces a high variance evaluation but for a low computational cost. In this work, we decide that ten-fold cross validation is appropriate for feature

selection and filter magnitude-cut analysis of the KNN algorithm, as the KNN algorithm has moderate computational training cost requirements. On the other hand, the SRKNN algorithm has higher computational training cost requirements especially when a large number of bootstrap resamples is used. From which, we decide that hold-out validation is more preferable for hyper-parameter tuning of the SRKNN algorithm. However, with increased computation we could consider using k-fold cross validation for hyper-parameter tuning in future work.

5 CONCLUSION

We present Z-Sequence, an empirical model that is composed of an ensemble of the k-nearest neighbours algorithm, known as the sequential random k-nearest neighbours algorithm. The model makes use of photometry data from observed galaxies within a specified search radius to estimate photometric redshifts of clusters. In this proof-of-concept study, we assembled training sets with cross-matched clusters detected in the Sloan Digital Sky Survey by the WHL12 and redMaPPer cluster catalogues, as using cross-matched clusters reduced the likelihood of having false detections in the training set. Whilst clusters that were not cross-matched were used to test the performance of the model. We demonstrated that employing an automated feature selection strategy, known as sequential forward selection, is effective at identifying predictive features from an initial set of features (i.e. filters and colours). We have shown that applying filter magnitude-cuts to the photometry data improved the overall accuracy of photometric redshift estimates, as this excluded galaxies with poor photometric measurements from model predictions. We examined the behaviour of each hyper-parameter setting for the SRKNN algorithm to understand how varying them affected model performance and generalisation. From which, we found that the choice of the number of nearest neighbours had the biggest impact, the choice of the number of initialised random features had moderate impact and the choice of the number of bootstrap resamples used had the least impact. The optimal values for each hyper-parameter setting were subsequently chosen for model testing. Our results showed that the tuned model performed well on clusters that were within the same redshift range (i.e. low and intermediate redshift) as the clusters in the training set and we also demonstrated that the tuned model is effective on clusters of all richnesses that were within the redshift training boundaries. We have shown the photometric redshift prediction error of Z-Sequence via the median value of $|\Delta z|/(1+z)$ on the WHL12 test samples (across a photometric redshift range of $0.05 \leq z \leq 0.599$) to be 0.0106 and on the redMaPPer test samples (across a photometric redshift range of $0.081 \leq z \leq 0.598$) to be 0.0123 within a 10 arcseconds search radius, where the photometric redshift prediction error for both test samples increased by 32 per cent and 47 per cent respectively when the search radius is enlarged to 32 arcseconds. In future work, we aim to apply our technique to imaging surveys as a tool to approximate redshifts for many clusters, such as LSST (Ivezić et al. 2019), Euclid Survey (Laureijs et al. 2011; Euclid Collaboration et al. 2019), Wide Field Instrument High Latitude Survey (Spergel et al. 2015), Hyper Suprime-Cam Subaru Strategic Survey (Aihara et al. 2017; Aihara et al. 2018; Aihara et al. 2019), Dark Energy Survey (Dark Energy Survey Collaboration et al. 2016; Abbott et al. 2018b) and XMM Cluster Survey (Mehrtens et al. 2012). It should be noted that our approach has no prerequisites which means that it is fully data driven. This is beneficial for photometric redshift estimation since Z-Sequence can be

adapted to any imaging survey and trained on galaxy photometry data from known cluster positions in existing cluster catalogues. To prepare for upcoming surveys, we intend to run Z-Sequence as a complementary tool to our own DEEP-CEE (Chan & Stott 2019) cluster finder to examine the entirety of the SDSS sky coverage in a preliminary data pipeline, where clusters detected directly from the astronomical images would be accompanied with estimated photometric redshifts.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for their thorough feedback which has improved the clarity of our paper.

We gratefully acknowledge the support from the Science and Technologies Facilities Council studentship funding and from the High End Computing facility at Lancaster University to perform extensive computations.

We would also like to thank the developers of Vizier (Ochsenbein et al. 2000), Risa Wechsler at Stanford University, TOPCAT (Taylor 2020), James Schombert at the University of Oregon, Edward L. Wright at the University of California, Los Angeles (Wright 2006) and Scikit-Learn (Pedregosa et al. 2011) for allowing the open distribution and free usage of their software for research.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is <http://www.sdss3.org/>.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

DATA AVAILABILITY

The photometry data used in this article is publicly available from the Sloan Digital Sky Survey at <https://vizier.u-strasbg.fr/viz-bin/VizieR?-source=V/139>. The WHL12 and redMaPPer v6.3 cluster catalogues can also be found in the public domain at <http://vizier.u-strasbg.fr/viz-bin/VizieR?-source=J/ApJS/199/34> and <http://risa.stanford.edu/redmapper/>.

REFERENCES

- Abazajian K., et al., 2004, *AJ*, 128, 502
- Abbott T. M. C., et al., 2018a, *Phys. Rev. D*, 98, 043526
- Abbott T. M. C., et al., 2018b, *ApJS*, 239, 18

- Aggarwal C., Hinneburg A., Keim D., 2002, First publ. in: Database theory, ICDDT 200, 8th International Conference, London, UK, January 4 - 6, 2001 / Jan Van den Bussche ... (eds.). Berlin: Springer, 2001, pp. 420-434 (=Lecture notes in computer science ; 1973)
- Aha D. W., Bankert R. L., 1996, in , Learning from data. Springer, New York, NY, USA, pp 199–206
- Aha W., Kibler D., Albert M., 1991, *Machine Learning*, 6, 37
- Ahn C. P., et al., 2012, *ApJS*, 203, 21
- Aihara H., et al., 2017, *Publications of the Astronomical Society of Japan*, 70, S4
- Aihara H., et al., 2018, *PASJ*, 70, S8
- Aihara H., et al., 2019, *PASJ*, 71, 114
- Alam S., et al., 2017, *MNRAS*, 470, 2617
- Amendola L., et al., 2018, *Living Reviews in Relativity*, 21, 2
- Ata M., et al., 2018, *MNRAS*, 473, 4773
- Babbedge T. S. R., et al., 2004, *Monthly Notices of the Royal Astronomical Society*, 353, 654
- Baldry I. K., et al., 2010, *MNRAS*, 404, 86
- Bayes T., Price n., 1763, *Philosophical Transactions of the Royal Society of London*, 53, 370
- Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2016, *MNRAS*, 460, 1371
- Bentley J. L., 1975, *Commun. ACM*, 18, 509
- Bergstra J., Bengio Y., 2012, *J. Mach. Learn. Res.*, 13, 281
- Beutler F., et al., 2017, *MNRAS*, 466, 2242
- Bilicki M., et al., 2018, *A&A*, 616, A69
- Blanton M. R., et al., 2017, *AJ*, 154, 28
- Böhlmann P., 2012, *Handbook of Computational Statistics*, pp 985–1022
- Bolzonella M., Miralles J. M., Pelló R., 2000, *A&A*, 363, 476
- Breiman L., 2001, *Machine Learning*, 45, 5
- Breiman L., Friedman J., Stone C., Olshen R., 1984, *Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series*, Taylor & Francis, Boca Raton, Florida, USA, <https://books.google.co.uk/books?id=JwQx-W0mSyQC>
- Briscoe E., Feldman J., 2011, *Cognition*, 118, 2
- Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, 432, 1483
- Chan M. C., Stott J. P., 2019, *MNRAS*, 489, 5770
- Colless M., et al., 2001, *MNRAS*, 328, 1039
- Comon P., 1994, *Signal Processing*, 36, 287
- Cover T. M., Hart P. E., 1967, *IEEE Trans. Inf. Theory*, 13, 21
- Dark Energy Survey Collaboration et al., 2016, *MNRAS*, 460, 1270
- Dietterich T. G., 2000, in International workshop on multiple classifier systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–15
- Doi M., et al., 2010, *AJ*, 139, 1628
- Dressler A., 1980, *ApJ*, 236, 351
- Dressler A., 1984, *ARA&A*, 22, 185
- Dressler A., Shectman S. A., 1987, *AJ*, 94, 899
- Duda R., Hart P., Stork D., 2001, *Pattern Classification*, 2 edn. John Wiley & Sons, New York
- Dudani S. A., 1976, *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6, 325
- Efron B., 1979, *The Annals of Statistics*, 7, 1
- Efron B., Tibshirani R., 1986, *Statistical Science*, 1, 54
- Efron B., Tibshirani R. J., 1994, *An introduction to the bootstrap*. CRC press, New York, NY, USA
- Eisenstein D. J., et al., 2011, *AJ*, 142, 72
- Euclid Collaboration et al., 2019, *A&A*, 627, A23
- Falco E. E., et al., 1999, *PASP*, 111, 438
- Fix E., 1951, Technical report, Discriminatory analysis: nonparametric discrimination, consistency properties. Randolph Field, Texas, USA
- Fotopoulou S., Paltani S., 2018, *A&A*, 619, A14
- Friedman J., Hastie T., Tibshirani R., 2001, *The elements of statistical learning*. Vol. 1, Springer New York
- Gil-Marín H., Percival W. J., Verde L., Brownstein J. R., Chuang C.-H., Kitaura F.-S., Rodríguez-Torres S. A., Olmstead M. D., 2017, *MNRAS*, 465, 1757
- Girardi M., Biviano A., Giuricin G., Mardirossian F., Mezzetti M., 1995, *ApJ*, 438, 527
- Gladders M. D., Yee H. K. C., 2000, *AJ*, 120, 2148
- Gladders M. D., López-Cruz O., Yee H. K. C., Kodama T., 1998, *ApJ*, 501, 571
- Goodfellow I., Bengio Y., Courville A., 2016, *Deep learning*. MIT press, Cambridge, Massachusetts, USA
- Gorecki A., Abate A., Ansari R., Barrau A., Baumont S., Moniez M., Ricol J.-S., 2014, *A&A*, 561, A128
- Gutierrez-Osuna R., 2011, Lecture notes in CSCE 666 Pattern Analysis, http://research.cs.tamu.edu/prism/lectures/pr/pr_111.pdf
- Guyon I., Elisseeff A., 2003, *J. Mach. Learn. Res.*, 3, 1157–1182
- Hamilton D., 1985, *ApJ*, 297, 371
- Hastie T., Tibshirani R., Friedman J., 2009, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media
- Hotelling H., 1933, *Journal of educational psychology*, 24, 417
- Howlett C., Ross A. J., Samushia L., Percival W. J., Manera M., 2015, *MNRAS*, 449, 848
- Hsieh B. C., Yee H. K. C., Lin H., Gladders M. D., 2005, *ApJS*, 158, 161
- Hu L., Huang M., Ke S., Tsai C., 2016, *SpringerPlus*, 5, 1304
- Huchra J. P., et al., 2012, *ApJS*, 199, 26
- Huss A., Jain B., Steinmetz M., 1999, *Monthly Notices of the Royal Astronomical Society*, 308, 1011
- Hyvärinen A., Oja E., 2000, *Neural Netw.*, 13, 411
- Ilbert O., et al., 2009, *ApJ*, 690, 1236
- Ivezić Ž., et al., 2004, *Astronomische Nachrichten*, 325, 583
- Ivezić Ž., et al., 2019, *ApJ*, 873, 111
- Jin S., Gu Q., Huang S., Shi Y., Feng L., 2014, *ApJ*, 787, 63
- Jones D. H., et al., 2009, *MNRAS*, 399, 683
- Joudaki S., et al., 2018, *MNRAS*, 474, 4894
- Joyce J., 2019, in Zalta E. N., ed., , *The Stanford Encyclopedia of Philosophy*, spring 2019 edn, Metaphysics Research Lab, Stanford University
- Kang H., 2013, *Korean journal of anesthesiology*, 64, 402
- Kauffmann G., et al., 2003, *MNRAS*, 341, 33
- Kodama T., Arimoto N., Barger A. J., Arag'ón-Salamanca A., 1998, *A&A*, 334, 99
- Kotsiantis S. B., Zaharakis I., Pintelas P., 2007, *Emerging artificial intelligence applications in computer engineering*, 160, 3
- Kravtsov A. V., Borgani S., 2012, *ARA&A*, 50, 353
- Laigle C., et al., 2016, *ApJS*, 224, 24
- Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193
- Lidman C., et al., 2008, *A&A*, 489, 981
- Lopes P. A. A., 2007, *Monthly Notices of the Royal Astronomical Society*, 380, 1608
- Lupton R. H., Gunn J. E., Szalay A. S., 1999, *AJ*, 118, 1406
- McCulloch W., Pitts W., 1943, *Bulletin of Mathematical Biophysics*, 5, 115
- Mehrtens N., et al., 2012, *MNRAS*, 423, 1024
- Mei S., et al., 2009, *ApJ*, 690, 42
- Mohanapriya M., Jayabalan L., 2018, *Journal of Physics: Conference Series*, 1142, 012011
- Neal B., 2019, arXiv e-prints, p. arXiv:1912.08286
- Newman A. B., Ellis R. S., Andreon S., Treu T., Raichoor A., Trinchieri G., 2014, *ApJ*, 788, 51
- Ochsenbein F., Bauer P., Marcout J., 2000, *A&AS*, 143, 23
- Oke J. B., Gunn J. E., 1983, *ApJ*, 266, 713
- Omohundro S. M., 1989, Technical report, Five balltree construction algorithms. Berkeley, California, USA
- Park C. H., Kim S. B., 2015, *Expert Systems with Applications*, 42, 2336
- Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, *A&A*, 621, A26
- Pearson K., 1900, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50, 157
- Pearson K., 1901, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 559
- Pedregosa F., et al., 2011, *Journal of Machine Learning Research*, 12, 2825
- Piattella O. F., 2018, *Lecture Notes in Cosmology*. UNITEXT for Physics, Springer International Publishing, Cham, Switzerland (arXiv:1803.00070), doi:10.1007/978-3-319-95570-4

- Planck Collaboration et al., 2016, *A&A*, **594**, A13
- Quinlan J. R., 1986, *Mach. Learn.*, **1**, 81
- Raschka S., 2014, Sebastian Racha. Disques, nd Web. Dec
- Raschka S., 2018, arXiv e-prints, p. [arXiv:1811.12808](https://arxiv.org/abs/1811.12808)
- Raudys S., Jain A., 1991, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **13**, 252
- Reitermanova Z., 2010, in WDS. MatfyzPress, Charles University, Prague, Czech Republic, pp 31–36
- Ross A. J., et al., 2017, *MNRAS*, **464**, 1168
- Rykoff E. S., et al., 2014, *ApJ*, **785**, 104
- Salvato M., Ilbert O., Hoyle B., 2019, *Nature Astronomy*, **3**, 212
- Sánchez C., et al., 2014, *Monthly Notices of the Royal Astronomical Society*, **445**, 1482
- Schawinski K., et al., 2009, *Monthly Notices of the Royal Astronomical Society*, **396**, 818
- Schlafly E. F., Finkbeiner D. P., 2011, *ApJ*, **737**, 103
- Schlegel D. J., Finkbeiner D. P., Davis M., 1998, *ApJ*, **500**, 525
- Segev N., El-Yaniv R., 2016, PhD thesis, Computer Science Department, Technion
- Spergel D., et al., 2015, arXiv e-prints, p. [arXiv:1503.03757](https://arxiv.org/abs/1503.03757)
- Stone M., 1974, *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, 111
- Stott J. P., Smail I., Edge A. C., Ebeling H., Smith G. P., Kneib J.-P., Pimblett K. A., 2007, *The Astrophysical Journal*, **661**, 95
- Stott J. P., Pimblett K. A., Edge A. C., Smith G. P., Wardlow J. L., 2009, *MNRAS*, **394**, 2098
- Strauss M. A., et al., 2002, *AJ*, **124**, 1810
- Strazzullo V., et al., 2016, *ApJ*, **833**, L20
- Strobl C., Malley J., Tutz G., 2009, *Psychological methods*, **14**, 323
- Tanaka M., et al., 2018, *PASJ*, **70**, S9
- Taylor M. B., 2020, in Ballester P., Ibsen J., Solar M., Shorridge K., eds, *Astronomical Society of the Pacific Conference Series Vol. 522, Astronomical Data Analysis Software and Systems XXVII*. p. 67
- Torrey L., Shavlik J., 2010, in , *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, Hershey, Pennsylvania, USA, pp 242–264
- Valencia-Zapata G., Mejia D., Klimeck G., Zentner M., Ersoy O., 2017, arXiv e-prints, p. [arXiv:1709.01439](https://arxiv.org/abs/1709.01439)
- VanderPlas J., 2016, *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., Sebastopol, California, USA
- Vilalta R., Carrier C., Brazdil P., Soares C. M., et al., 2010, *Inductive transfer*. Springer US, Boston, Massachusetts, USA, pp 545–548, doi:10.1007/978-0-387-30164-8_401, https://doi.org/10.1007/978-0-387-30164-8_401
- Walcher J., Groves B., Budavari T., Dale D., 2011, *Astrophysics and Space Science*, **331**, 1
- Webb G. I., 2010, *Lazy Learning*. Springer US, Boston, Massachusetts, USA, pp 571–572, doi:10.1007/978-0-387-30164-8_443, https://doi.org/10.1007/978-0-387-30164-8_443
- Weinstein M. A., et al., 2004, *ApJS*, **155**, 243
- Wen Z. L., Han J. L., Liu F. S., 2009, *ApJS*, **183**, 197
- Wen Z. L., Han J. L., Liu F. S., 2012, *ApJS*, **199**, 34
- Wold H., 1983, Iiasa collaborative paper, *Systems Analysis by Partial Least Squares*, <http://pure.iiasa.ac.at/id/eprint/2336/>. IIASA, Laxenburg, Austria, <http://pure.iiasa.ac.at/id/eprint/2336/>
- Wold S., Sjöström M., Eriksson L., 2001, *Chemometrics and Intelligent Laboratory Systems*, **58**, 109
- Wolf C., et al., 2009, in Wang W., Yang Z., Luo Z., Chen Z., eds, *Astronomical Society of the Pacific Conference Series Vol. 408, The Starburst-AGN Connection*. p. 248 ([arXiv:0906.0306](https://arxiv.org/abs/0906.0306))
- Wright E. L., 2006, *PASP*, **118**, 1711
- Wu J., Chen X.-Y., Zhang H., Xiong L.-D., Lei H., Deng S.-H., 2019, *Journal of Electronic Science and Technology*, **17**, 26
- Yee H. K. C., Gladders M. D., López-Cruz O., 1999, in Weymann R., Storrie-Lombardi L., Sawicki M., Brunner R., eds, *Astronomical Society of the Pacific Conference Series Vol. 191, Photometric Redshifts and the Detection of High Redshift Galaxies*. p. 166 ([arXiv:astro-ph/9908001](https://arxiv.org/abs/astro-ph/9908001))
- York D. G., et al., 2000, *AJ*, **120**, 1579
- da Costa L. N., et al., 1998, *AJ*, **116**, 1
- de Haan T., et al., 2016, *ApJ*, **832**, 95
- de Propris R., Stanford S. A., Eisenhardt P. R., Dickinson M., Elston R., 1999, *AJ*, **118**, 719
- van der Maaten L., Hinton G., 2008, *Journal of Machine Learning Research*, **9**, 2579

SUPPLEMENTARY MATERIAL (ONLINE)

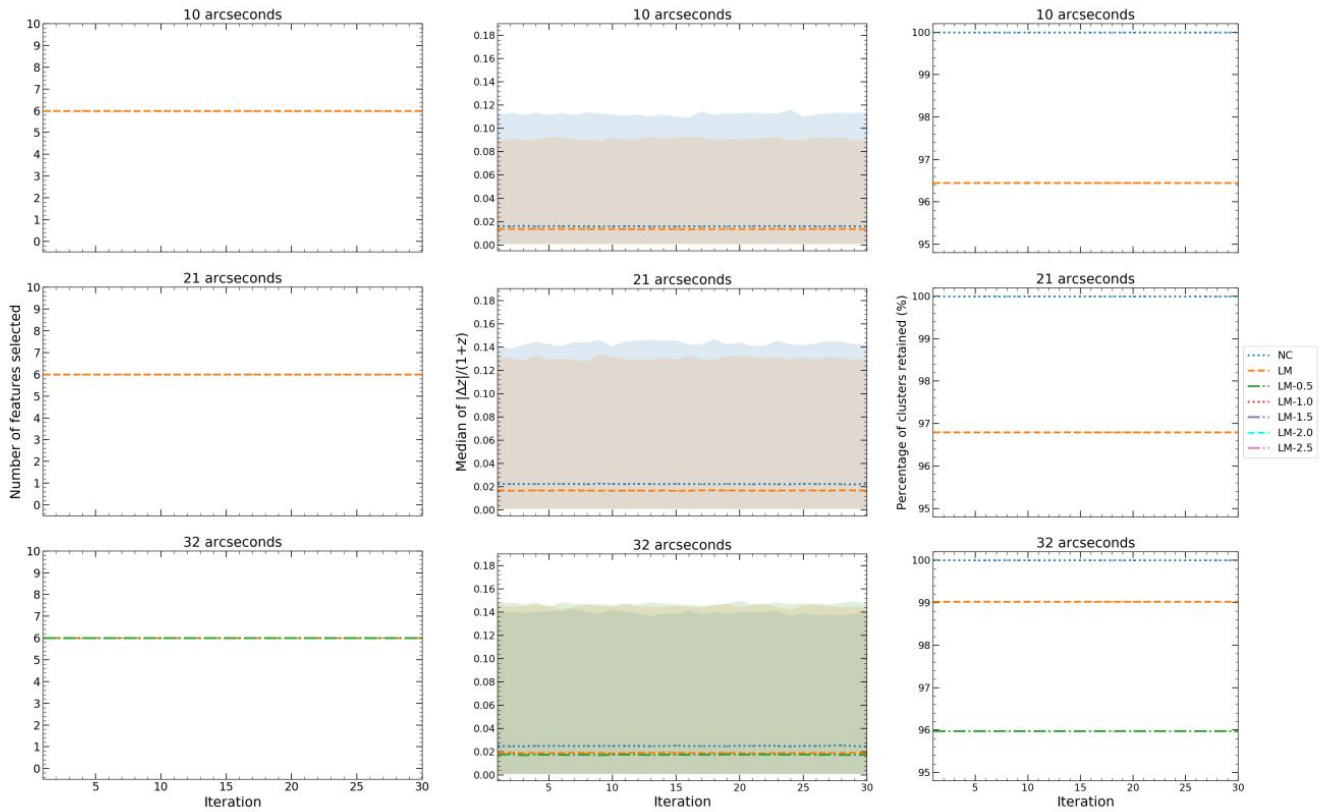


Figure S1. Plots displaying the results from applying filter magnitude-cuts to the MWAR training set using a single KNN algorithm with a control group of manually selected features for each search radii (10 arcseconds on the top row, 21 arcseconds on the middle row and 32 arcseconds on the bottom row). `NC' represents a dataset with no filter magnitude-cuts applied and `LM' represents the MWAR dataset with SFS selected features where filter magnitude-cuts are applied to the limiting magnitude of SDSS. In addition, `LM' is the faintest filter magnitude-cut whilst `LM-2.5' is the brightest filter magnitude-cut. Left column: Number of features selected for the control group feature subset in ten-fold cross validation across thirty repeats. Middle column: Median of photometric redshift prediction errors ($|\Delta z|/(1+z)$) across all tested clusters for the control group feature subset in ten-fold cross validation across thirty repeats, where the shaded regions represent 95 per cent confidence intervals. Right column: Percentage of test clusters retained after filter magnitude-cuts are applied with the control group feature subset in ten-fold cross validation across thirty repeats. It should be noted that if the percentage of clusters retained, after filter magnitude-cuts are applied, do not satisfy the 95 per cent cluster retention threshold we would not display the corresponding results in the other columns.

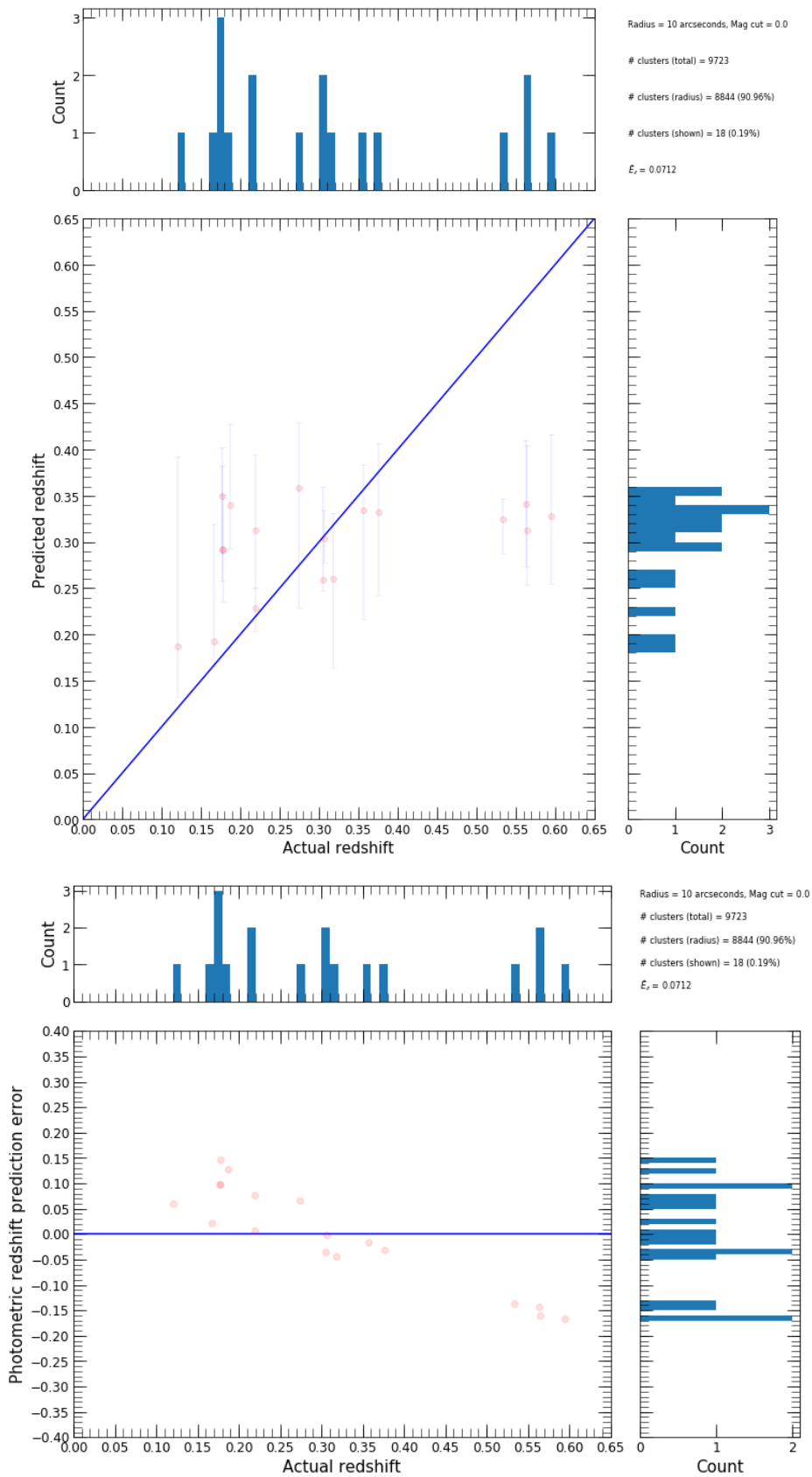


Figure S2. Plots displaying the performance of photometric redshift predictions of clusters for the WNMR dataset that had partial bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters in the WNMR dataset, '# clusters (radius)' represents the number of clusters in the WNMR test set that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters in the WNMR test set that have observed galaxies within a 10 arcseconds search radius with partial bootstrap resamples returned, $\bar{\epsilon}_z$ represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

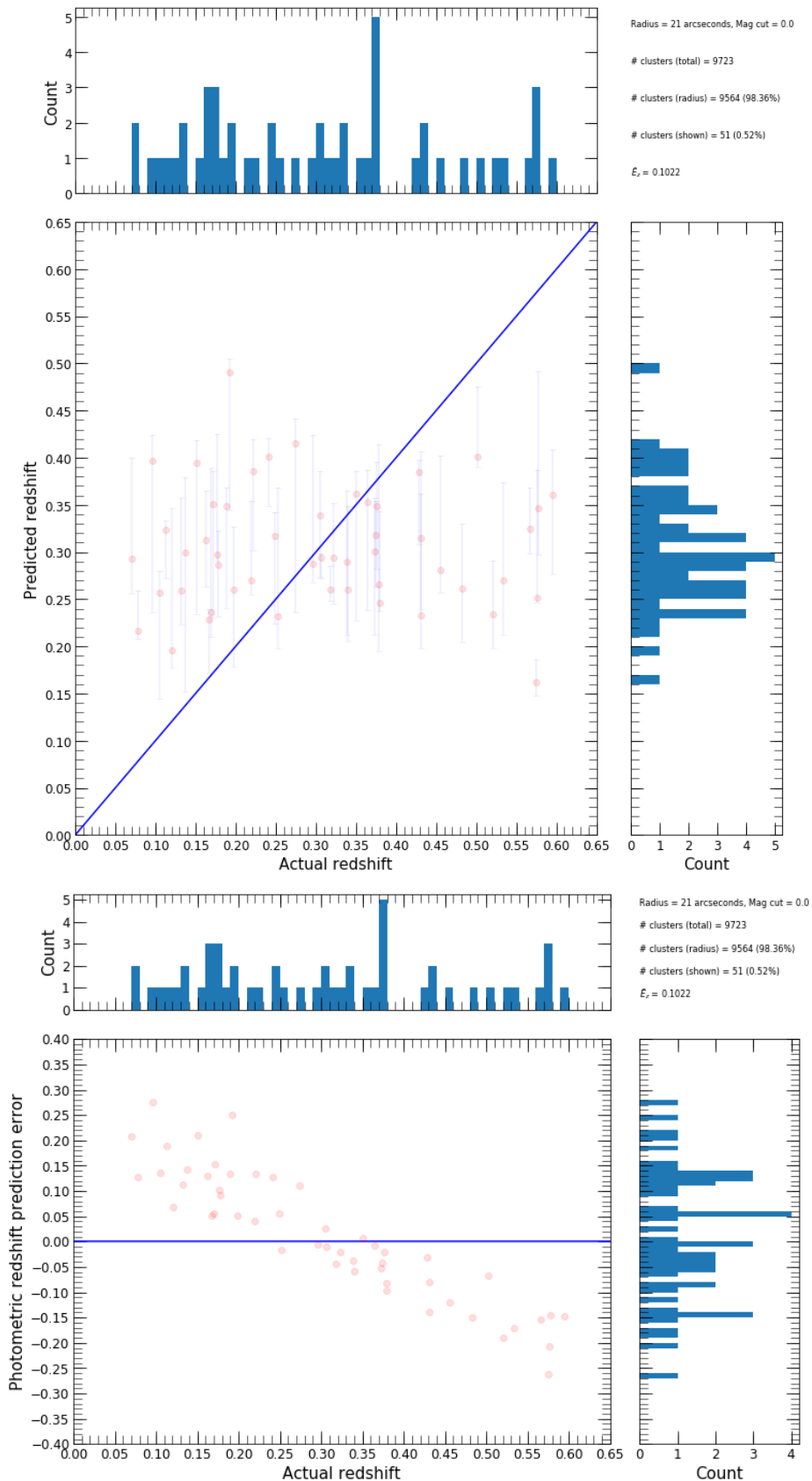


Figure S3. Plots displaying the performance of photometric redshift predictions of clusters for the WNMR dataset that had partial bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters in the WNMR dataset, '# clusters (radius)' represents the number of clusters in the WNMR test set that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters in the WNMR test set that have observed galaxies within a 21 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with partial bootstrap resamples returned.

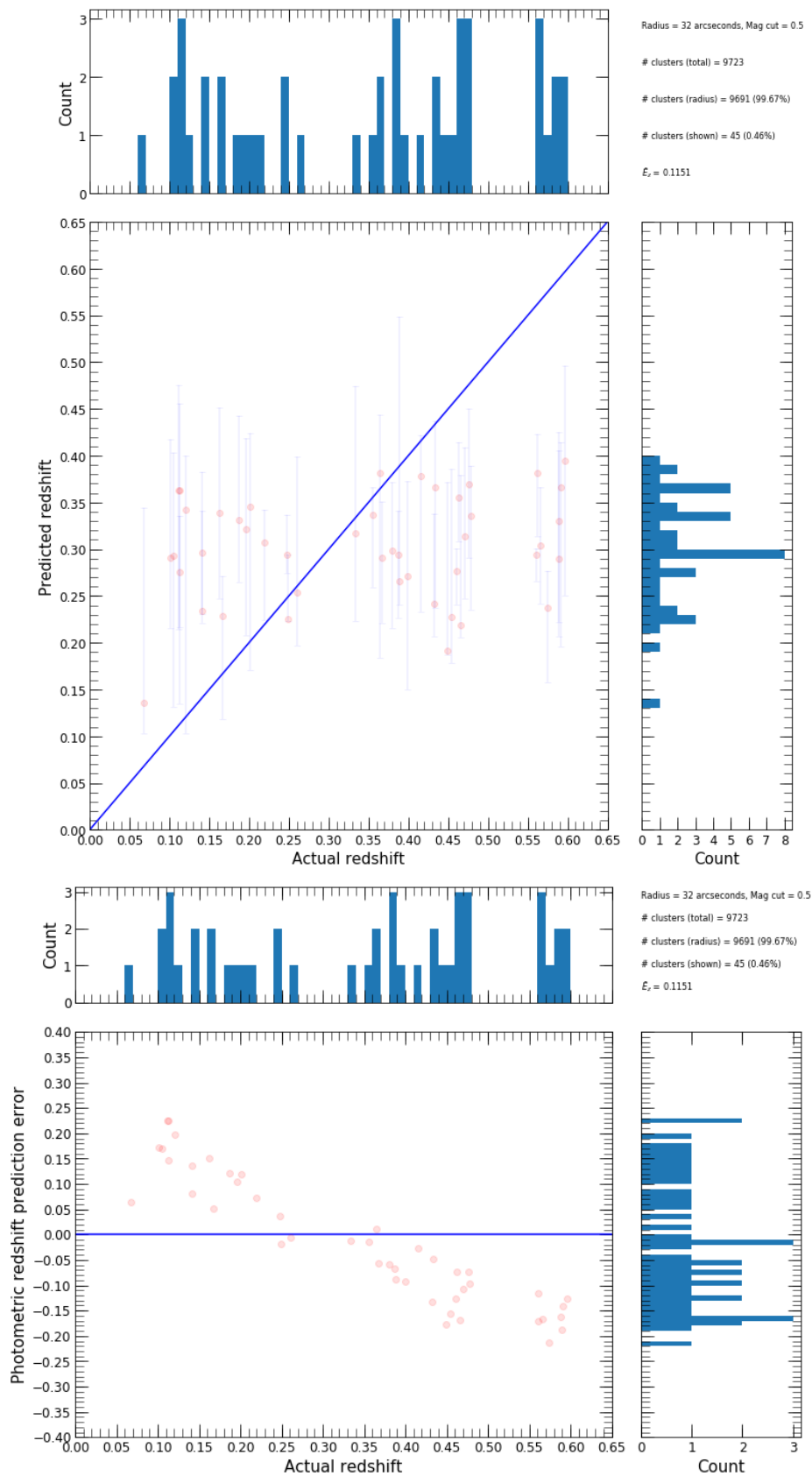


Figure S4. Plots displaying the performance of photometric redshift predictions of clusters for the WNMR dataset that had partial bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters in the WNMR dataset, '# clusters (radius)' represents the number of clusters in the WNMR test set that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters in the WNMR test set that have observed galaxies within a 32 arcseconds search radius with partial bootstrap resamples returned, \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with partial bootstrap resamples returned.

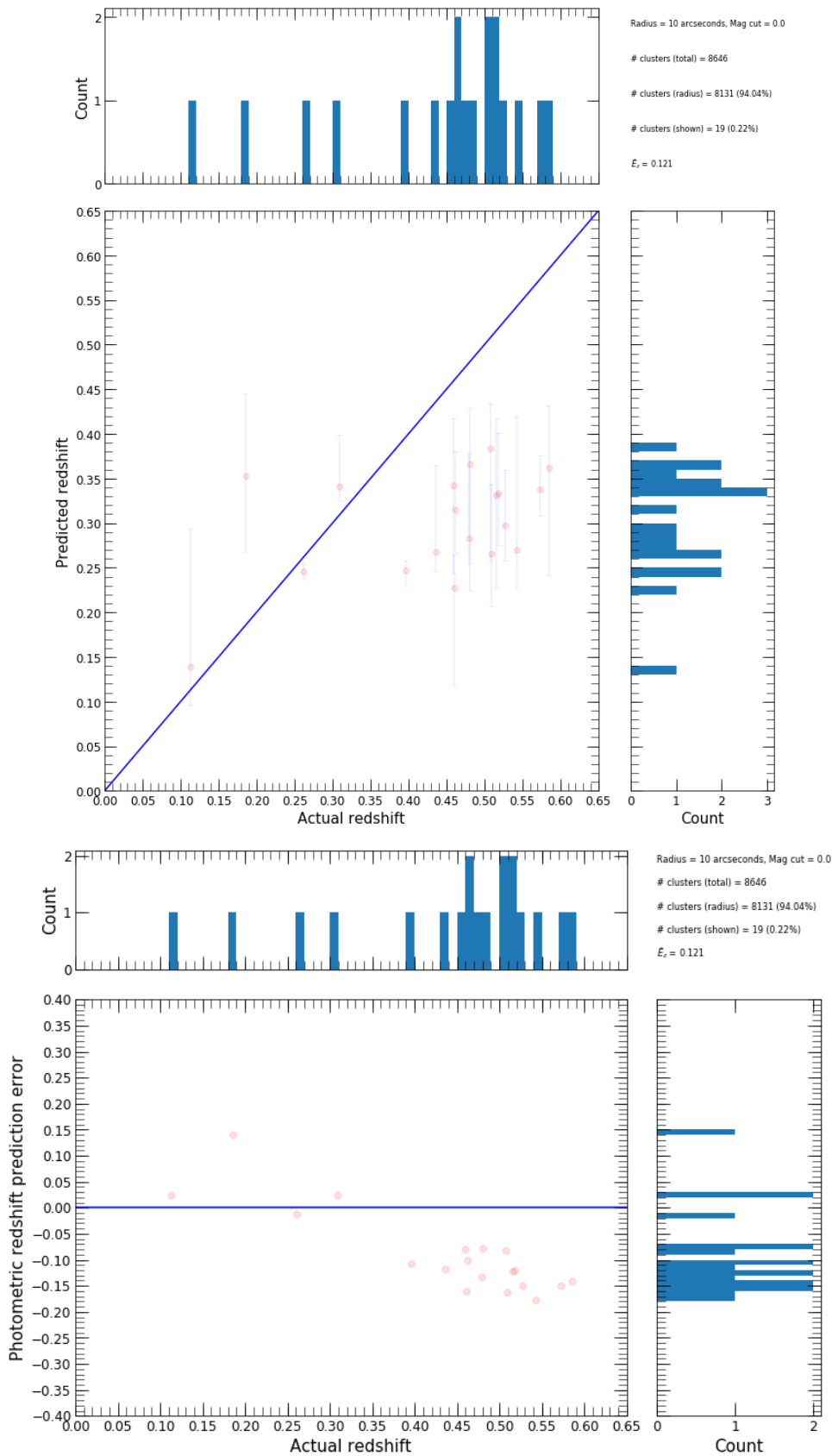


Figure S5. Plots displaying the performance of photometric redshift predictions of clusters for the RNMW dataset that had partial bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters in the RNMW dataset, '# clusters (radius)' represents the number of clusters in the RNMW test set that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters in the RNMW test set that have observed galaxies within a 10 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

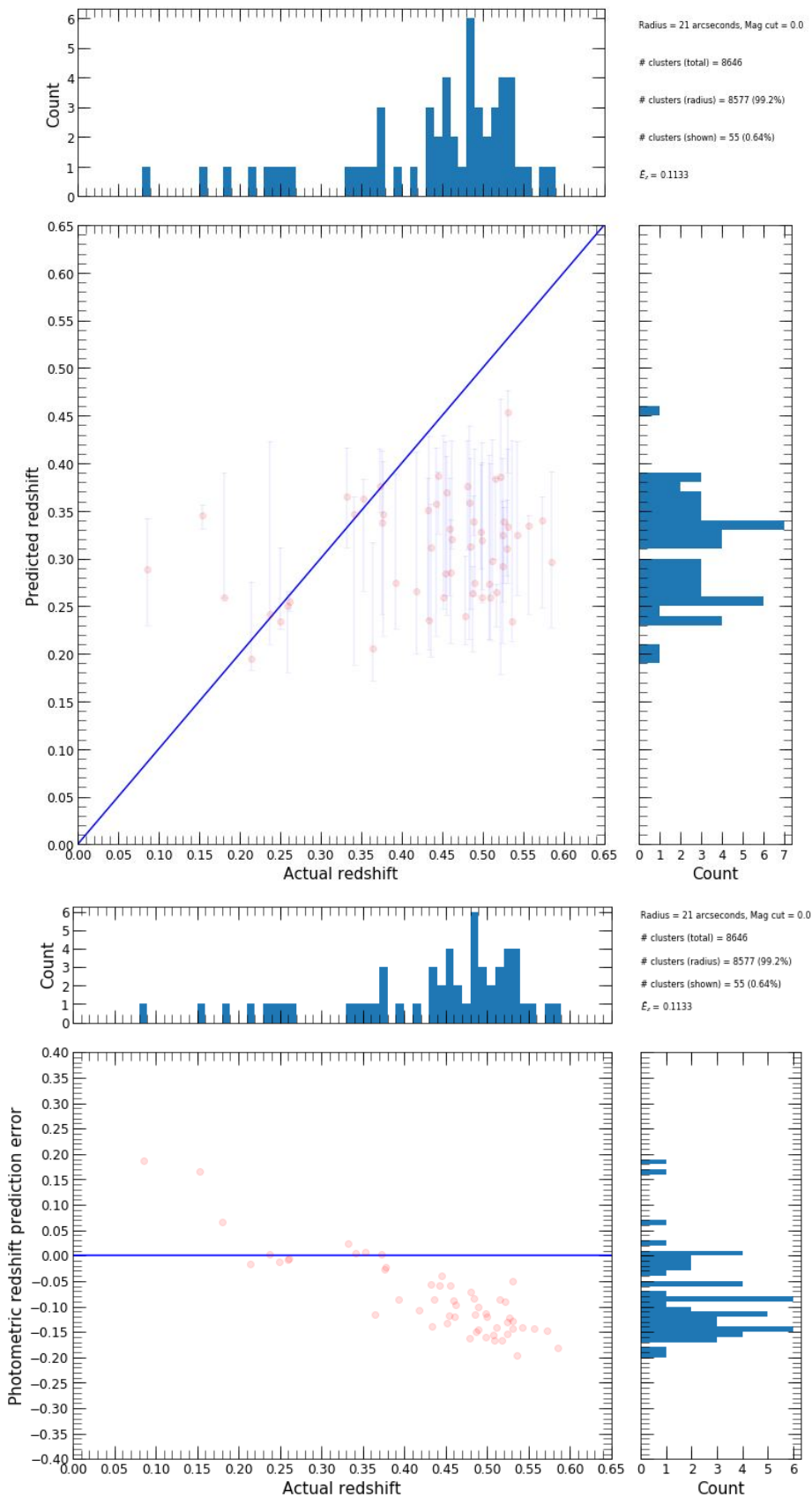


Figure S6. Plots displaying the performance of photometric redshift predictions of clusters for the RNMW dataset that had partial bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters in the RNMW dataset, '# clusters (radius)' represents the number of clusters in the RNMW test set that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters in the RNMW test set that have observed galaxies within a 21 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with partial bootstrap resamples returned.

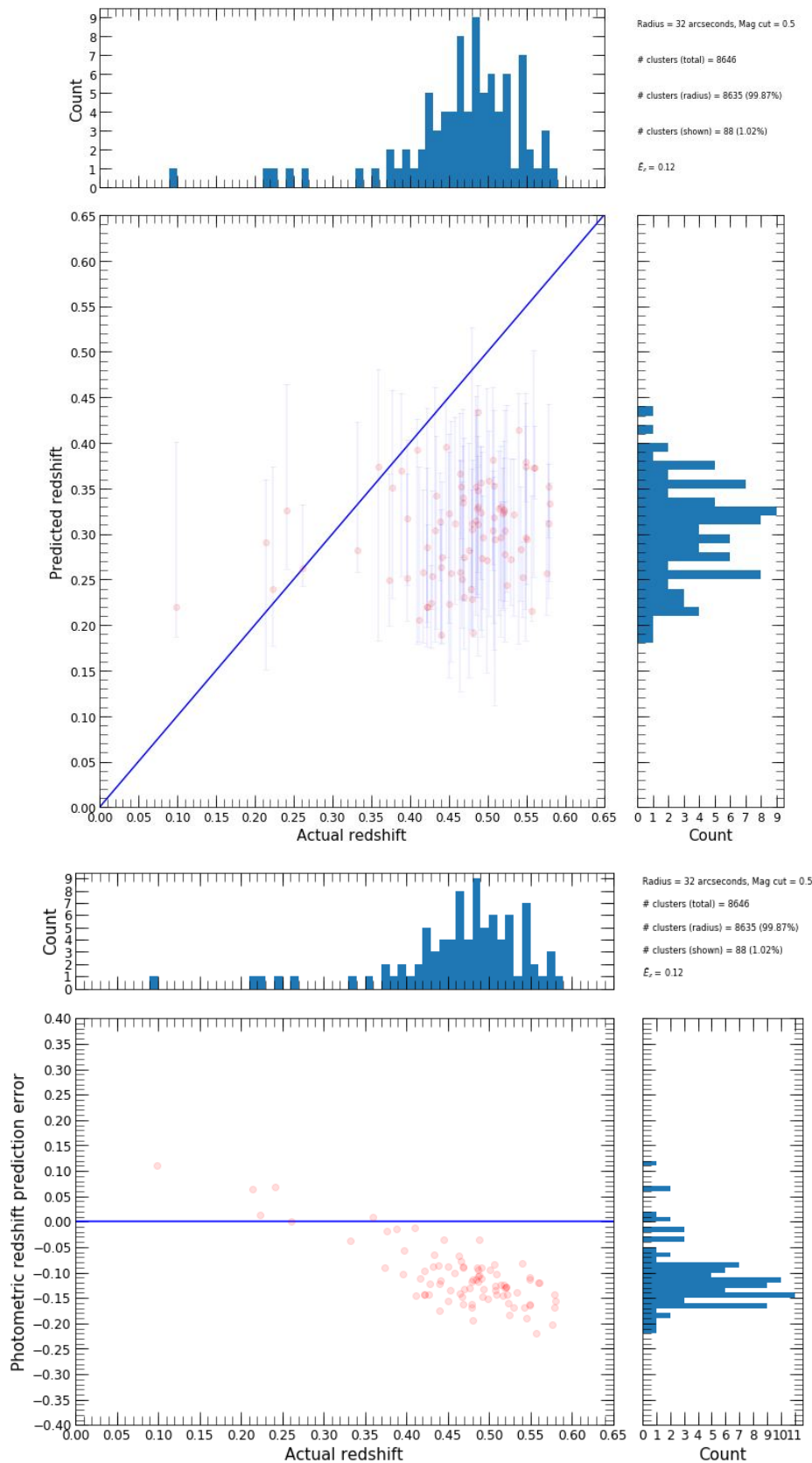


Figure S7. Plots displaying the performance of photometric redshift predictions of clusters for the RNMW dataset that had partial bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters in the RNMW dataset, '# clusters (radius)' represents the number of clusters in the RNMW test set that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters in the RNMW test set that have observed galaxies within a 32 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with partial bootstrap resamples returned.

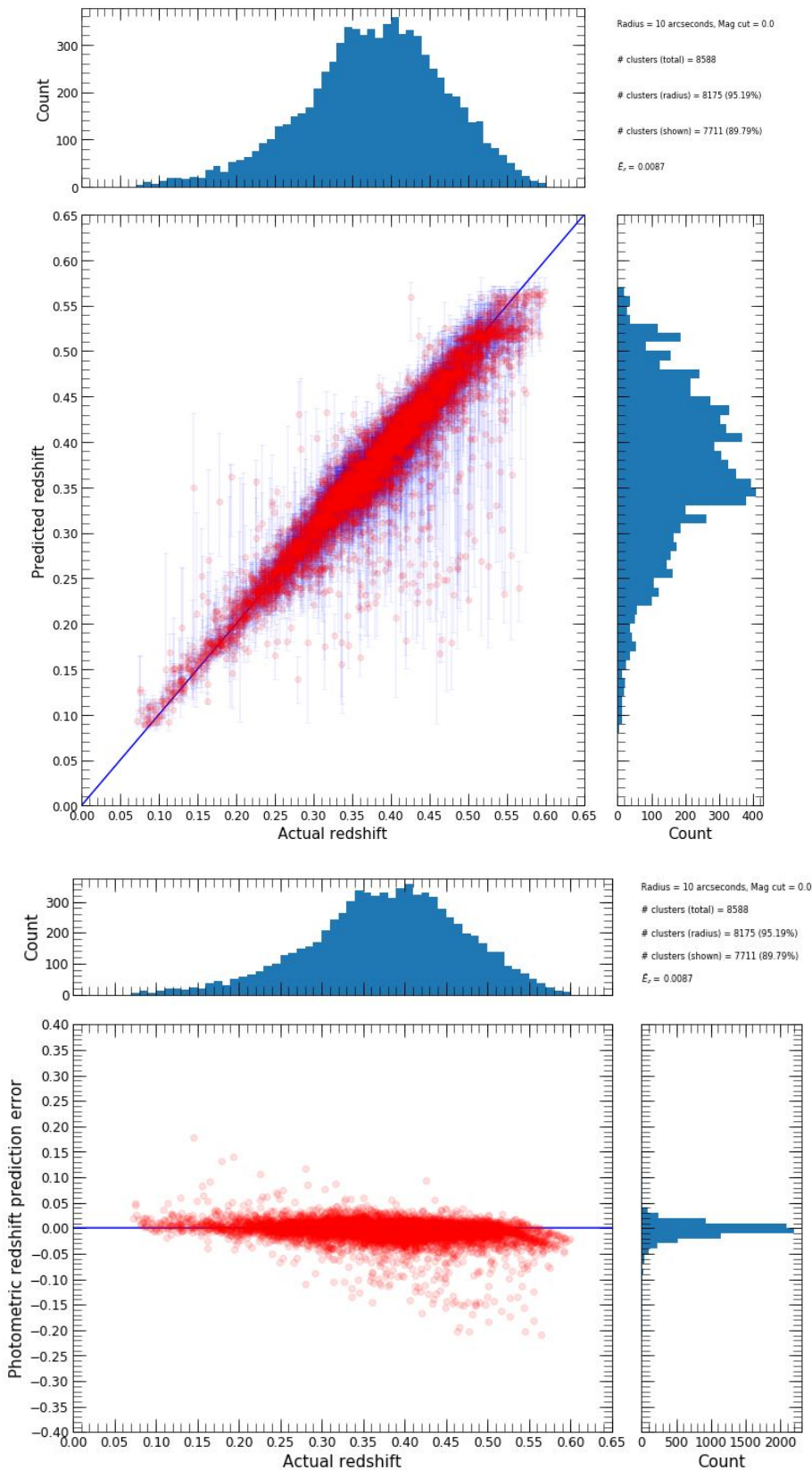


Figure S8. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the MWAR dataset, that had full bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the MWAR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 10 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with full bootstrap resamples returned.

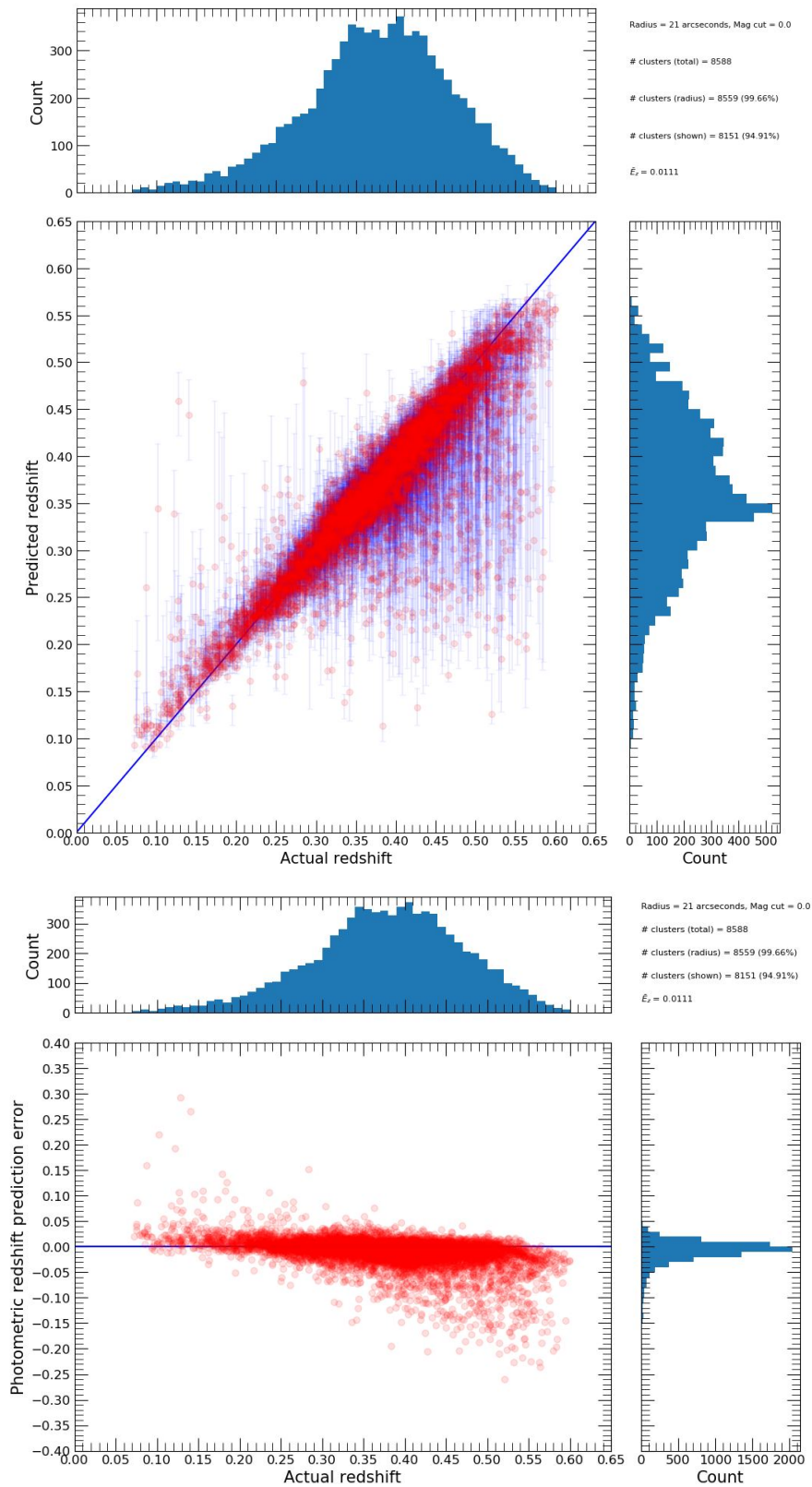


Figure S9. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the MWAR dataset, that had full bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the MWAR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 21 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with full bootstrap resamples returned.

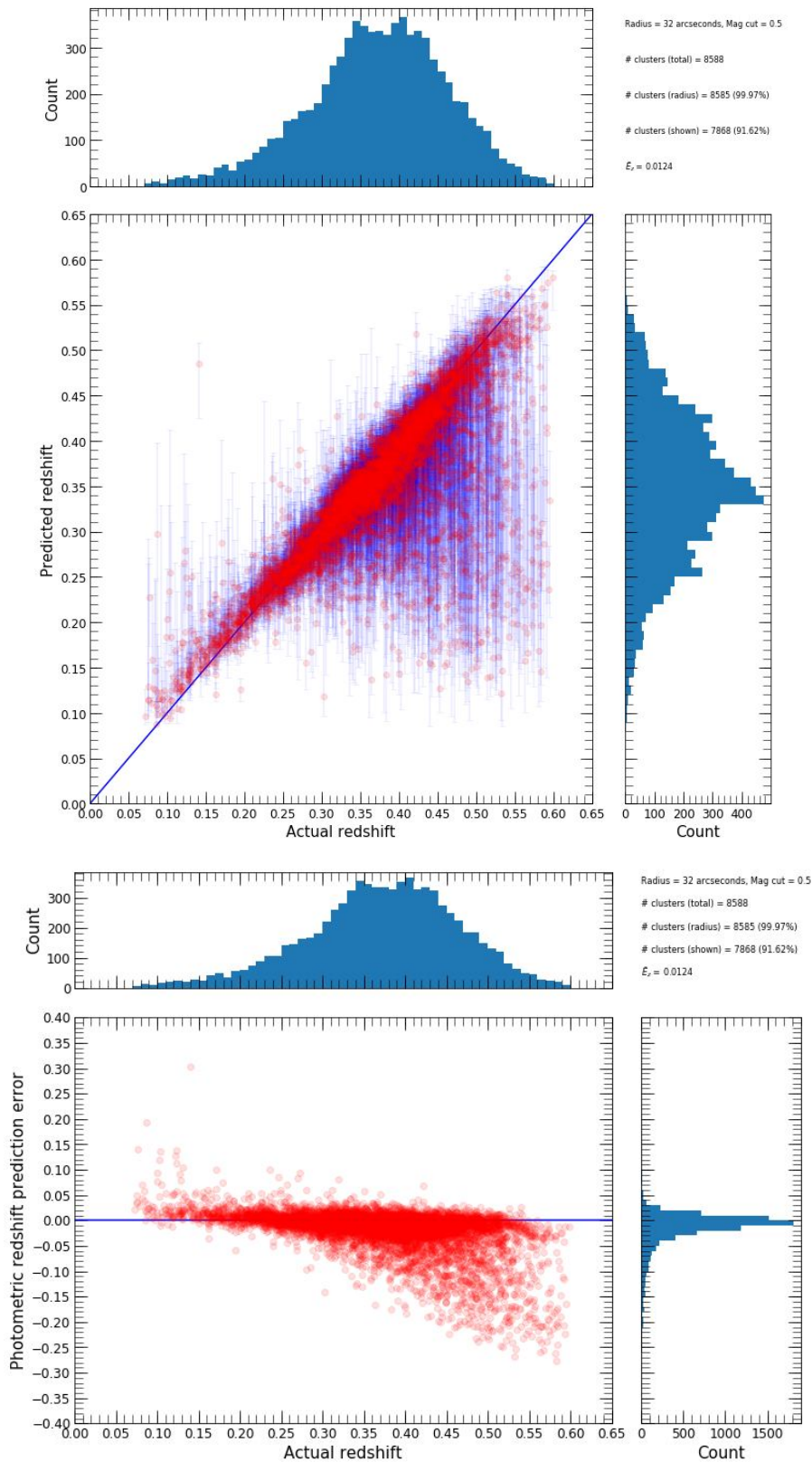


Figure S10. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the MWAR dataset, that had full bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the MWAR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 32 arcseconds search radius with full bootstrap resamples returned, \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with full bootstrap resamples returned.

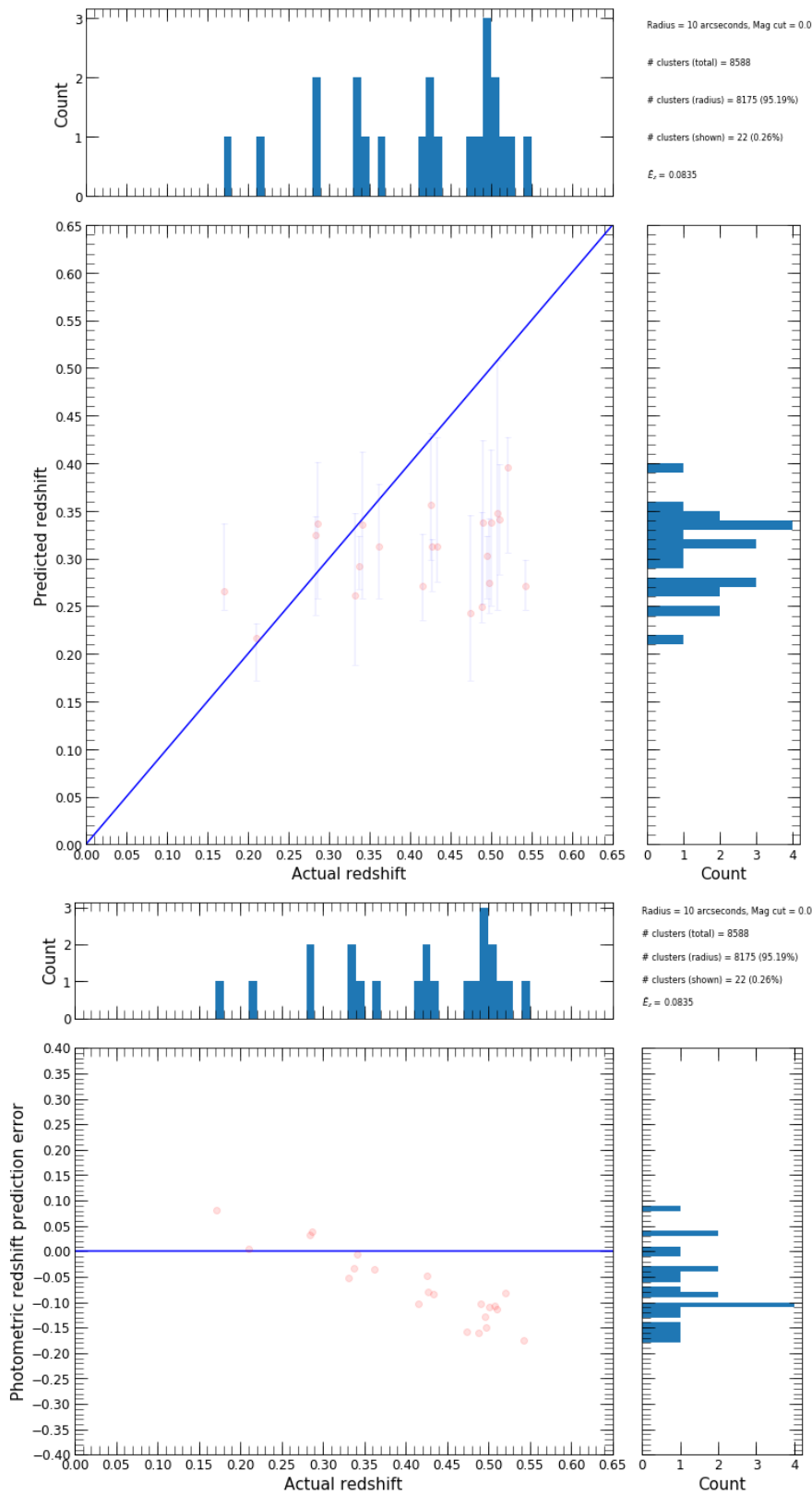


Figure S11. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the MWAR dataset, that had partial bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the MWAR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 10 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

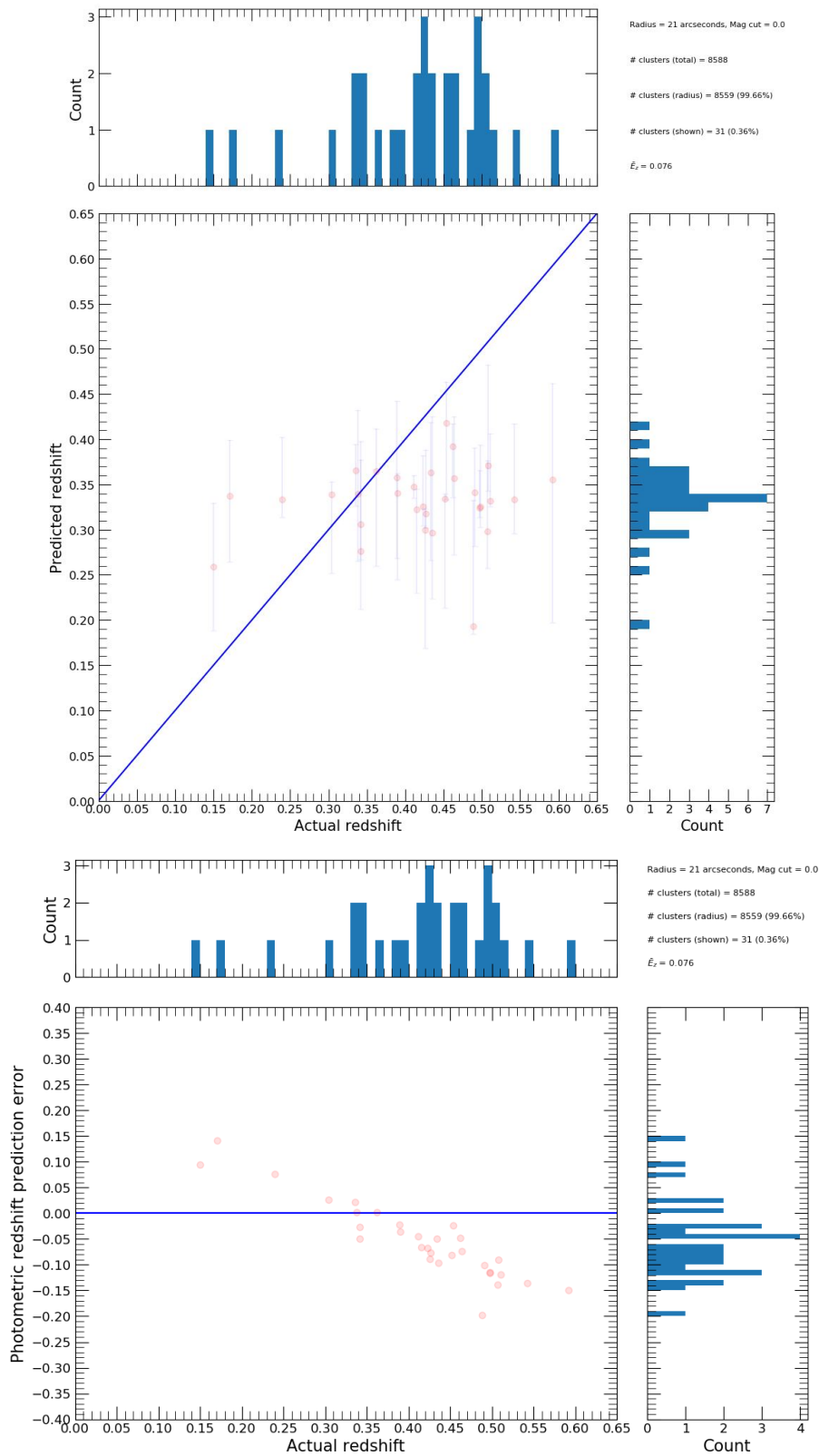


Figure S12. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the MWAR dataset, that had partial bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the MWAR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 21 arcseconds search radius with partial bootstrap resamples returned, \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with partial bootstrap resamples returned.

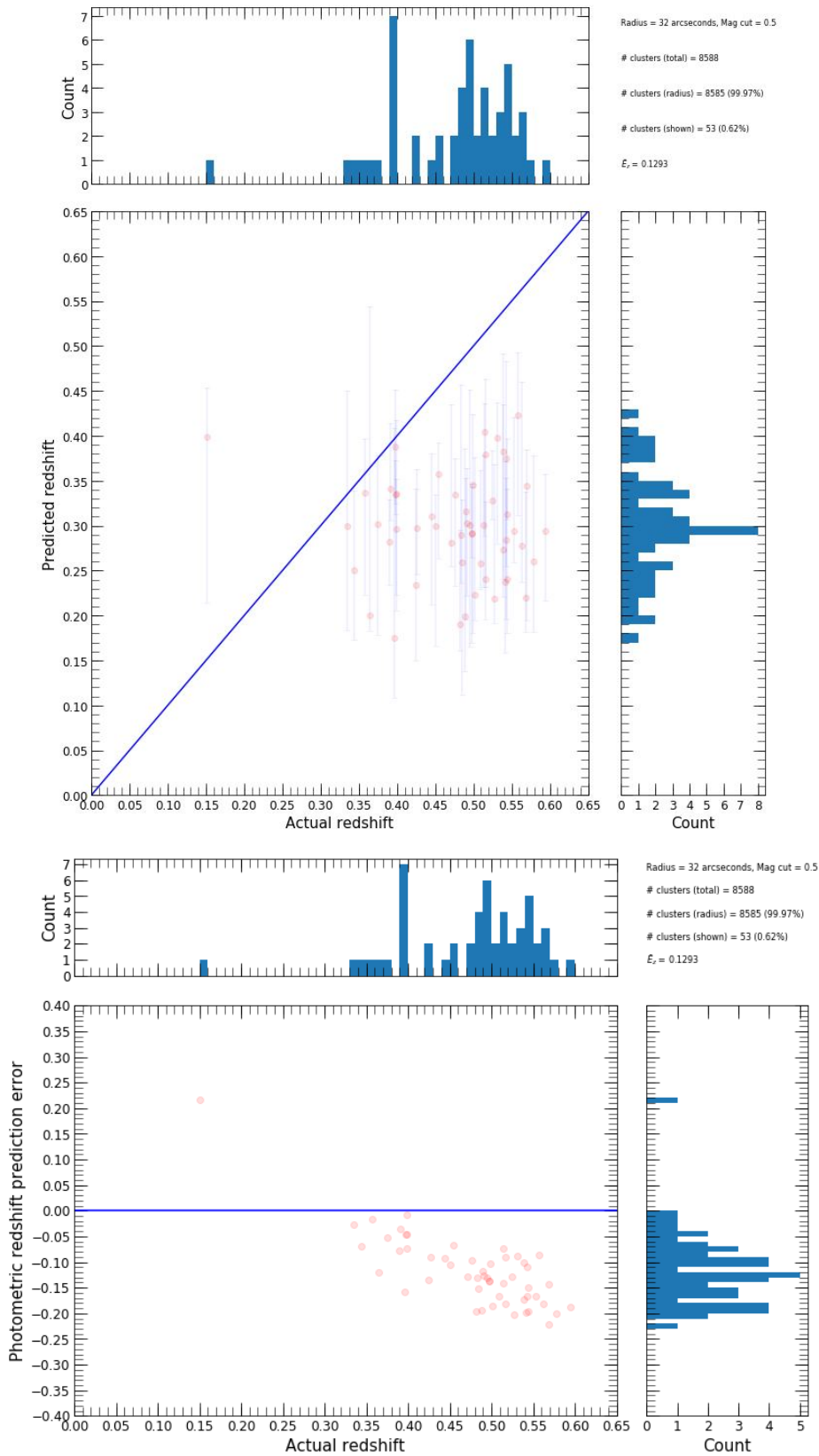


Figure S13. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the MWAR dataset, that had partial bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the MWAR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 32 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with partial bootstrap resamples returned.

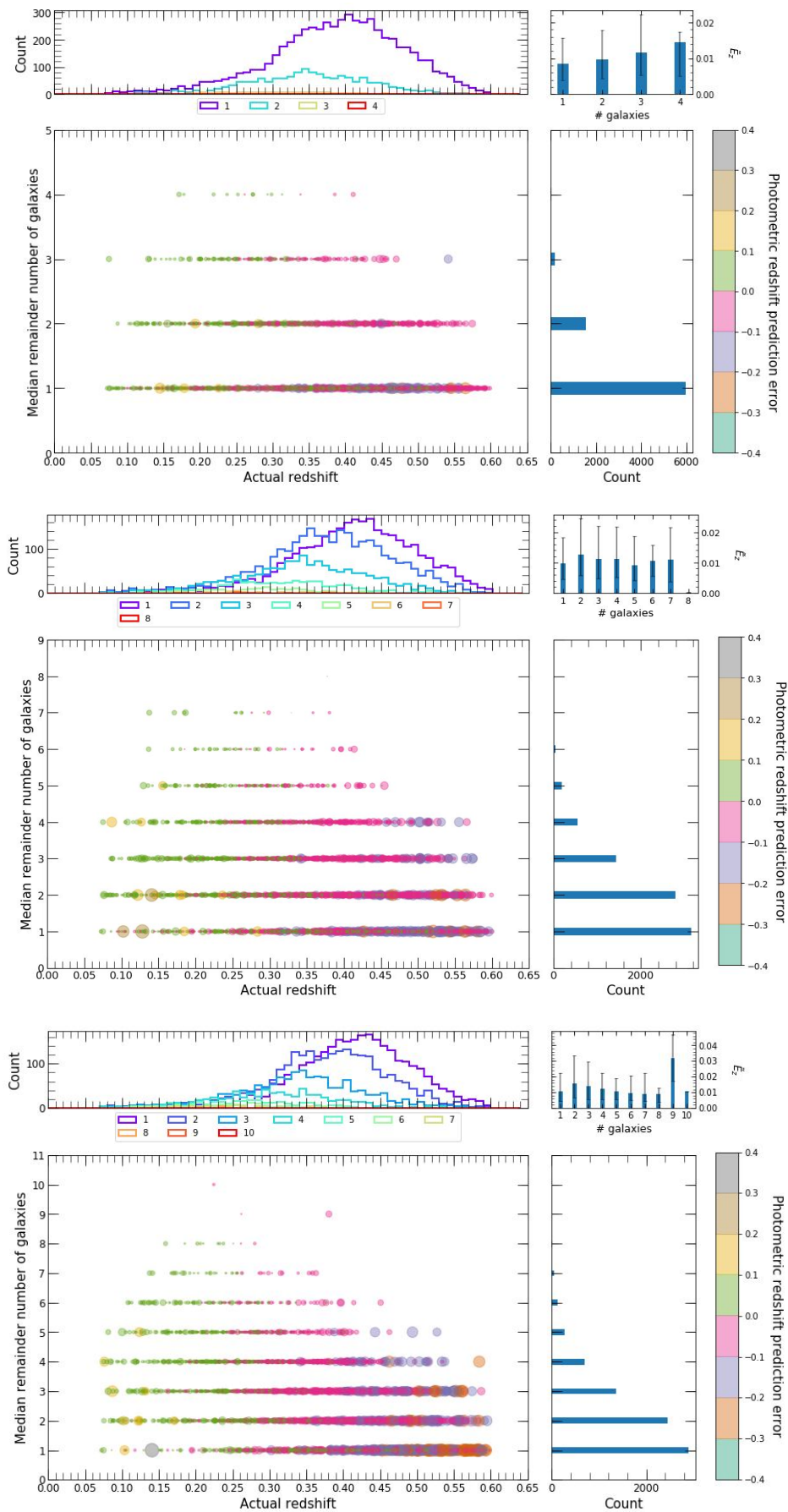


Figure S14. Plots displaying the number of galaxies used in photometric redshift predictions of clusters with low richness versus 'actual' redshift of tested clusters, which did not qualify for the MWAR dataset, where predictions had full bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

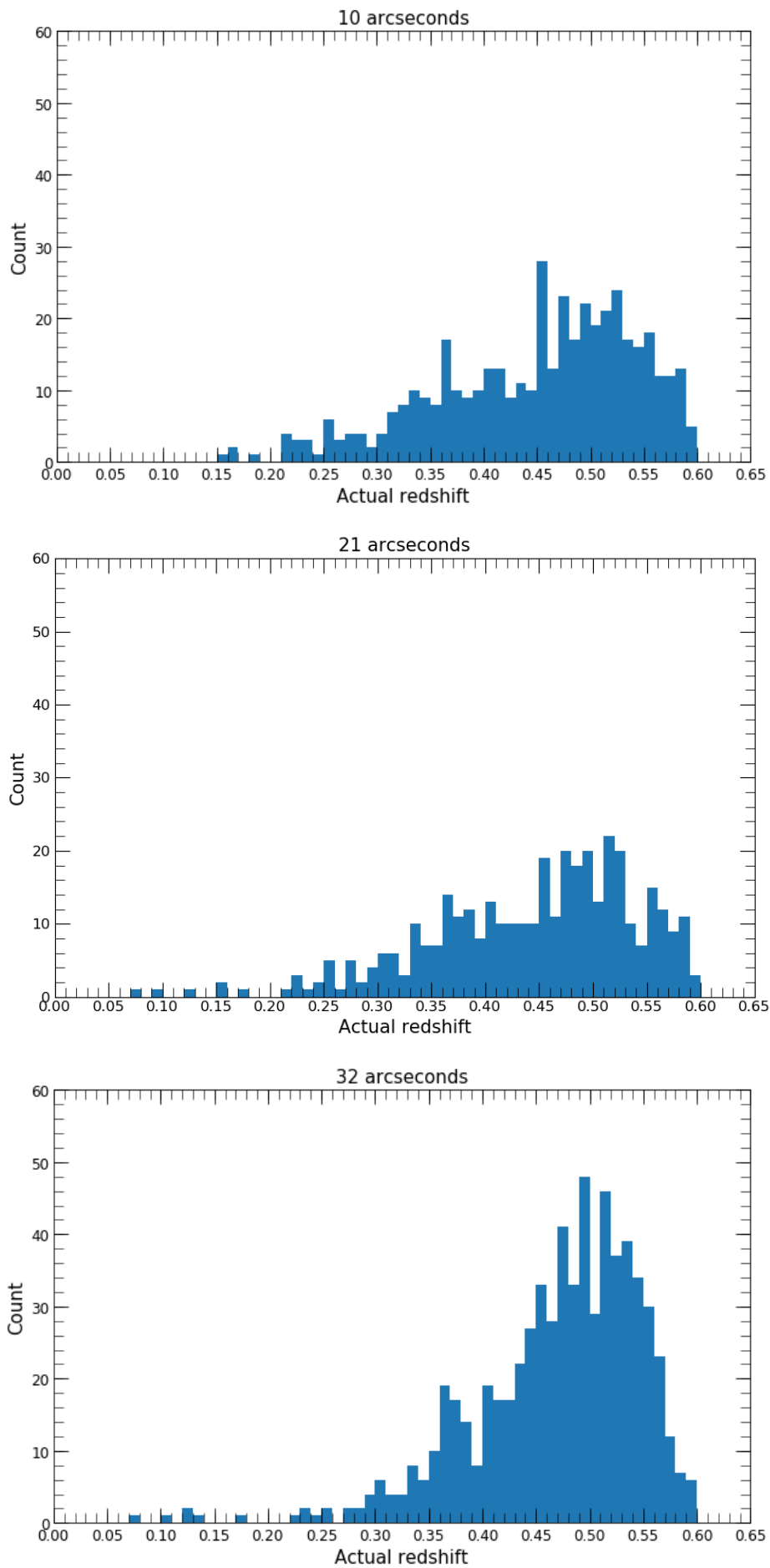


Figure S15. Frequency histograms displaying the ‘actual’ redshift distributions of clusters with low richness, which did not qualify for the MWAR dataset, that had no bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius.

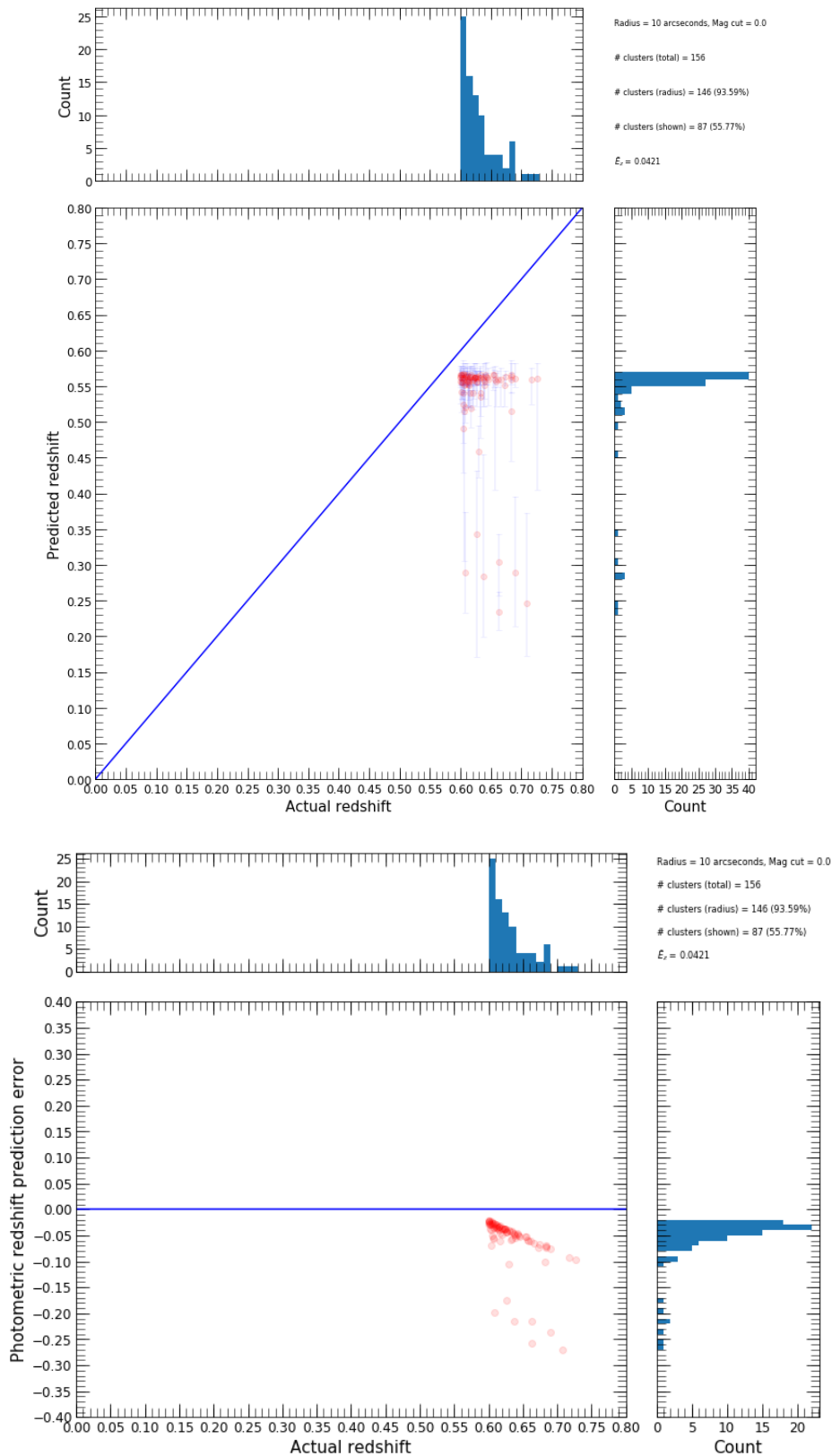


Figure S16. Plots displaying the performance of photometric redshift predictions of clusters at high redshift, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with full bootstrap resamples returned.

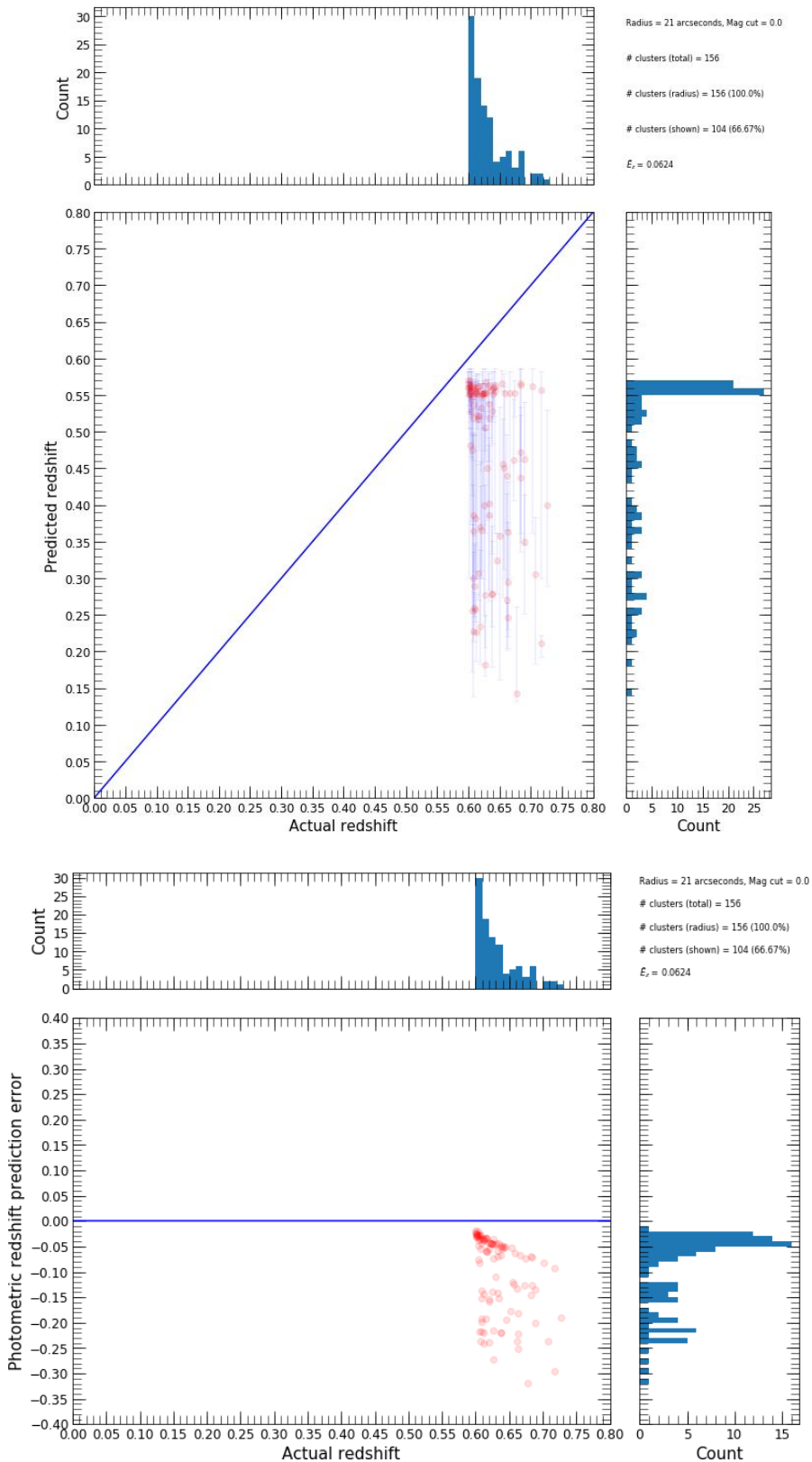


Figure S17. Plots displaying the performance of photometric redshift predictions of clusters at high redshift, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with full bootstrap resamples returned.

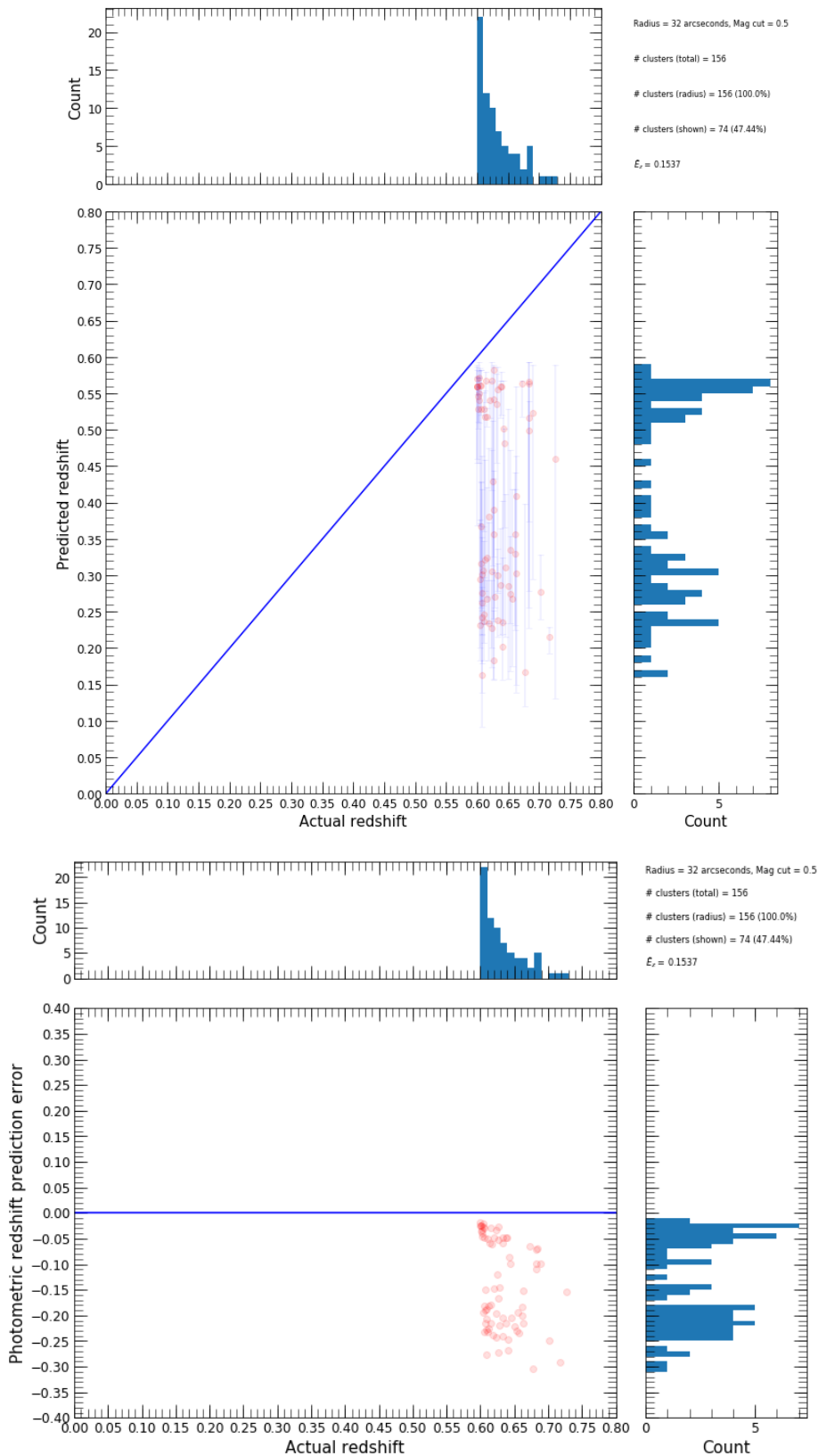


Figure S18. Plots displaying the performance of photometric redshift predictions of clusters at high redshift, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with full bootstrap resamples returned.

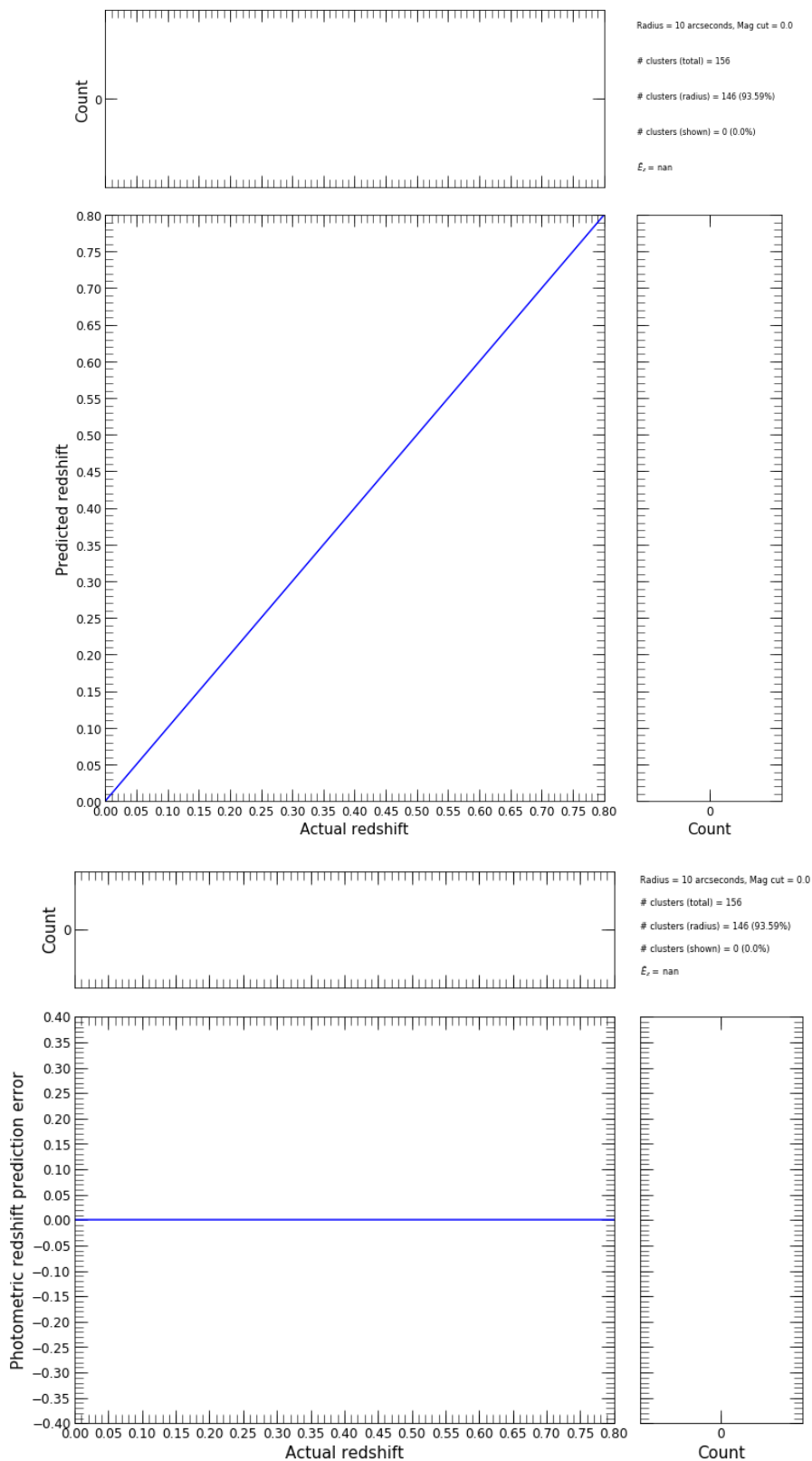


Figure S19. Plots displaying the performance of photometric redshift predictions of clusters at high redshift, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 10 arcseconds search radius. It should be noted that in this figure there were no resultant predictions made by the tuned model, as none of the clusters met the conditions. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius with partial bootstrap resamples returned, E_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

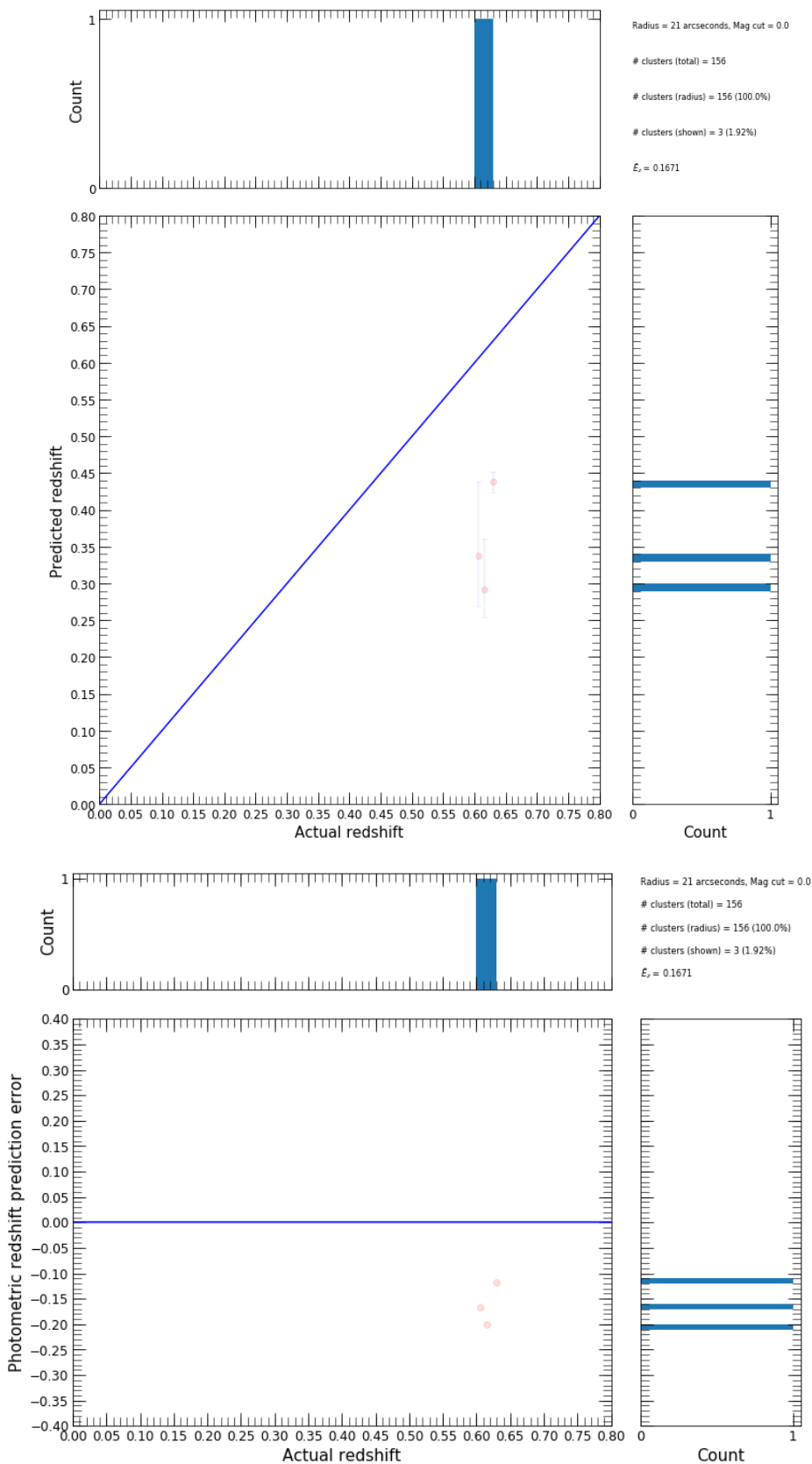


Figure S20. Plots displaying the performance of photometric redshift predictions of clusters at high redshift, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with partial bootstrap resamples returned.

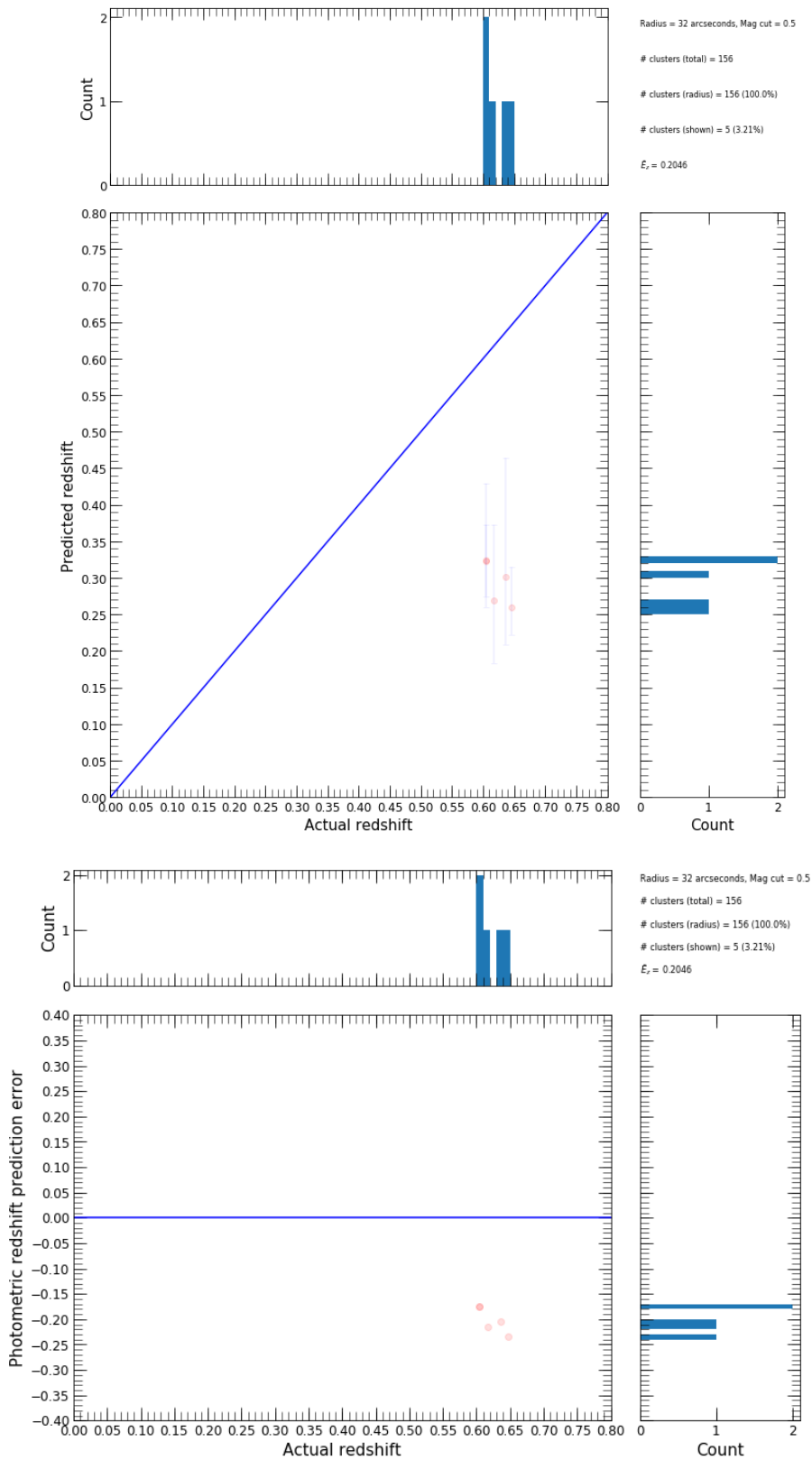


Figure S21. Plots displaying the performance of photometric redshift predictions of clusters at high redshift, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with partial bootstrap resamples returned.

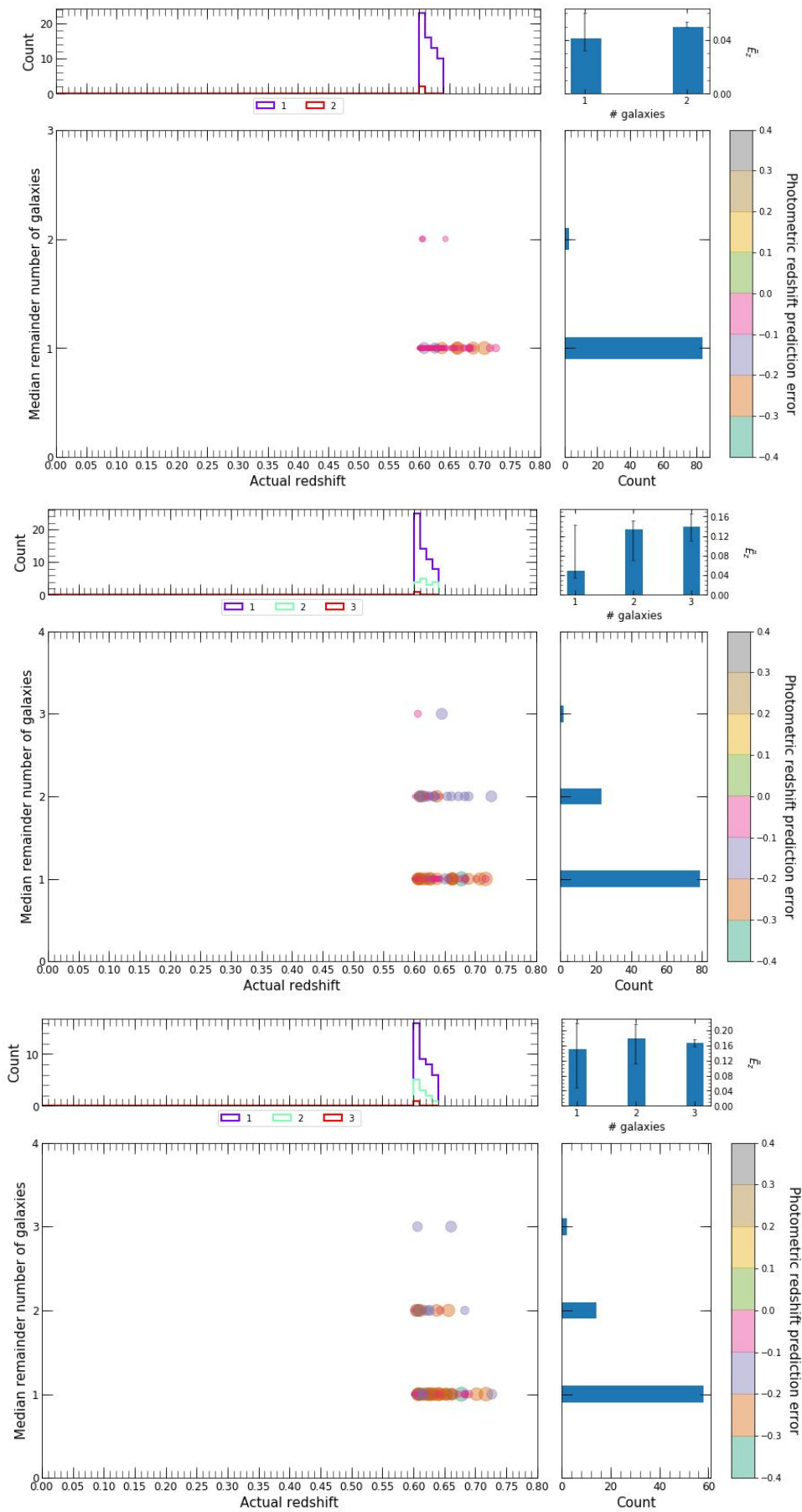


Figure S22. Plots displaying the number of galaxies used in photometric redshift predictions of clusters at high redshift versus 'actual' redshift of tested clusters, which did not qualify for the WNMR dataset, where predictions had full bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

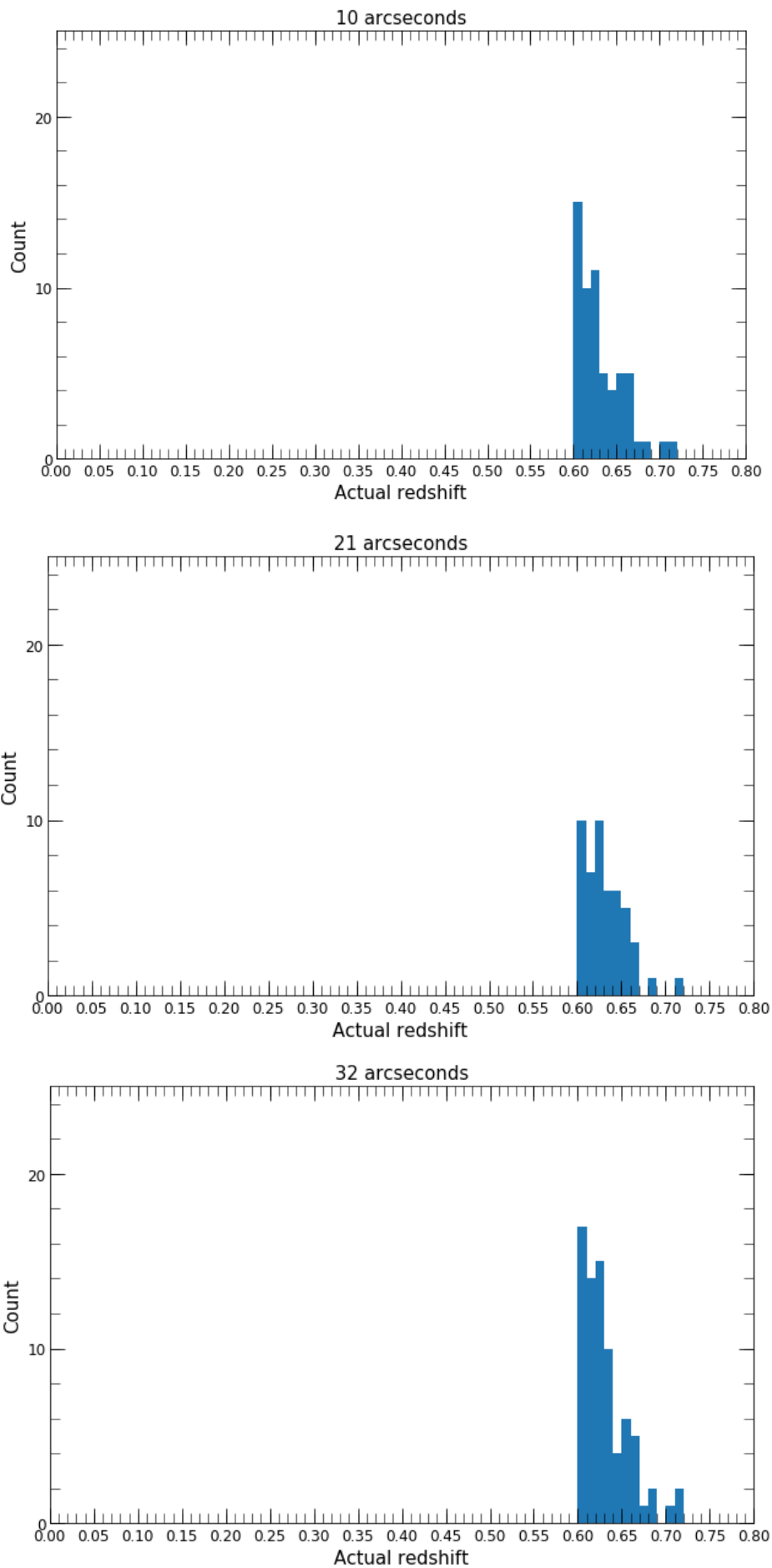


Figure S23. Frequency histograms displaying the ‘actual’ redshift distributions of clusters at high redshift, which did not qualify for the WNMR dataset, that had no bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius.

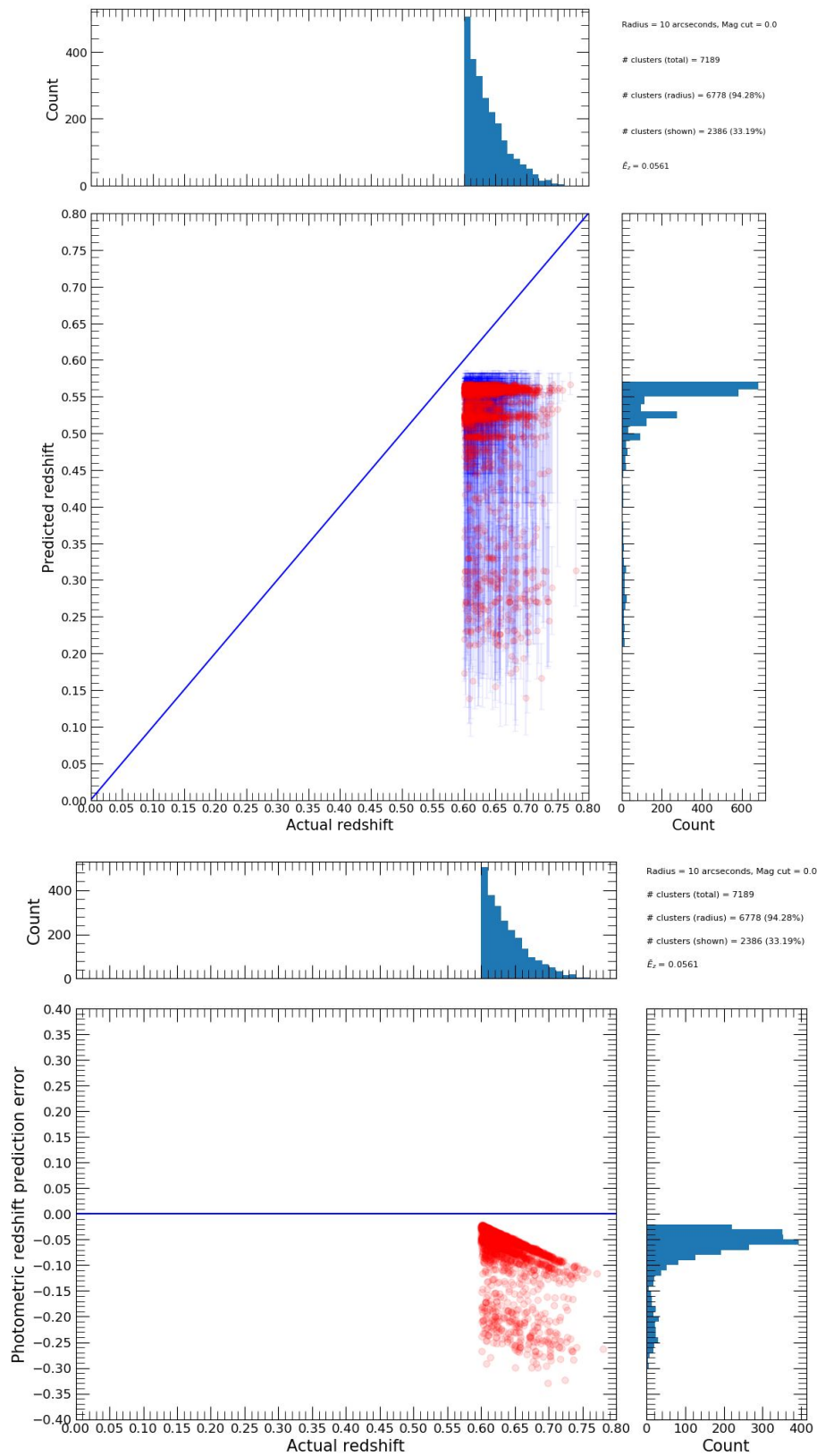


Figure S24. Plots displaying the performance of photometric redshift predictions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with full bootstrap resamples returned.

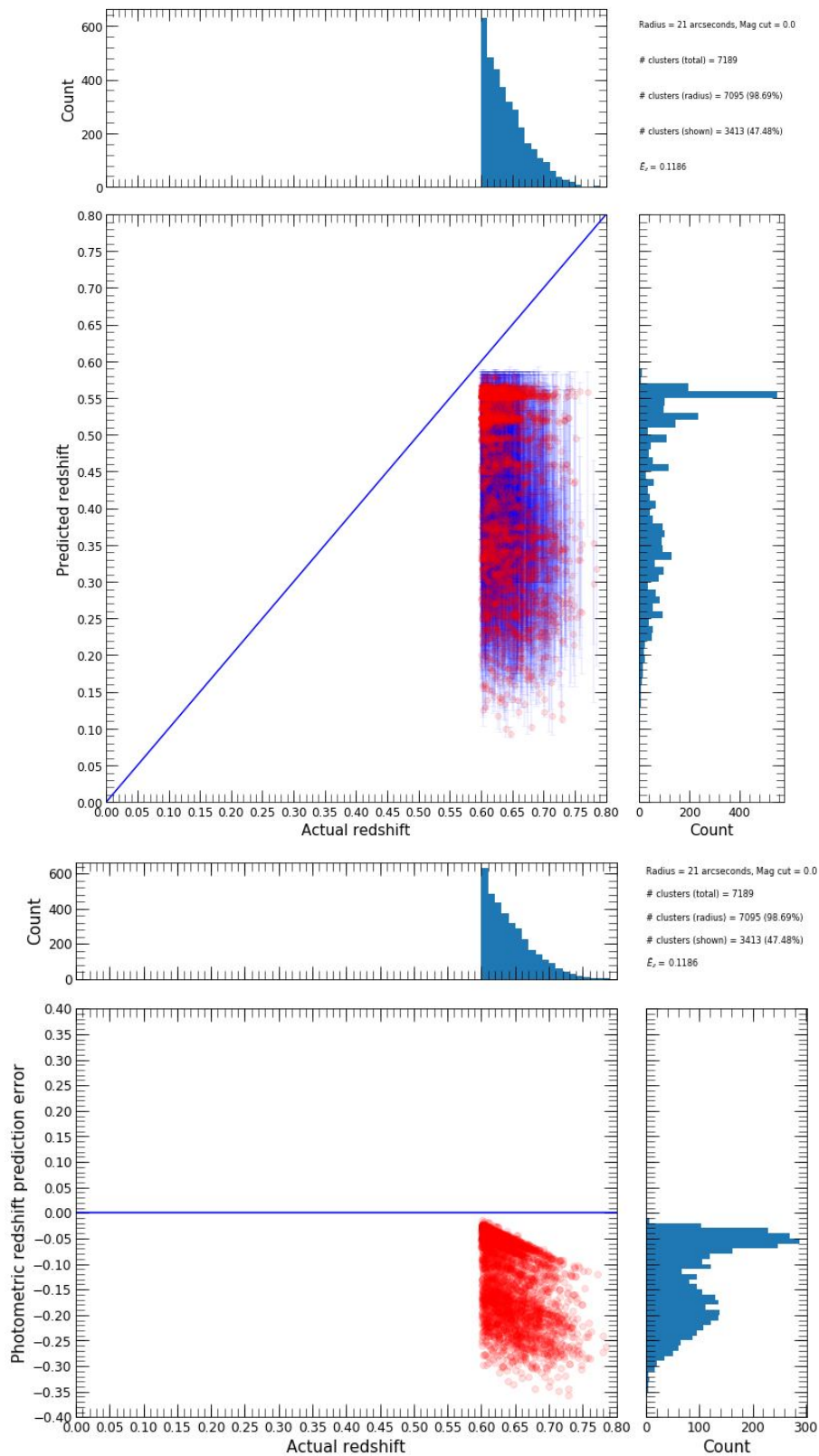


Figure S25. Plots displaying the performance of photometric redshift predictions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius with full bootstrap resamples returned, $\bar{\epsilon}_z$ represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with full bootstrap resamples returned.

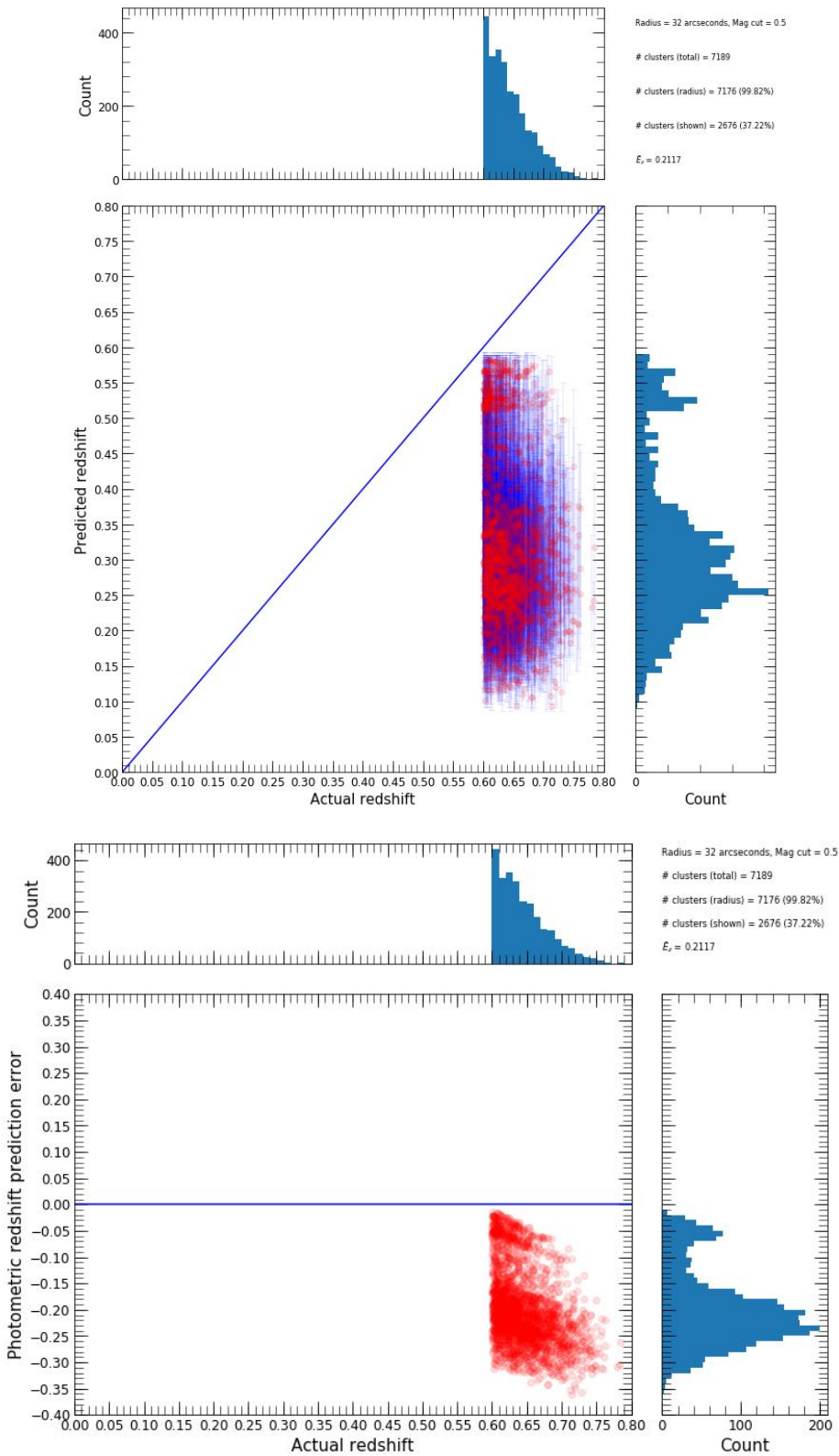


Figure S26. Plots displaying the performance of photometric redshift predictions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with full bootstrap resamples returned.

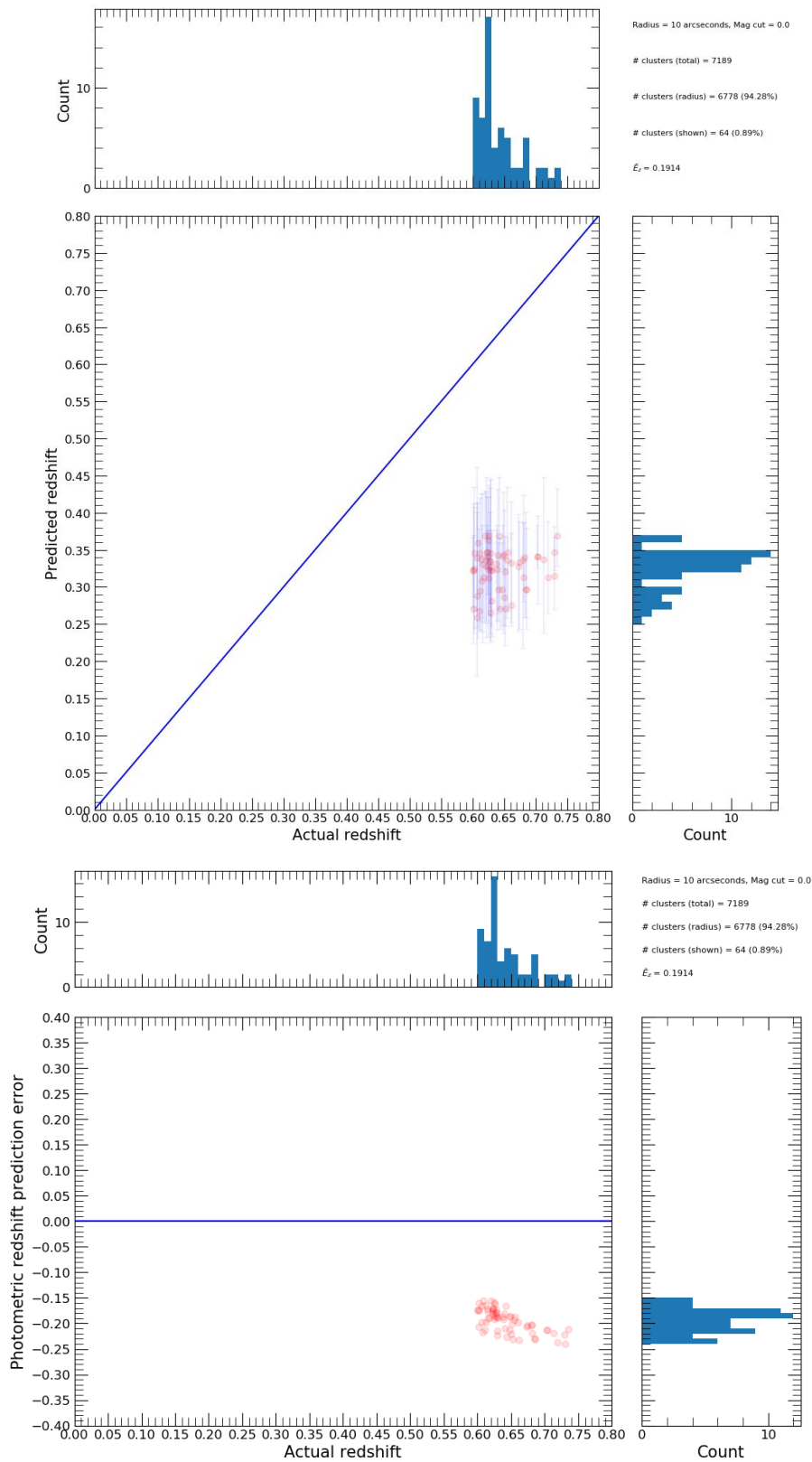


Figure S27. Plots displaying the performance of photometric redshift predictions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius with partial bootstrap resamples returned, \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

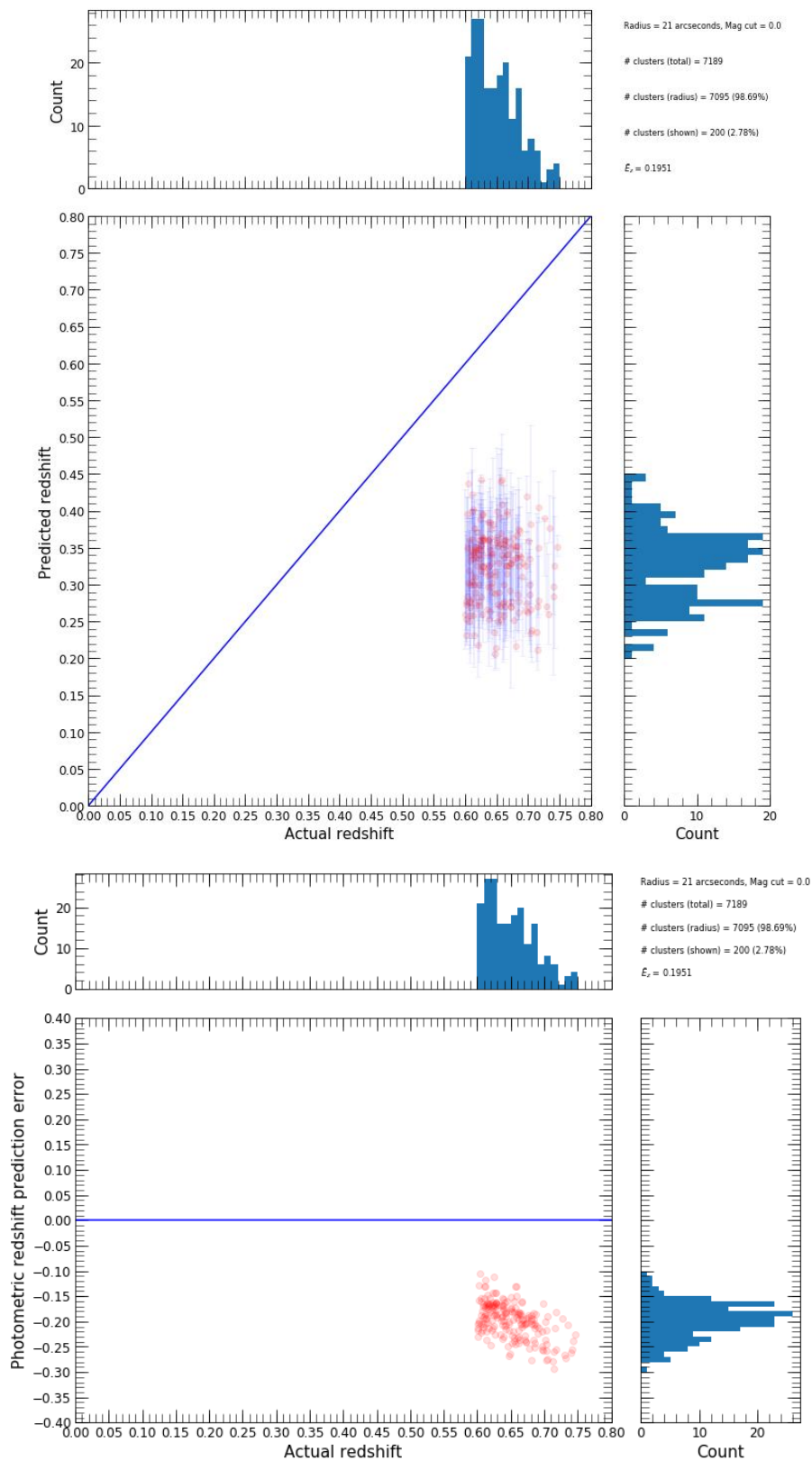


Figure S28. Plots displaying the performance of photometric redshift predictions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with partial bootstrap resamples returned.

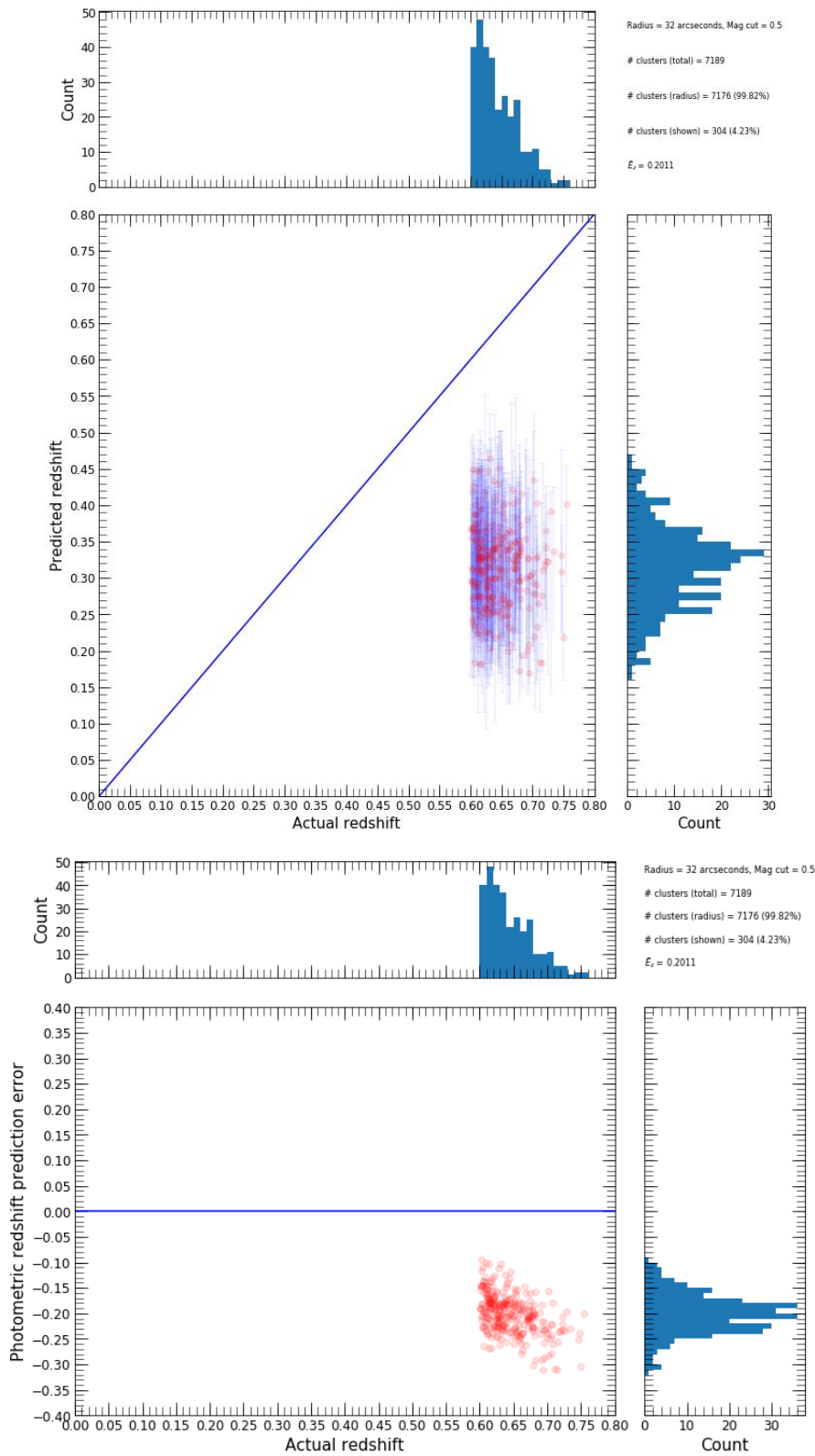


Figure S29. Plots displaying the performance of photometric redshift predictions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with partial bootstrap resamples returned.

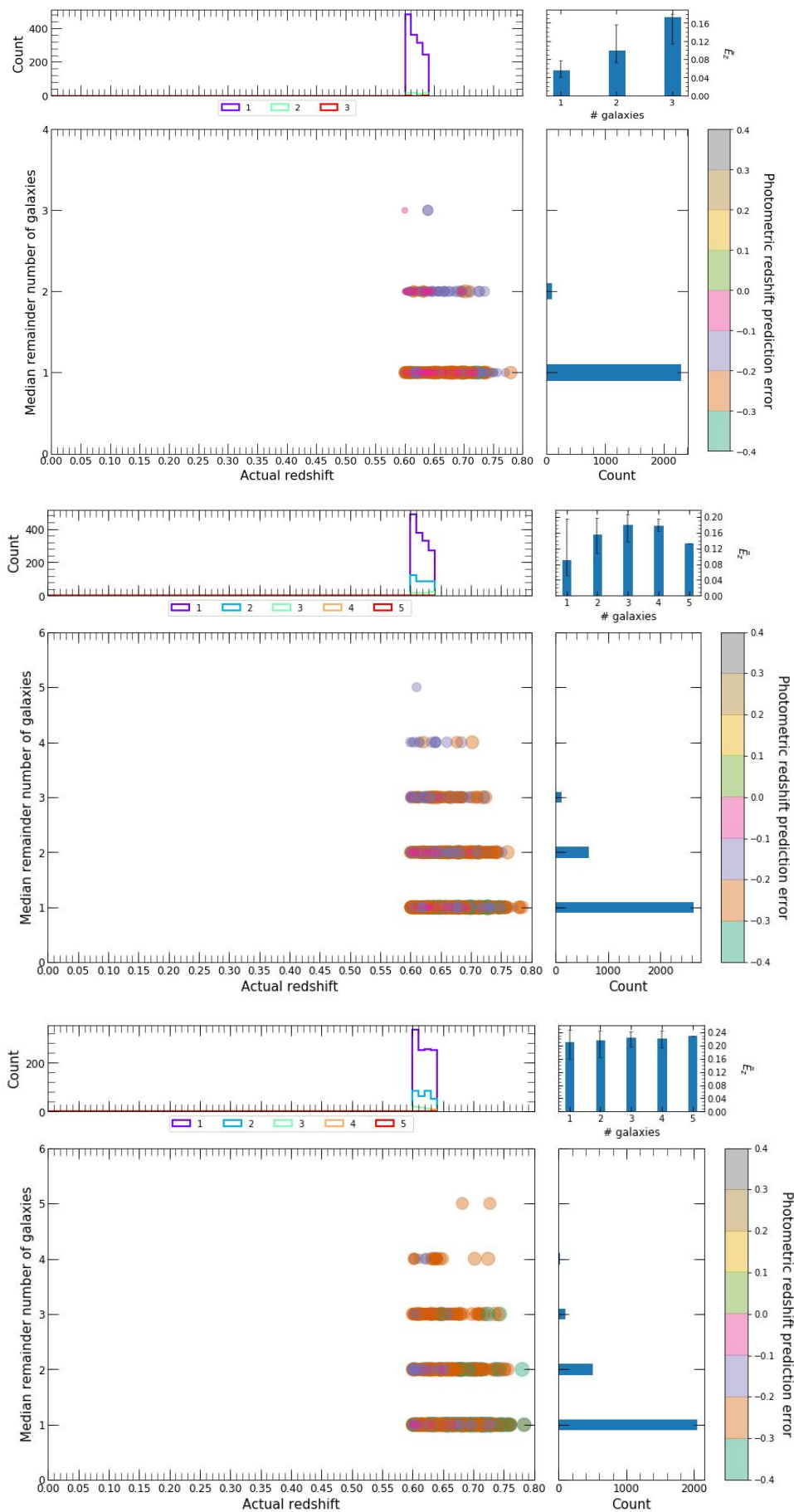


Figure S30. Plots displaying the number of galaxies used in photometric redshift predictions of clusters at high redshift with low richness versus 'actual' redshift of tested clusters, which did not qualify for the WNMR dataset, where predictions had full bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

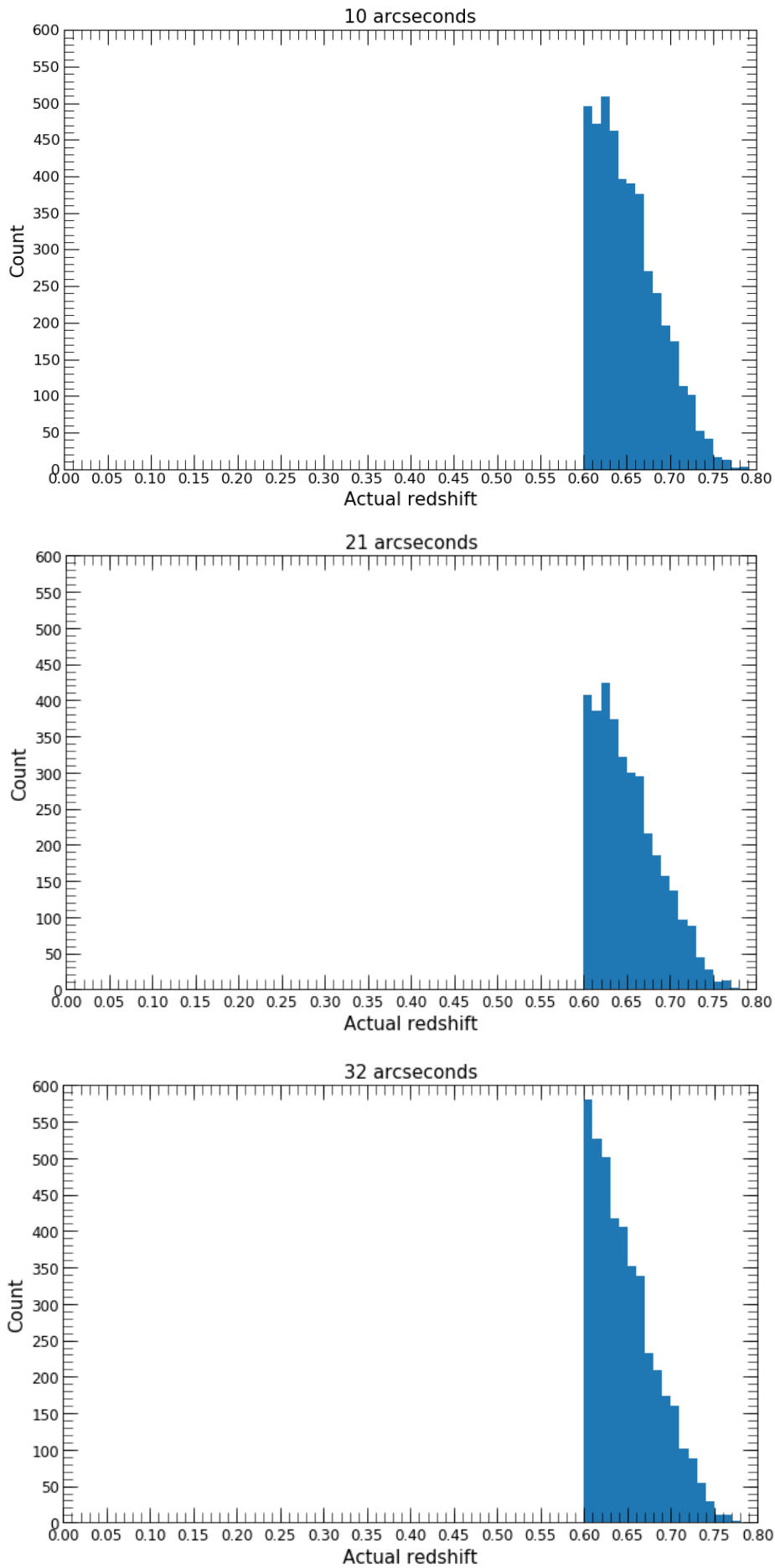


Figure S31. Frequency histograms displaying the 'actual' redshift distributions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had no bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius.

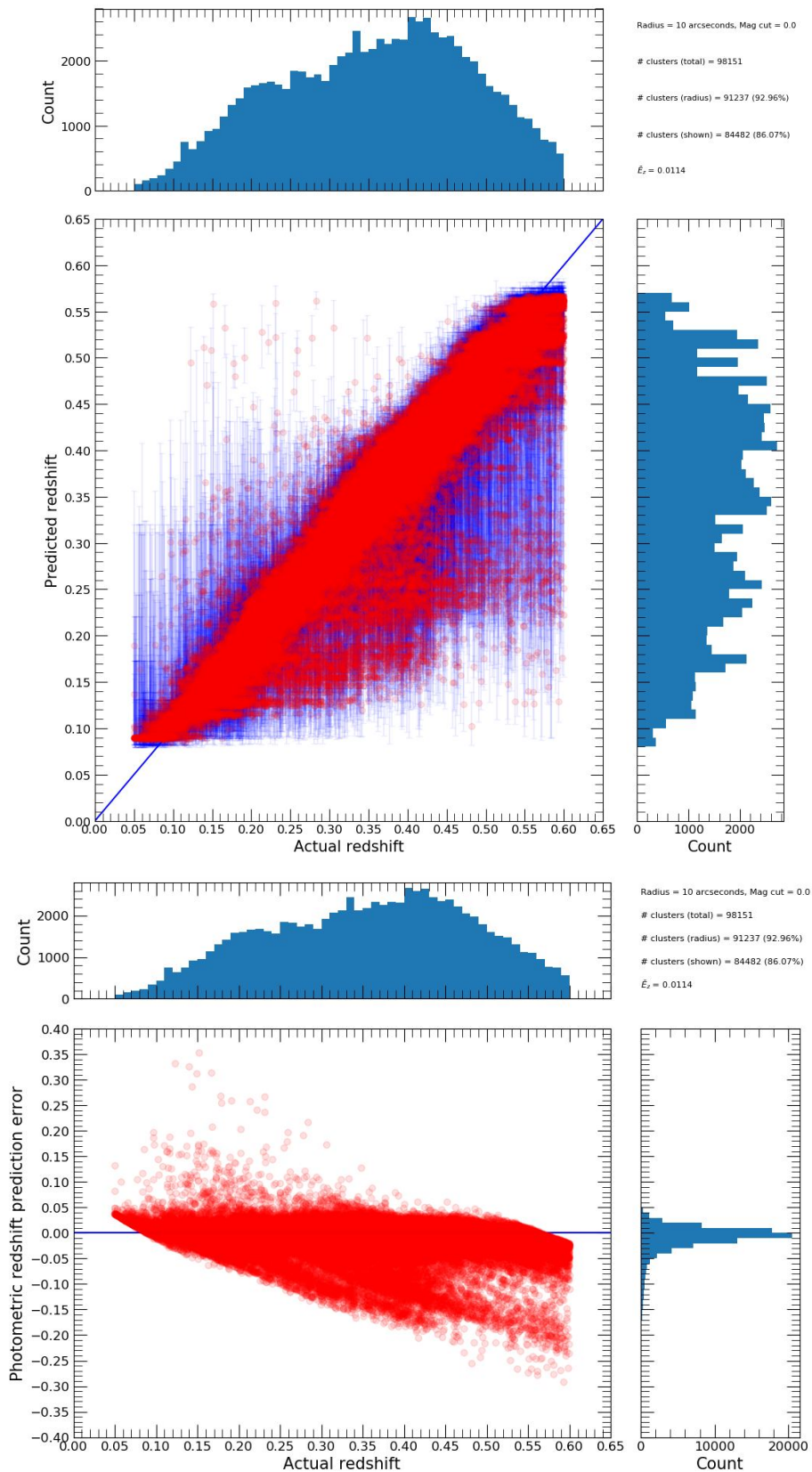


Figure S32. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius with full bootstrap resamples returned, \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with full bootstrap resamples returned.

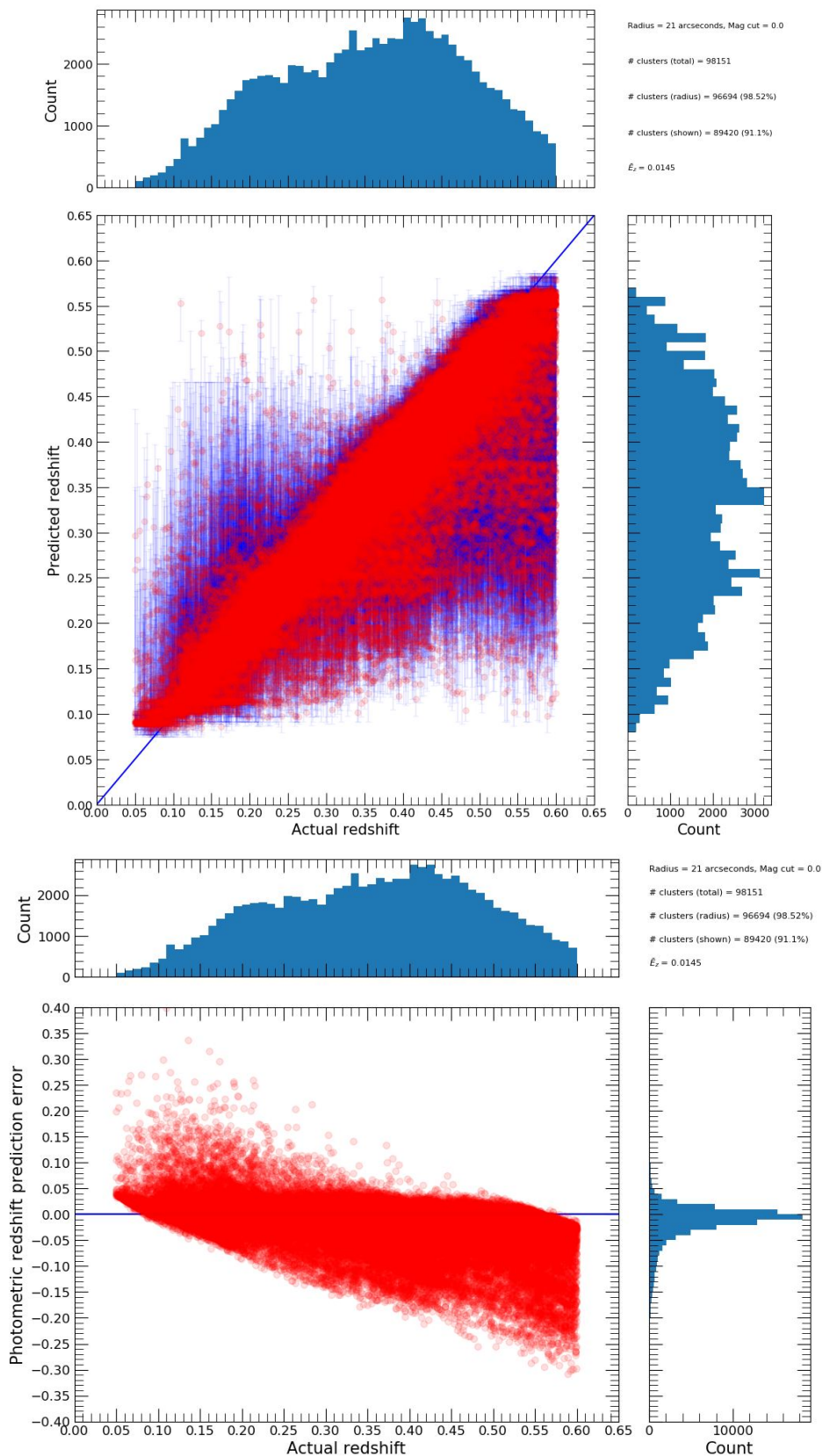


Figure S33. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with full bootstrap resamples returned.

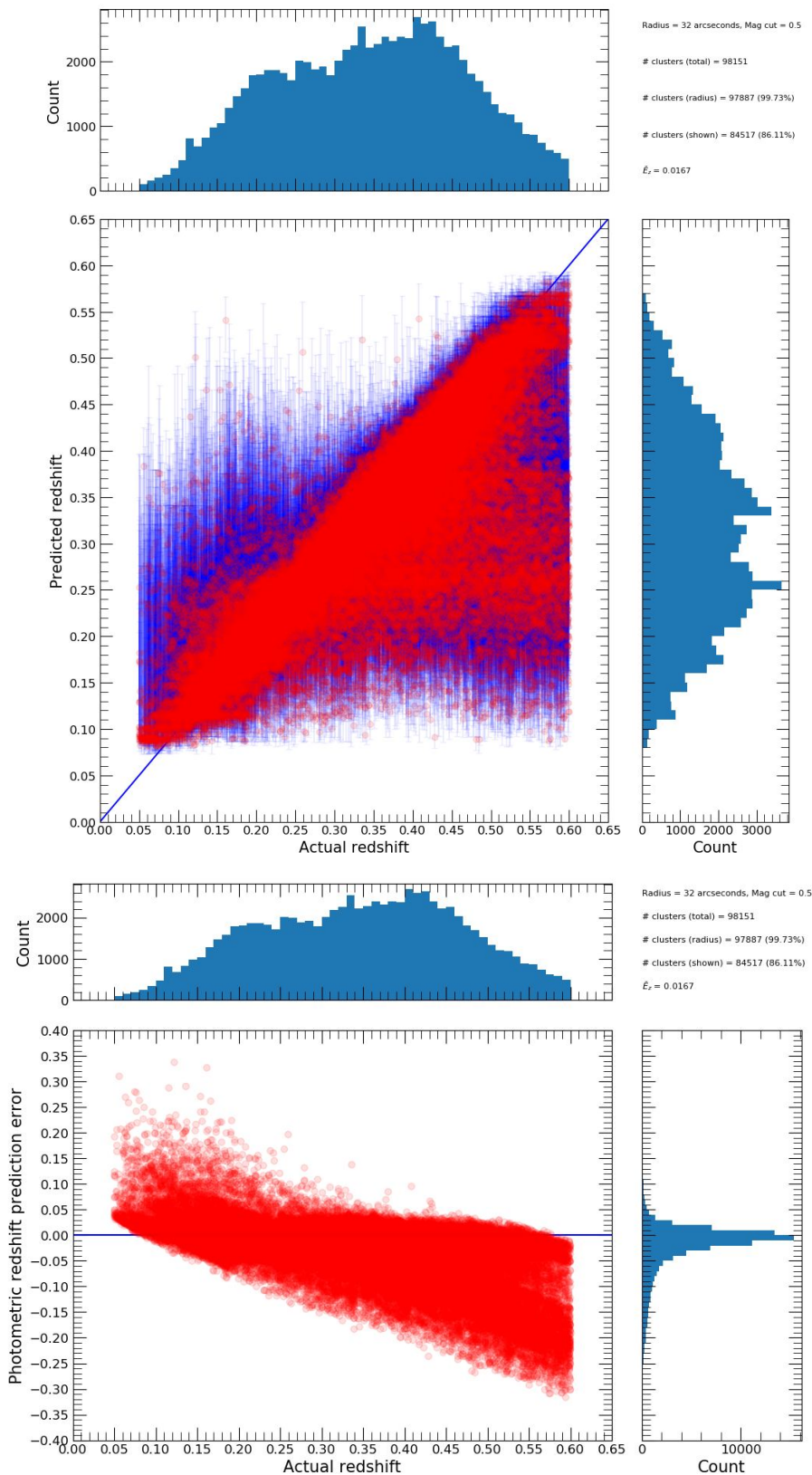


Figure S34. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius with full bootstrap resamples returned, \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with full bootstrap resamples returned.

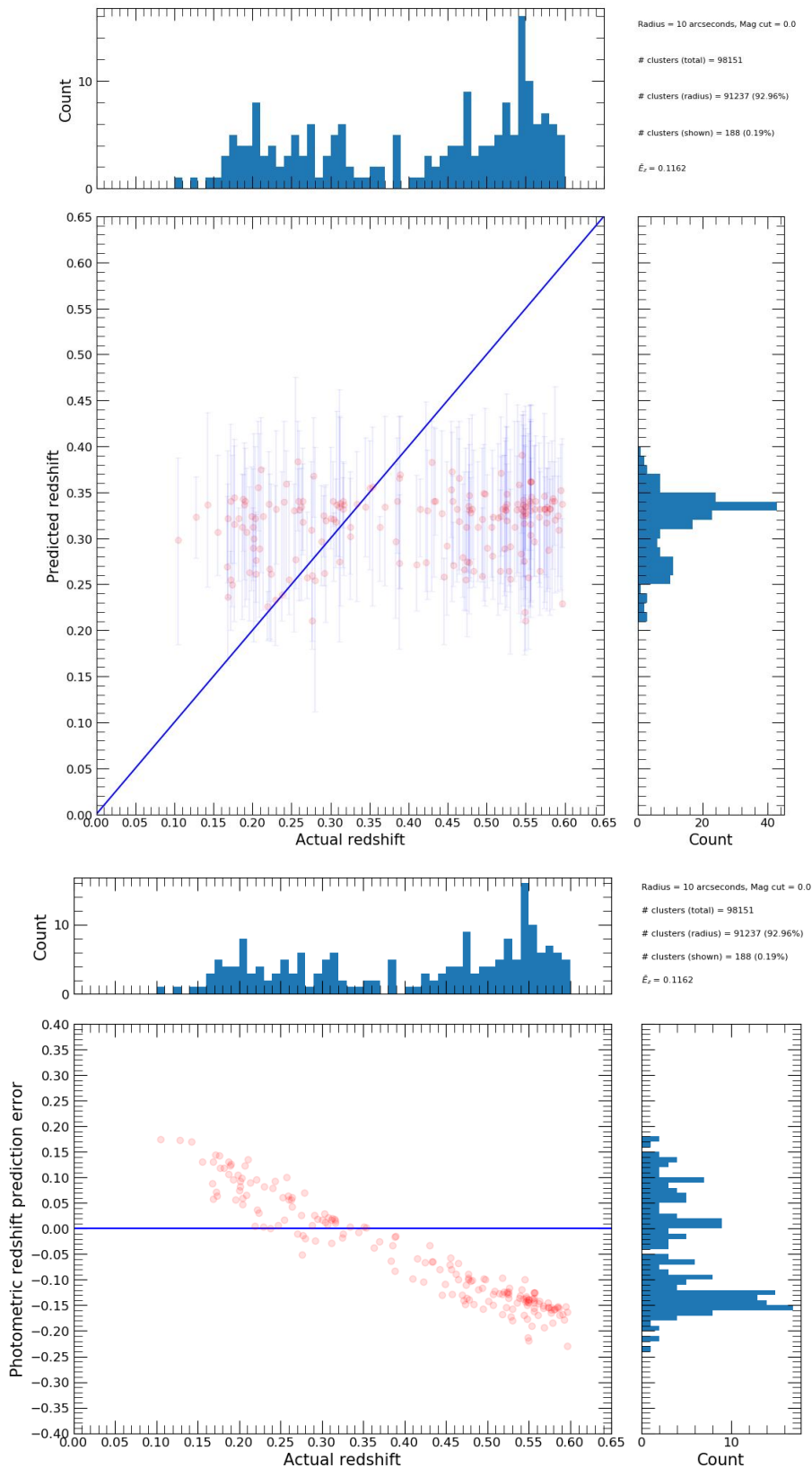


Figure S35. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10 arcseconds search radius with partial bootstrap resamples returned, \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

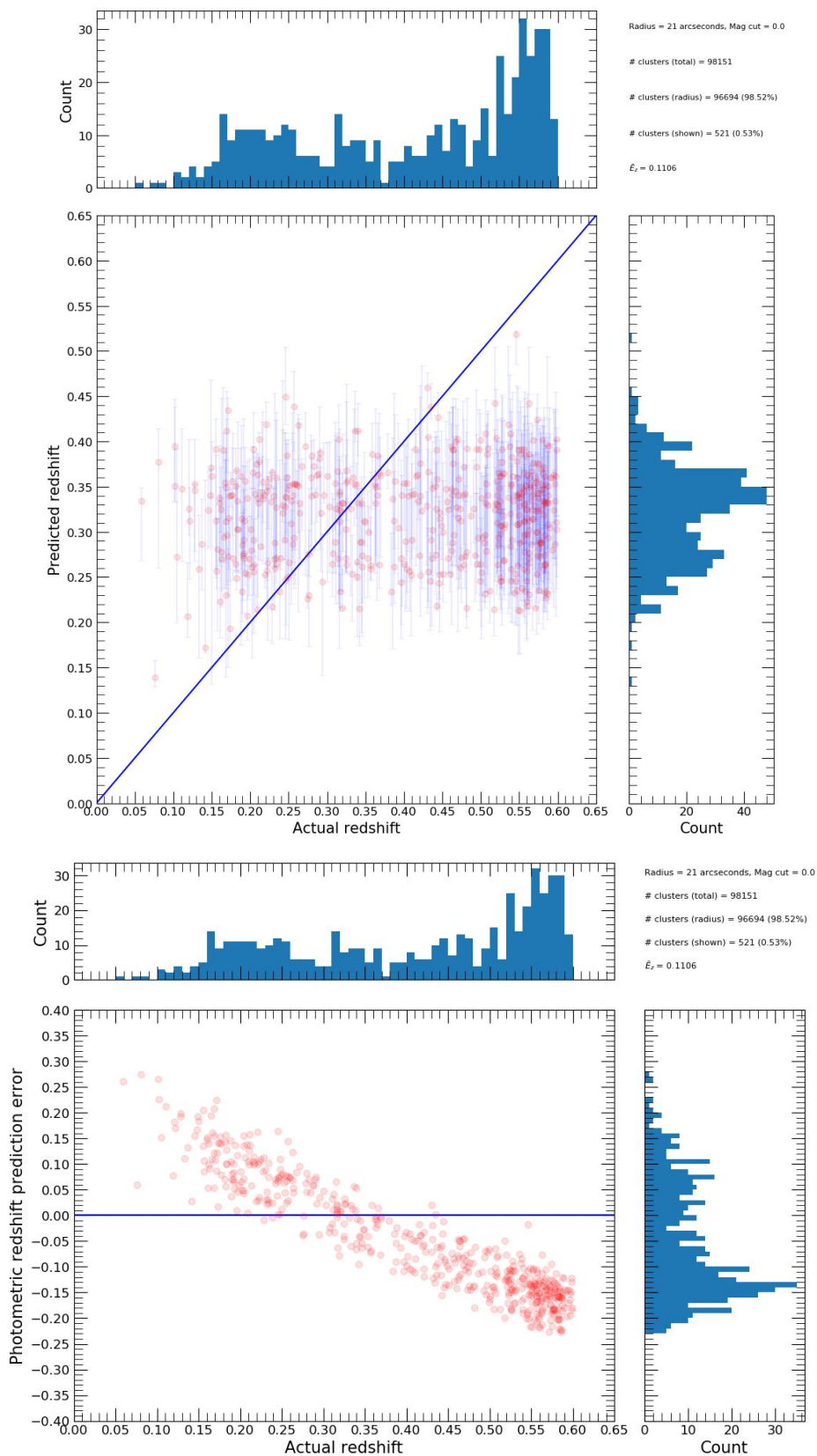


Figure S36. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 21 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 21 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 21 arcseconds search radius with partial bootstrap resamples returned.

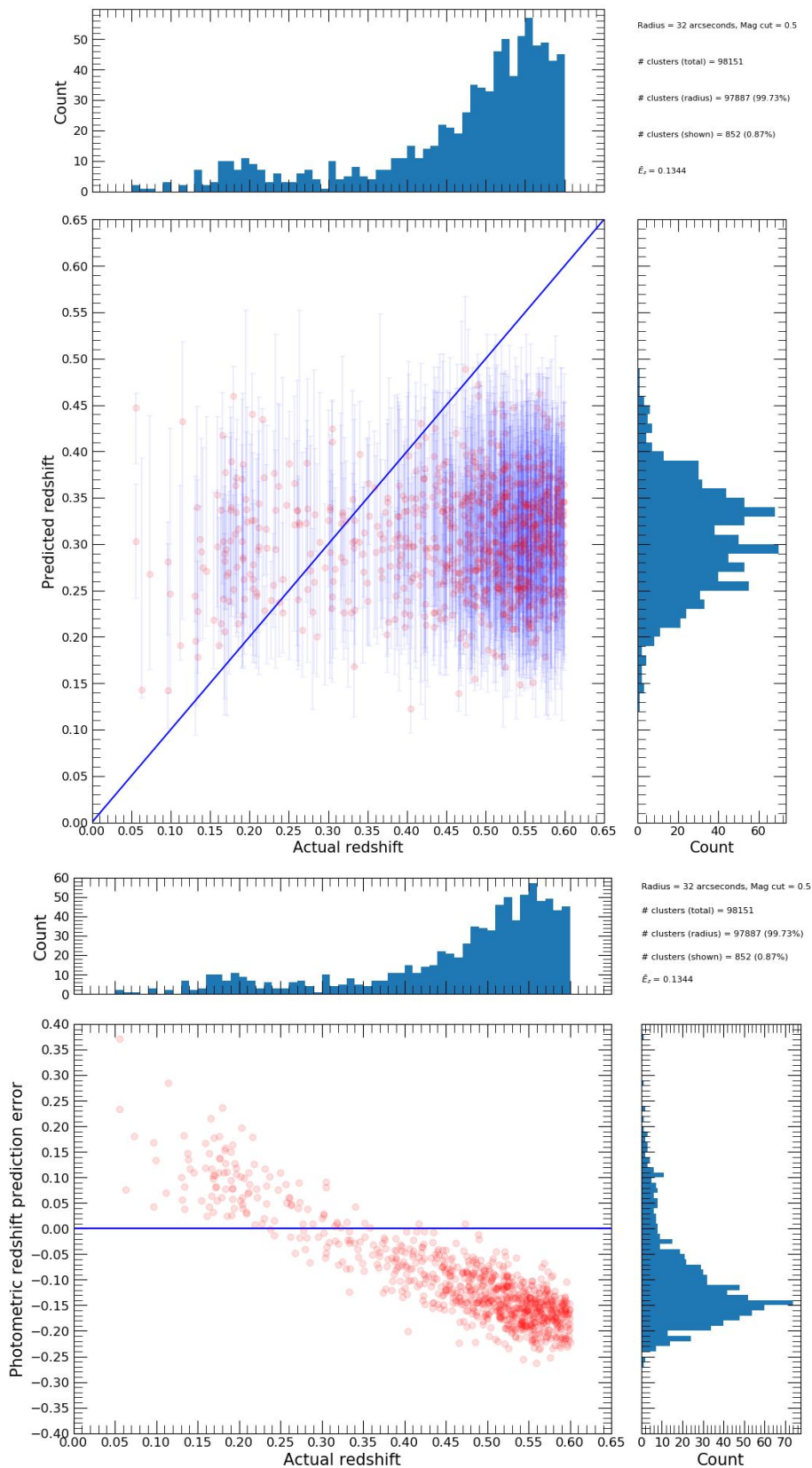


Figure S37. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the WNMR dataset, that had partial bootstrap resamples returned within a 32 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the WNMR dataset, '# clusters (radius)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius, '# clusters (shown)' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 32 arcseconds search radius with partial bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 32 arcseconds search radius with partial bootstrap resamples returned.

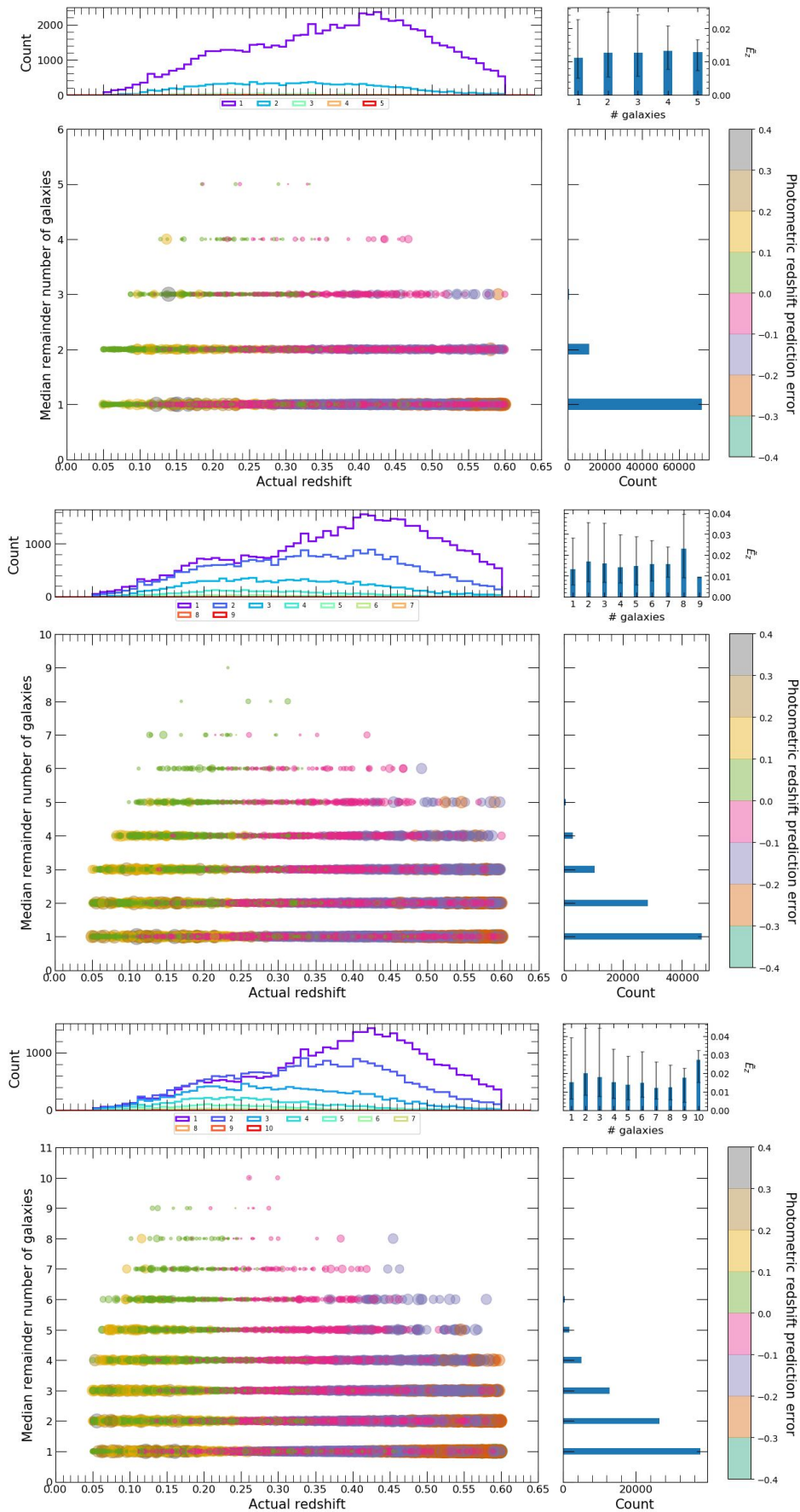


Figure S38. Plots displaying the number of galaxies used in photometric redshift predictions of clusters with low richness versus 'actual' redshift of tested clusters, which did not qualify for the WNMR dataset, where predictions had full bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

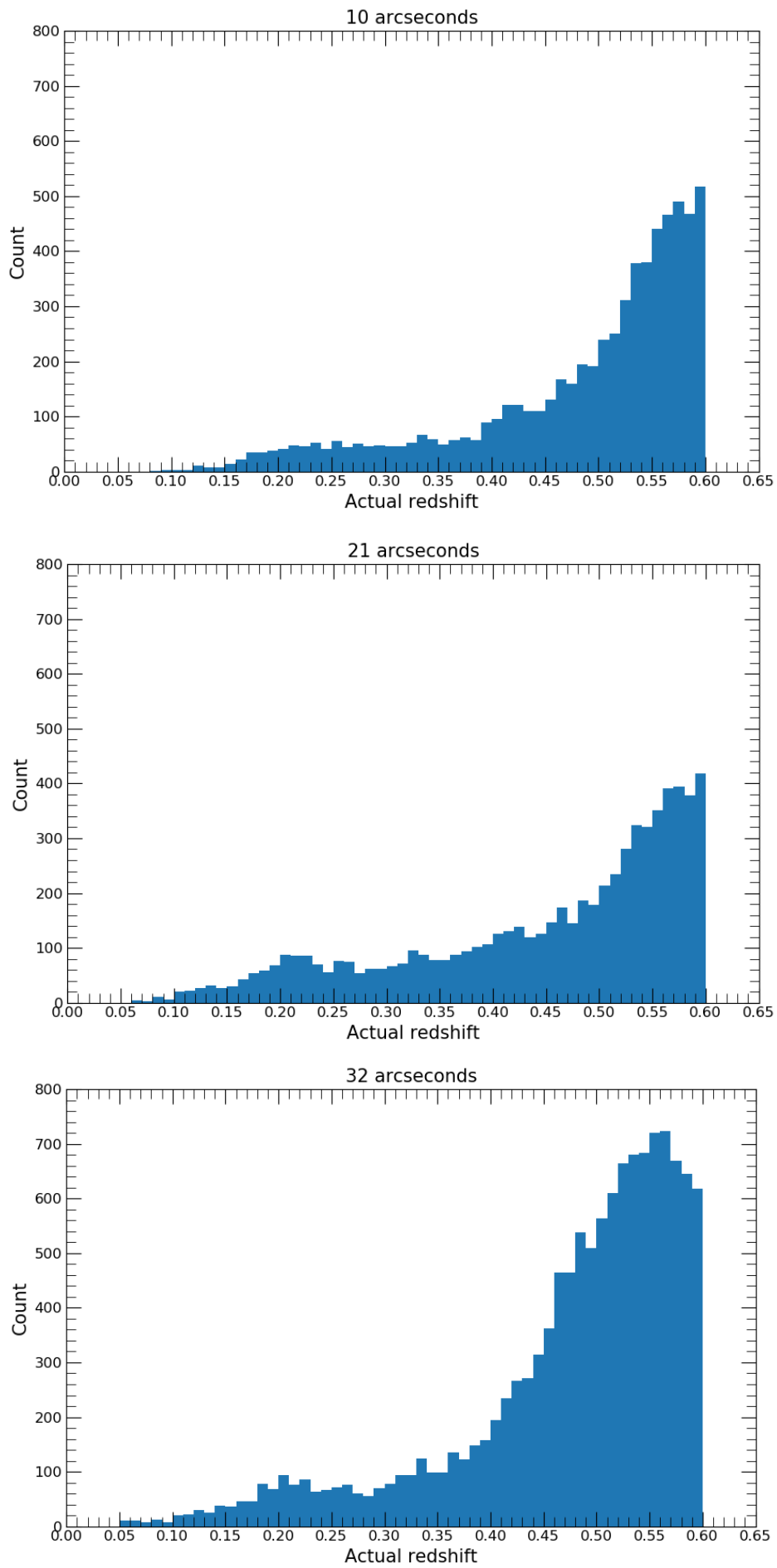


Figure S39. Frequency histograms displaying the 'actual' redshift distributions of clusters with low richness, which did not qualify for the WNMR dataset, that had no bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius.

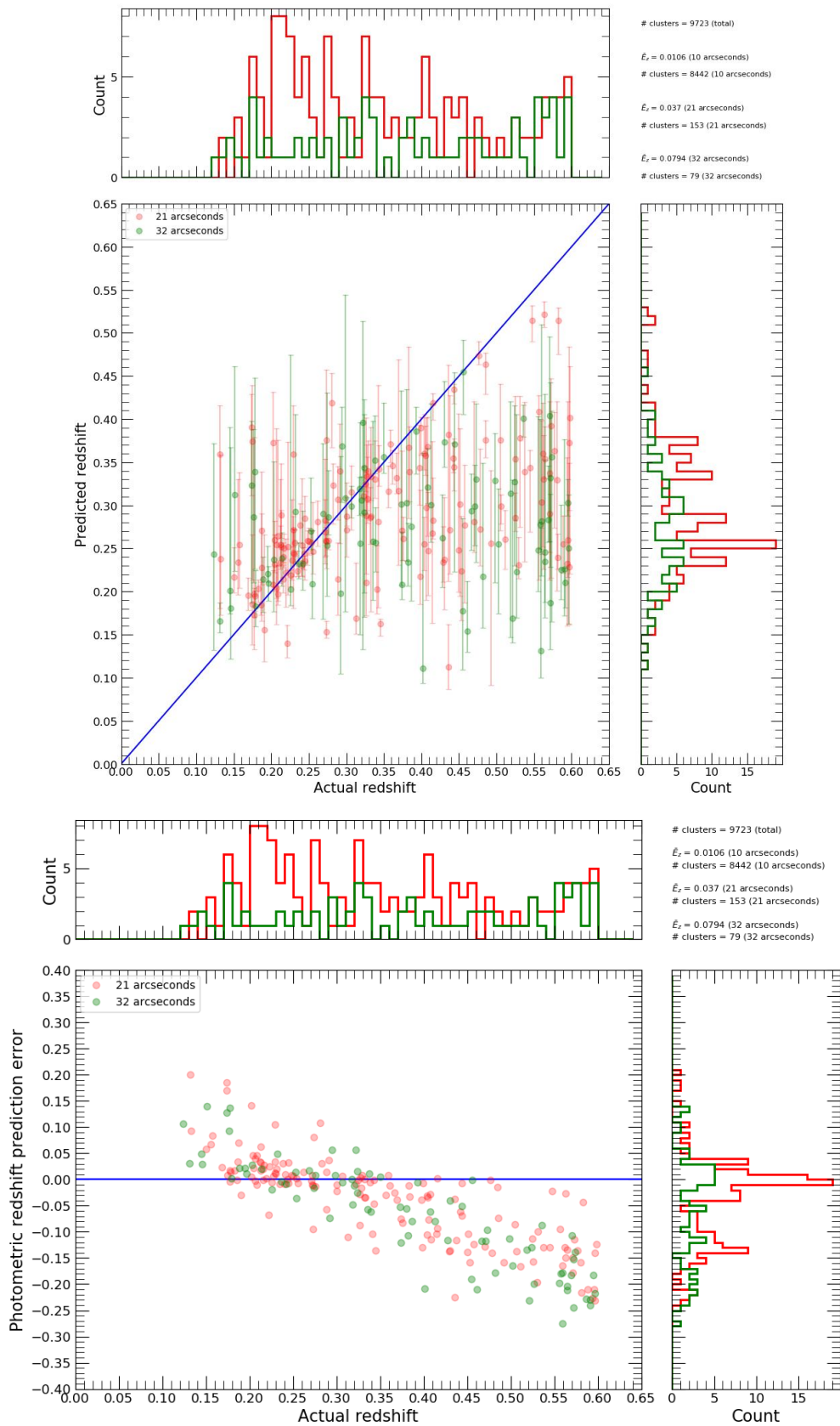


Figure S40. Plots displaying the performance of photometric redshift predictions of clusters for the WNMR test set that had full bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radius. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters for the WNMR test set, '# clusters' represents the number of clusters for the WNMR test set that have observed galaxies within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned.

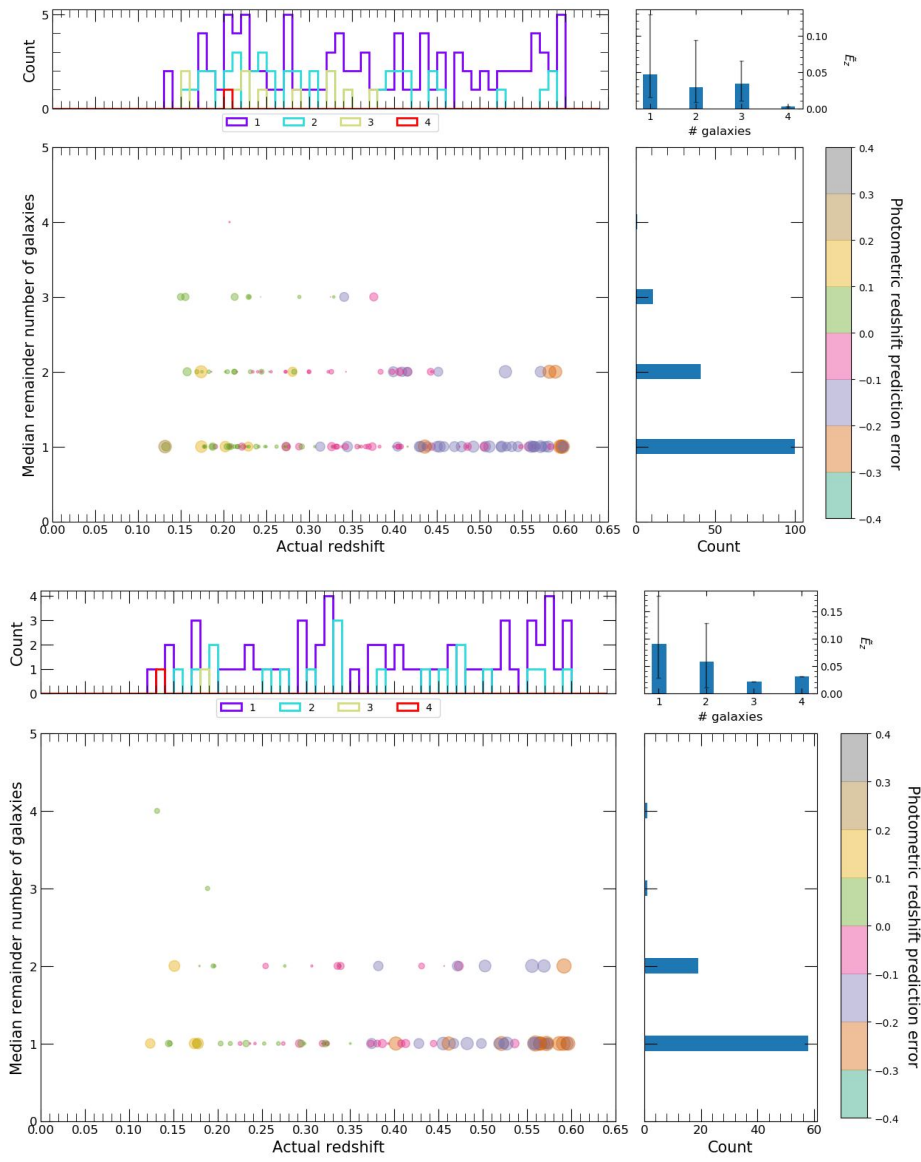


Figure S41. Plots displaying the number of galaxies used in photometric redshift predictions versus 'actual' redshift of tested clusters for the WNMR test set, where predictions had full bootstrap resamples returned within a 21 (top row) or 32 (bottom row) arcseconds search radius. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

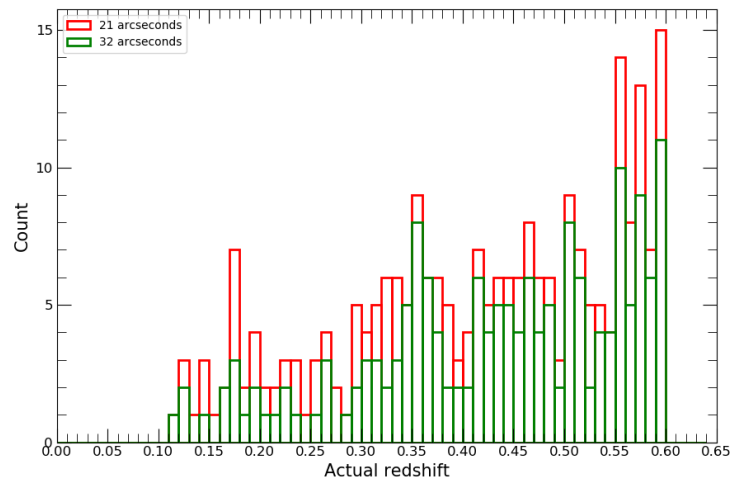


Figure S42. Frequency histograms displaying the 'actual' redshift distributions of the remaining clusters from the WNMR test set that had no bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radius.

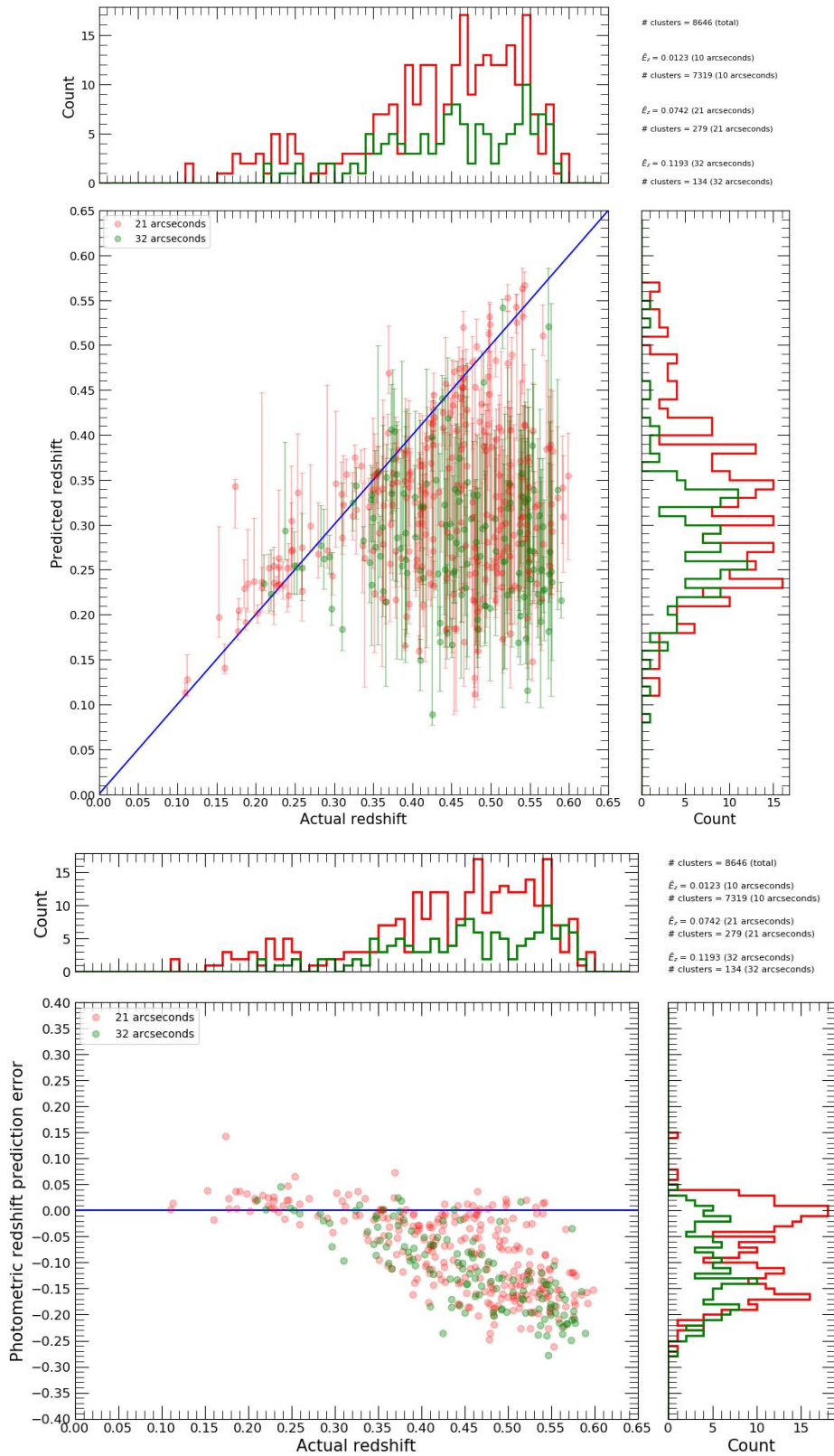


Figure S43. Plots displaying the performance of photometric redshift predictions of clusters for the RNMW test set that had full bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radii. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters for the RNMW test set, '# clusters' represents the number of clusters for the RNMW test set that have observed galaxies within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned.

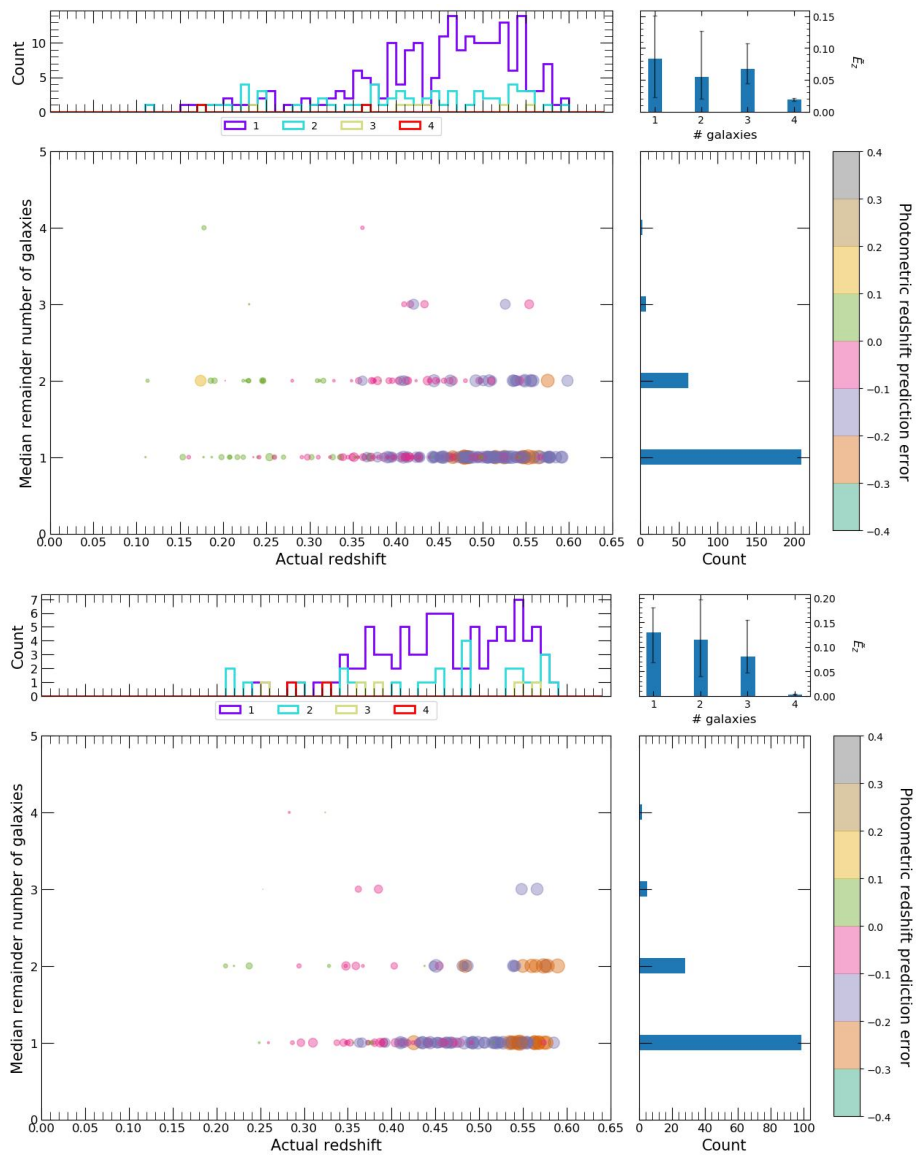


Figure S44. Plots displaying the number of galaxies used in photometric redshift predictions versus ‘actual’ redshift of tested clusters for the RNMW test set, where predictions had full bootstrap resamples returned within a 21 (top row) or 32 (bottom row) arcseconds search radius. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

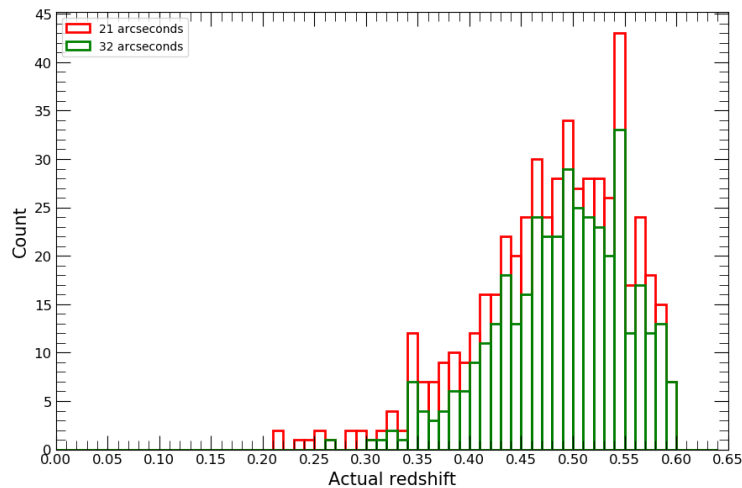


Figure S45. Frequency histograms displaying the ‘actual’ redshift distributions of the remaining clusters from the RNMW test set that had no bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radius.

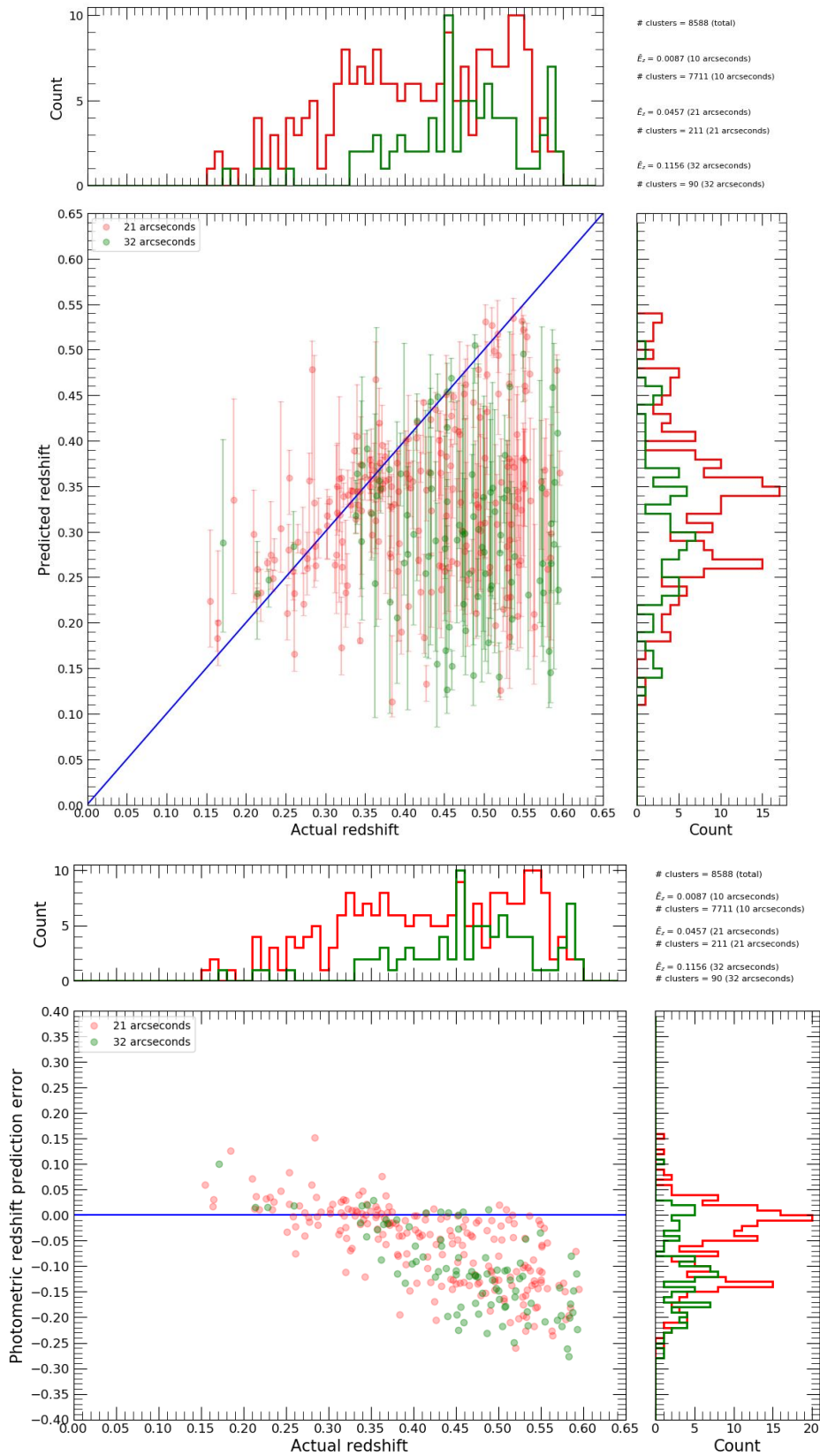


Figure S46. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the MWAR dataset, that had full bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radii. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the MWAR dataset, '# clusters' represents the number of clusters with low richness which did not qualify for the MWAR dataset that have observed galaxies within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned.

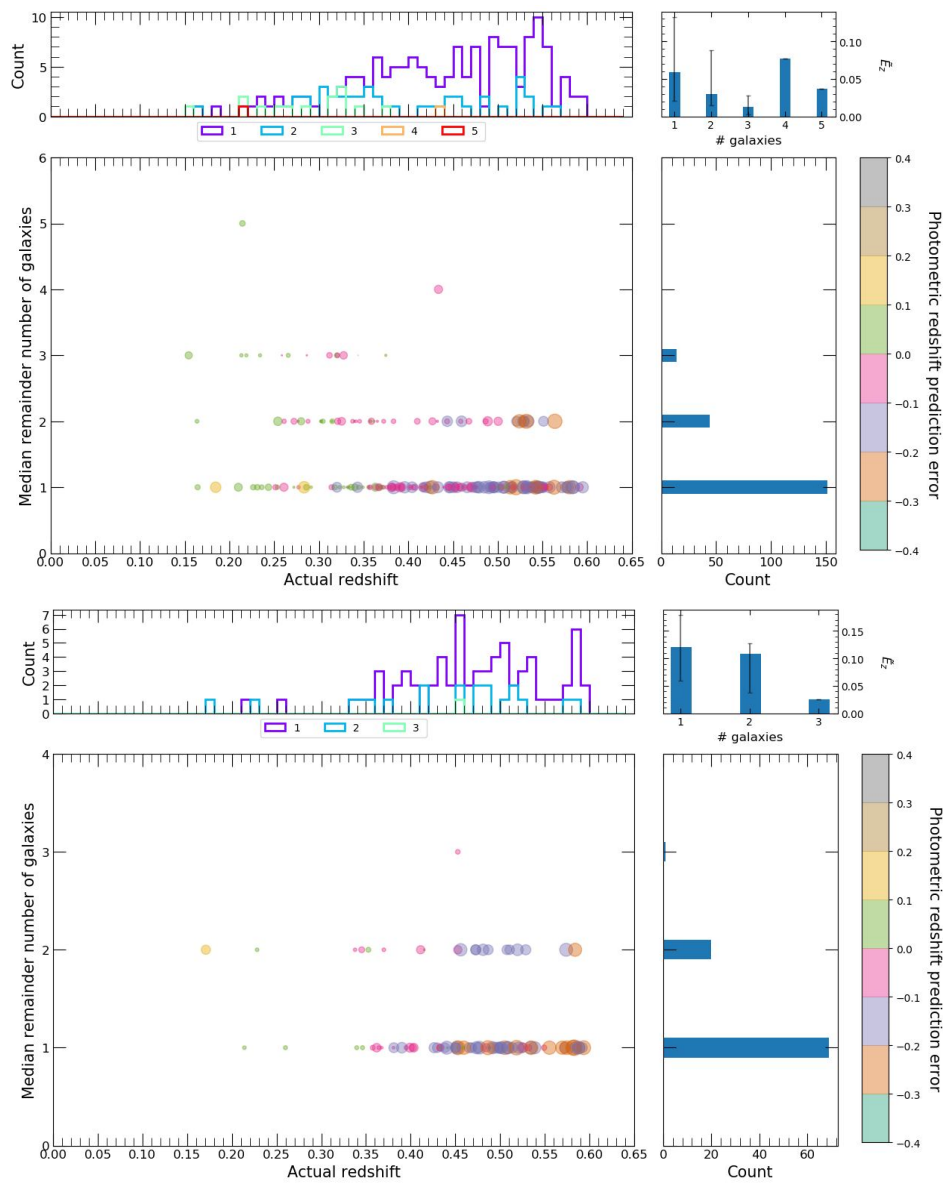


Figure S47. Plots displaying the number of galaxies used in photometric redshift predictions of clusters with low richness versus ‘actual’ redshift of tested clusters, which did not qualify for the MWAR dataset, where predictions had full bootstrap resamples returned within a 21 (top row) or 32 (bottom row) arcseconds search radius. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

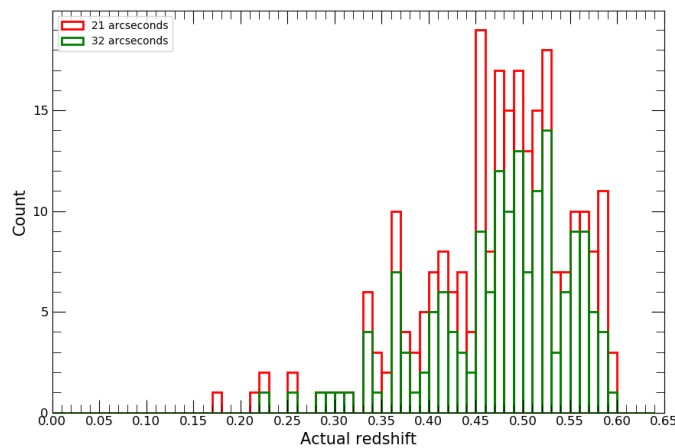


Figure S48. Frequency histograms displaying the ‘actual’ redshift distributions of clusters with low richness, which did not qualify for the MWAR dataset, that had no bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radius.

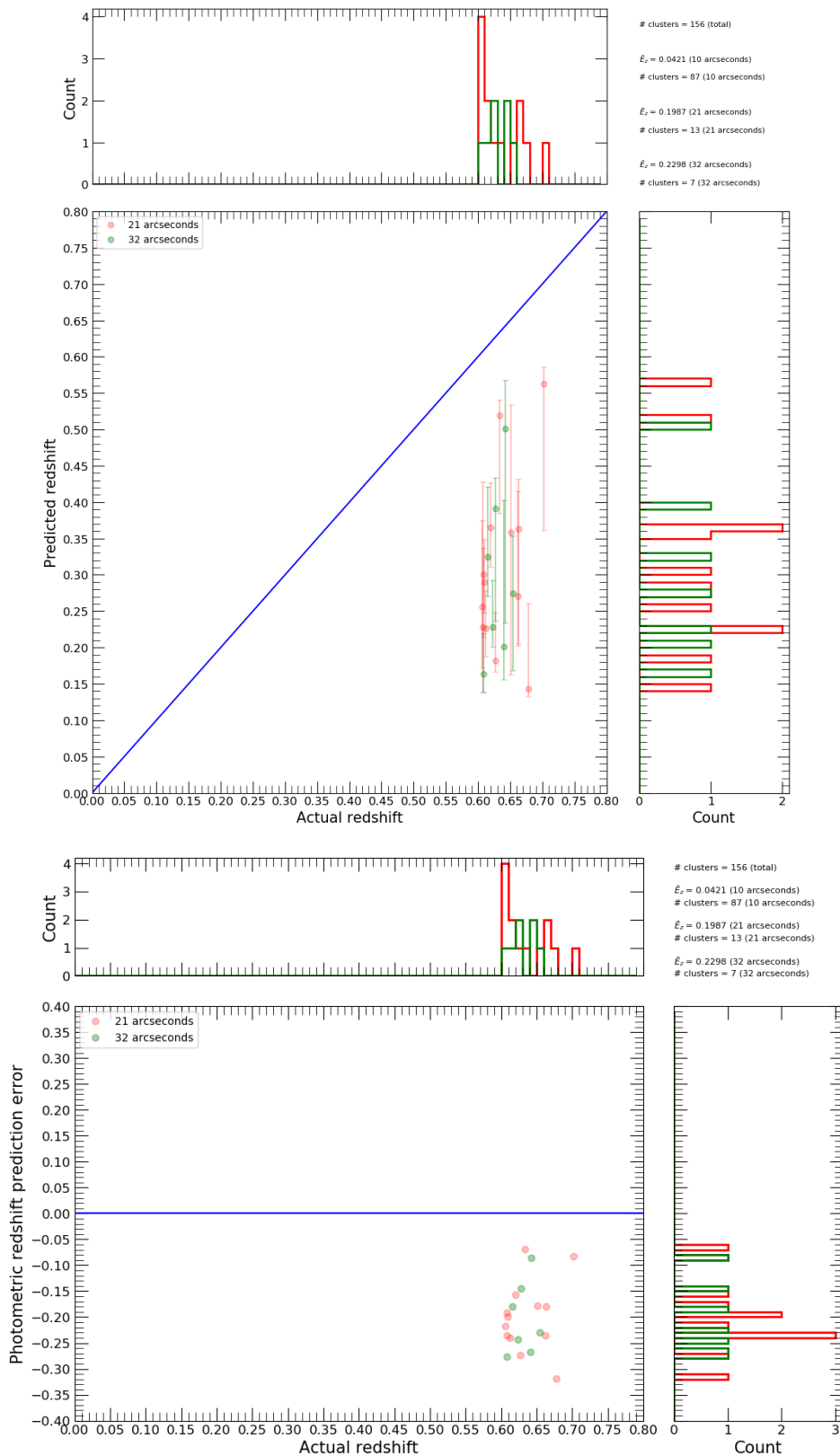


Figure S49. Plots displaying the performance of photometric redshift predictions of clusters at high redshift, which did not qualify for the WNMN dataset, that had full bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radii. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift which did not qualify for the WNMN dataset, '# clusters' represents the number of clusters at high redshift which did not qualify for the WNMN dataset that have observed galaxies within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned.

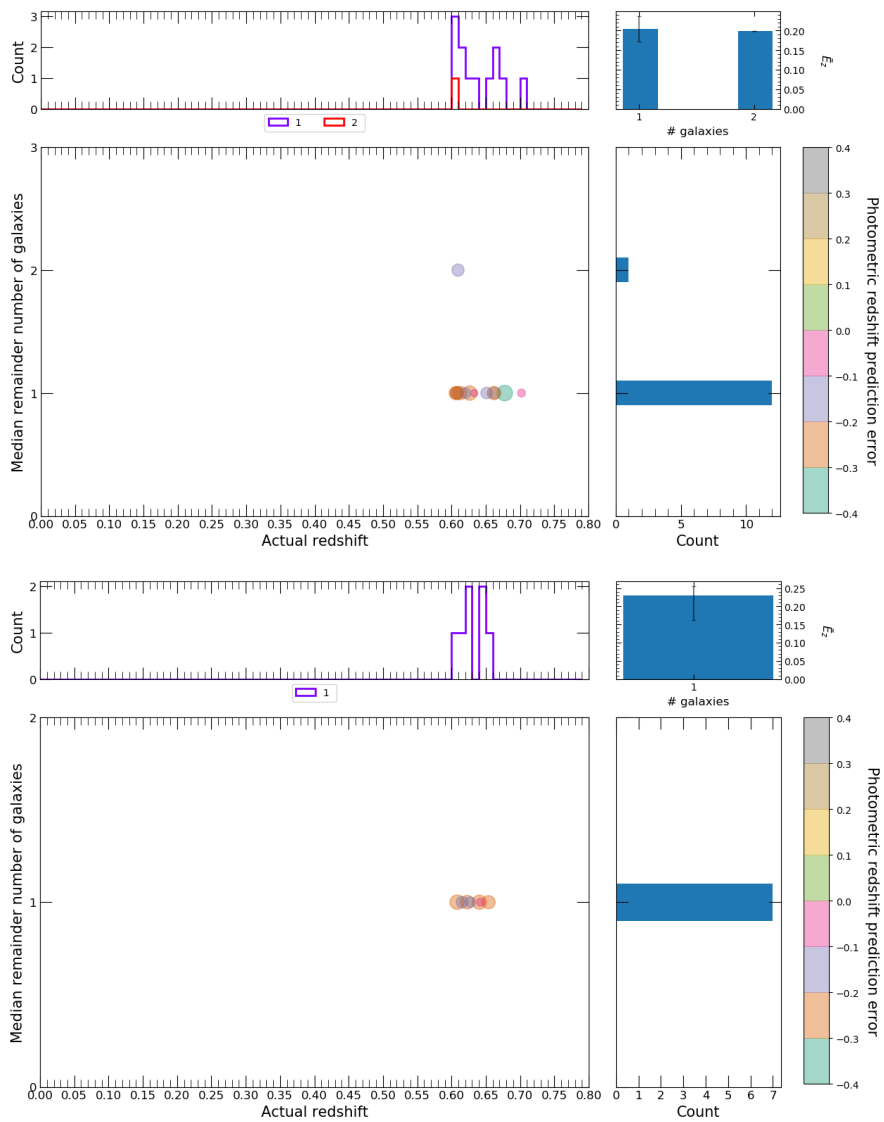


Figure S50. Plots displaying the number of galaxies used in photometric redshift predictions of clusters at high redshift versus 'actual' redshift of tested clusters, which did not qualify for the WNMR dataset, where predictions had full bootstrap resamples returned within a 21 (top row) or 32 (bottom row) arcseconds search radius. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

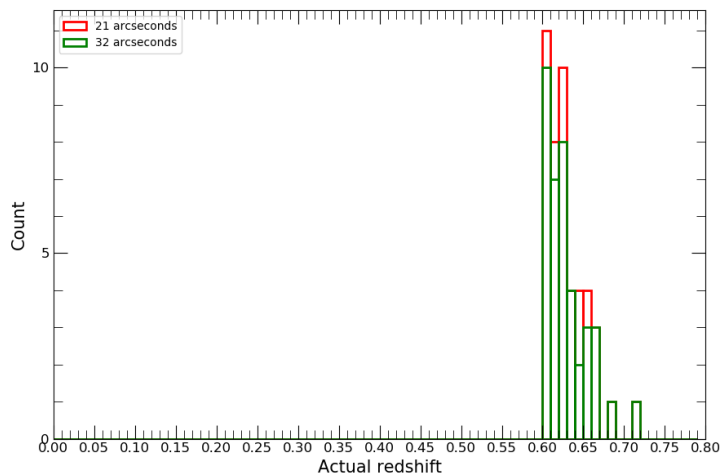


Figure S51. Frequency histograms displaying the 'actual' redshift distributions of clusters at high redshift, which did not qualify for the WNMR dataset, that had no bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radius.

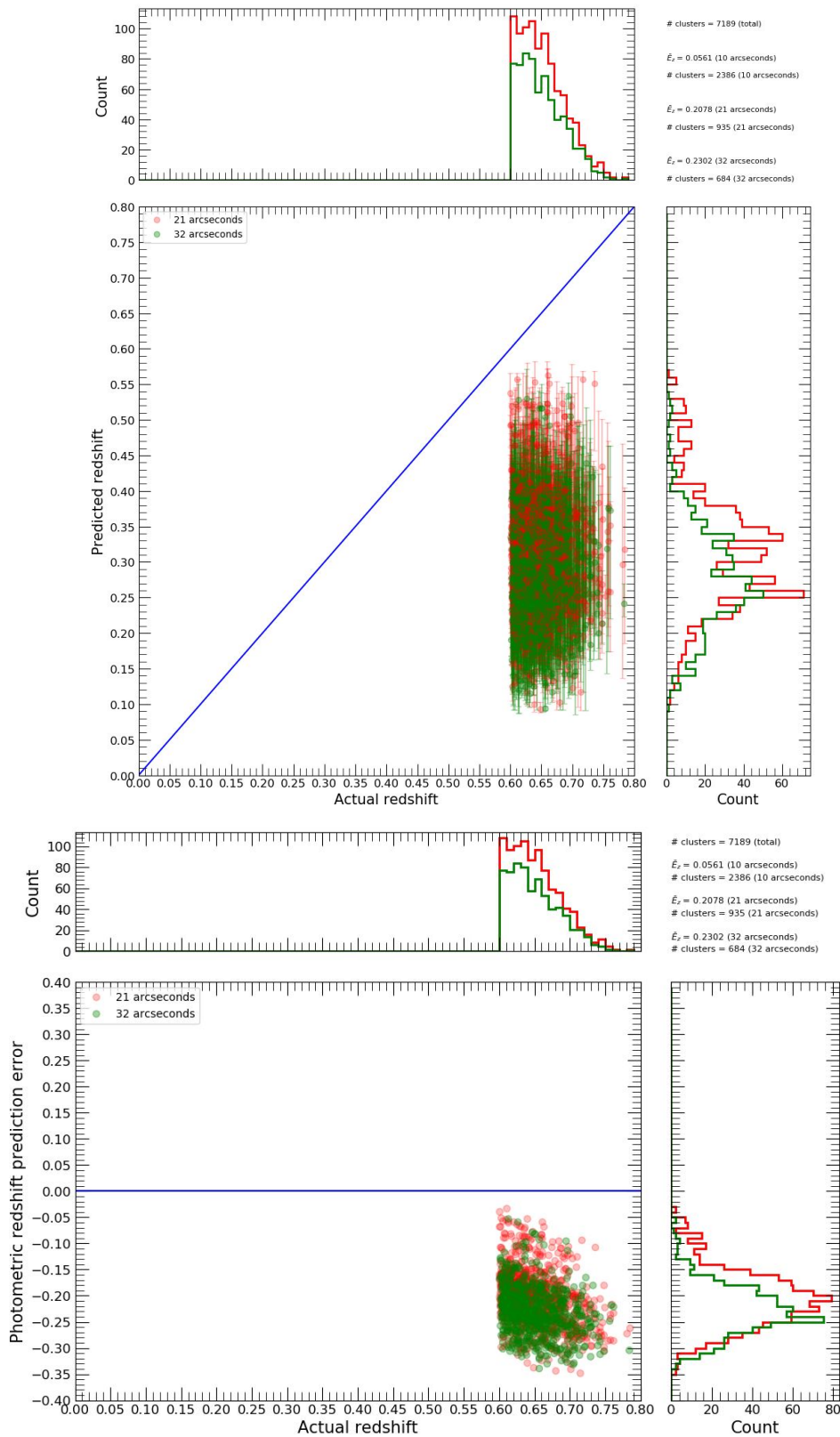


Figure S52. Plots displaying the performance of photometric redshift predictions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radii. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters at high redshift with low richness which did not qualify for the WNMR dataset, '# clusters' represents the number of clusters at high redshift with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned.

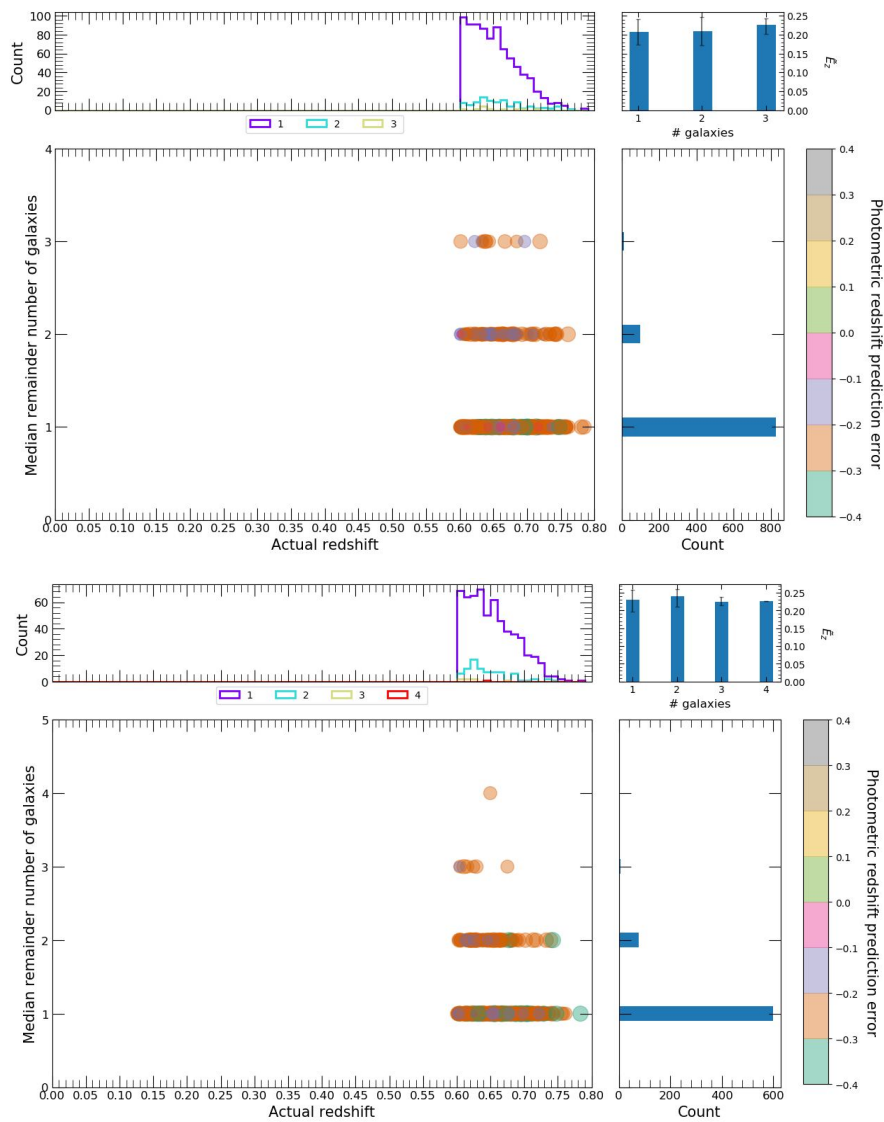


Figure S53. Plots displaying the number of galaxies used in photometric redshift predictions of clusters at high redshift with low richness versus 'actual' redshift of tested clusters, which did not qualify for the WNMR dataset, where predictions had full bootstrap resamples returned within a 21 (top row) or 32 (bottom row) arcseconds search radius. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

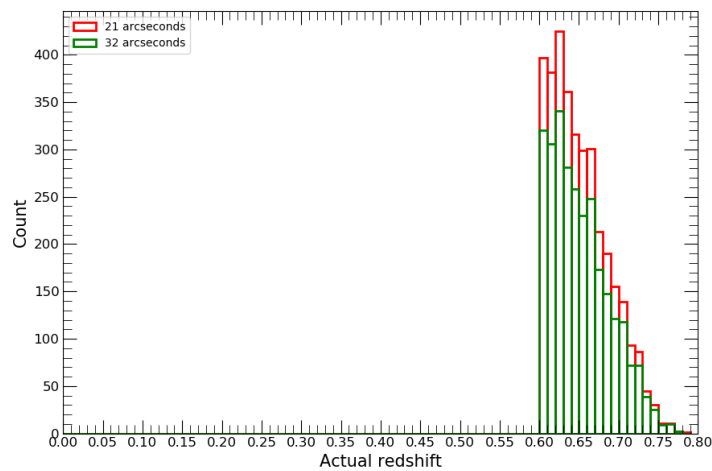


Figure S54. Frequency histograms displaying the 'actual' redshift distributions of clusters at high redshift with low richness, which did not qualify for the WNMR dataset, that had no bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radius.

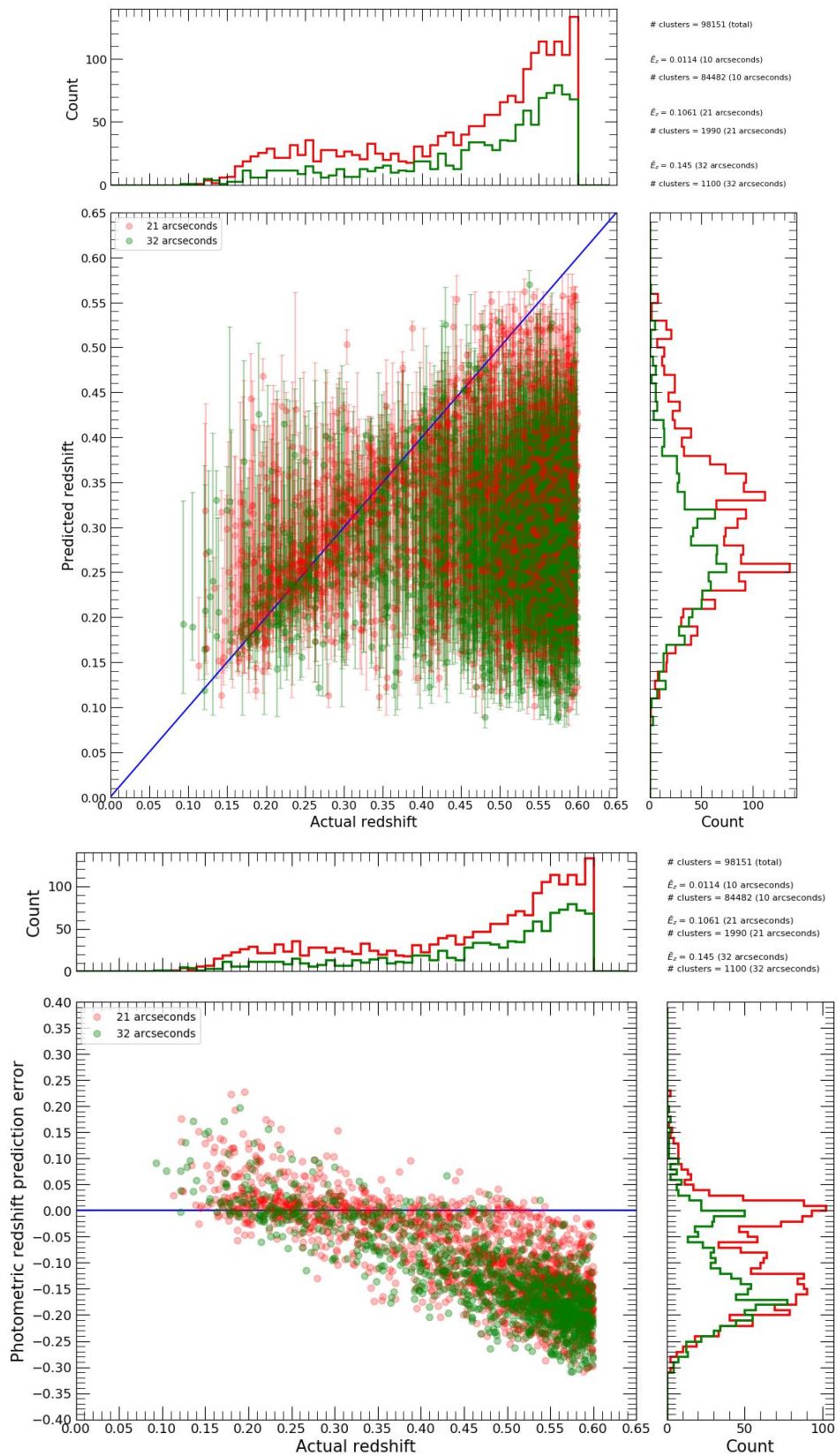


Figure S55. Plots displaying the performance of photometric redshift predictions of clusters with low richness, which did not qualify for the WNMR dataset, that had full bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radii. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters with low richness which did not qualify for the WNMR dataset, '# clusters' represents the number of clusters with low richness which did not qualify for the WNMR dataset that have observed galaxies within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned, \bar{E}_z represents the median of photometric redshift prediction errors across all tested clusters within a 10, 21 and 32 arcseconds search radii with full bootstrap resamples returned.

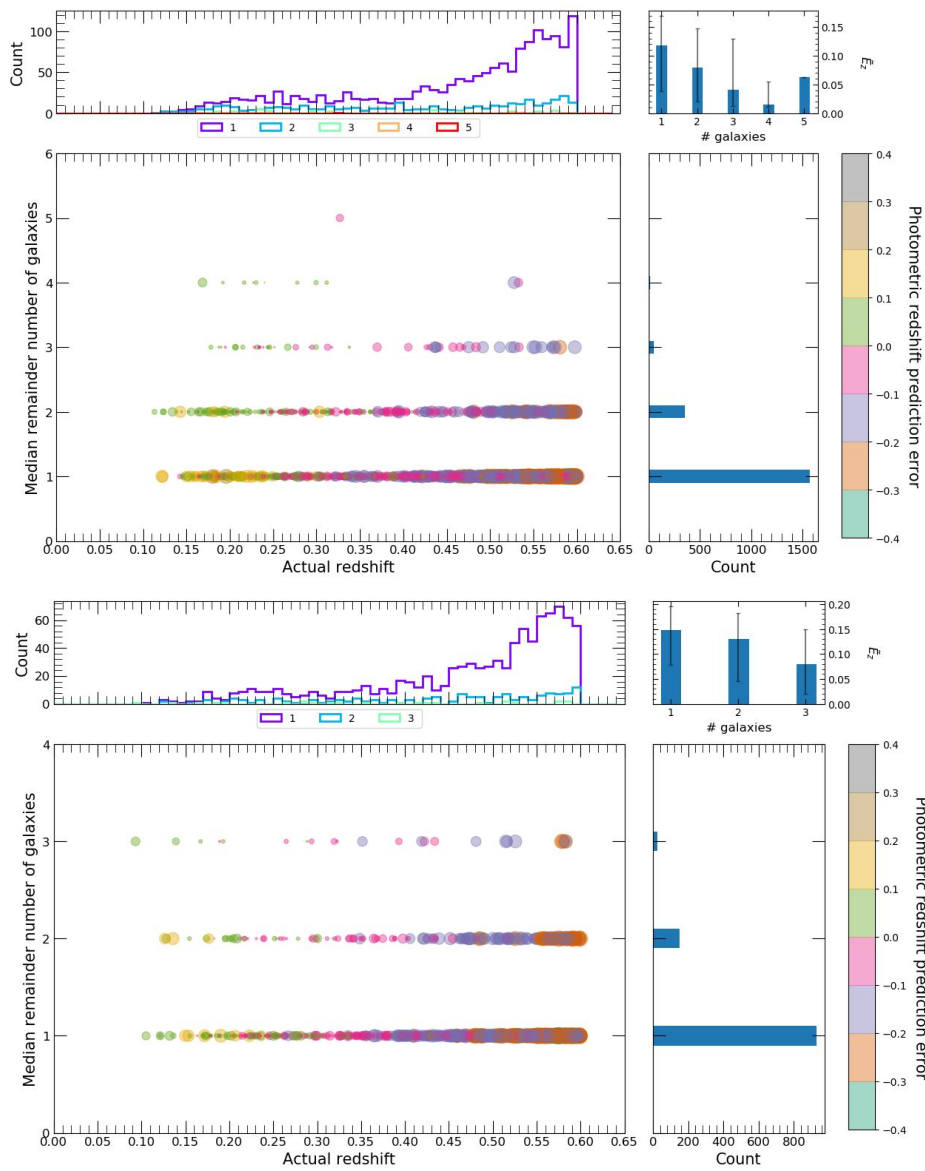


Figure S56. Plots displaying the number of galaxies used in photometric redshift predictions of clusters with low richness versus ‘actual’ redshift of tested clusters, which did not qualify for the WNMR dataset, where predictions had full bootstrap resamples returned within a 21 (top row) or 32 (bottom row) arcseconds search radius. If a cluster has no or partial bootstrap resamples returned at 10 arcseconds then the search radius is increased until a prediction with full bootstrap resamples returned is obtained. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. \tilde{E}_z represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

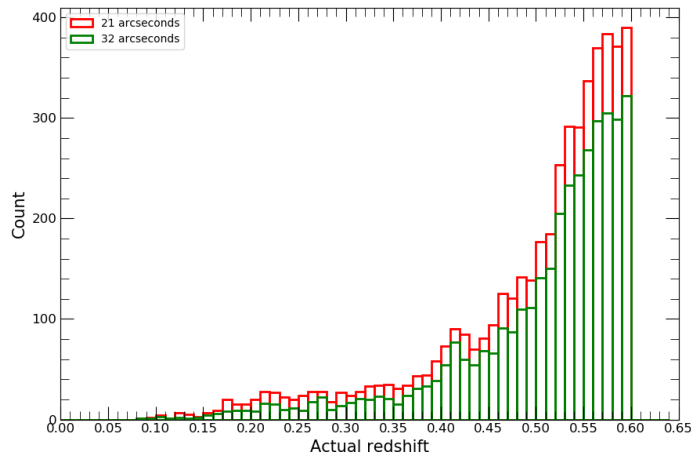


Figure S57. Frequency histograms displaying the ‘actual’ redshift distributions of clusters with low richness, which did not qualify for the WNMR dataset, that had no bootstrap resamples returned within a 21 (red) or 32 (green) arcseconds search radius.

APPENDIX

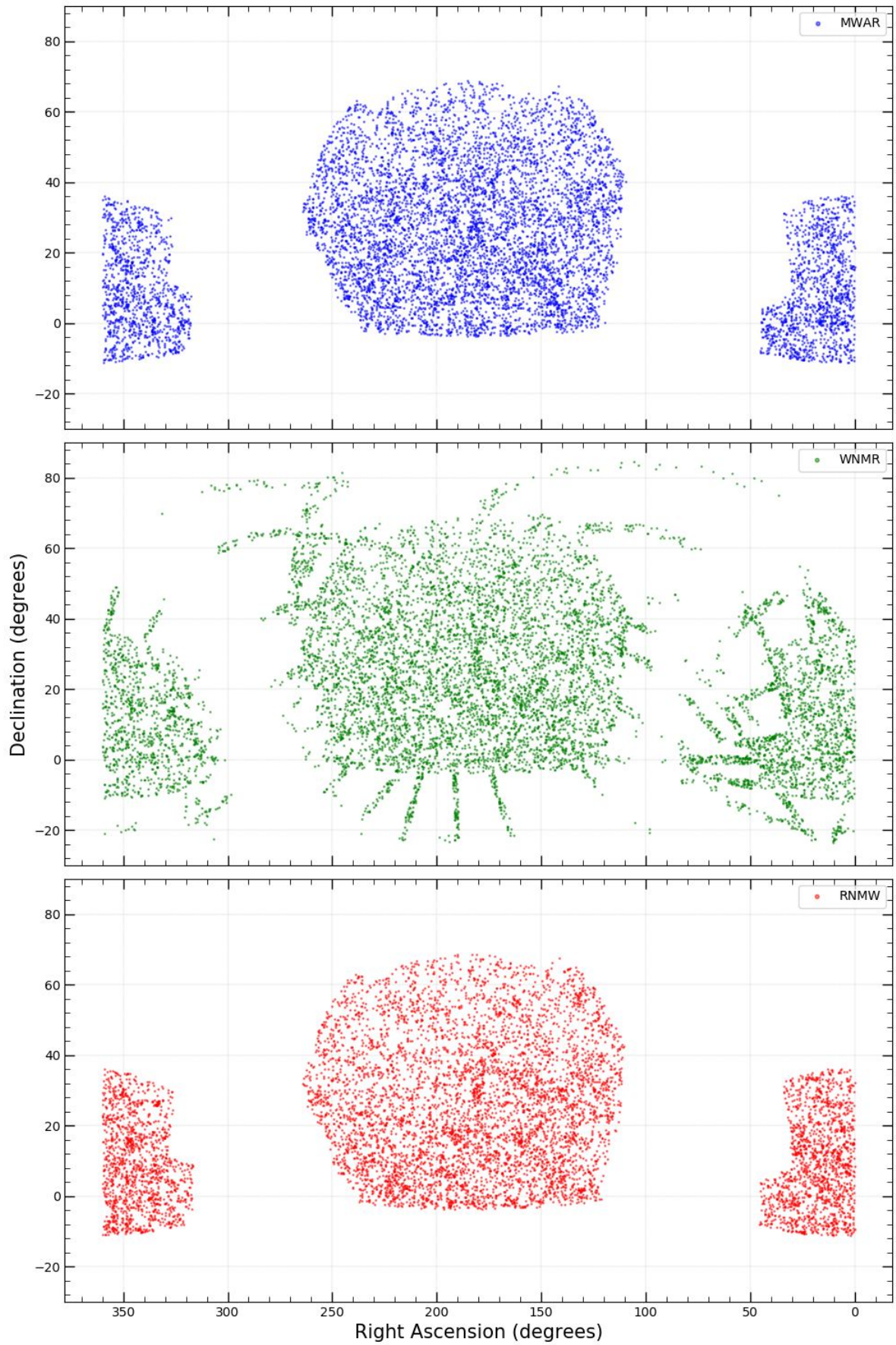


Figure SA1. Astronomical sky maps with the J2000 epoch coordinate system displaying the observed positions of clusters in the MWAR (top row), WNMR (middle row) and RNMW (bottom row) datasets.

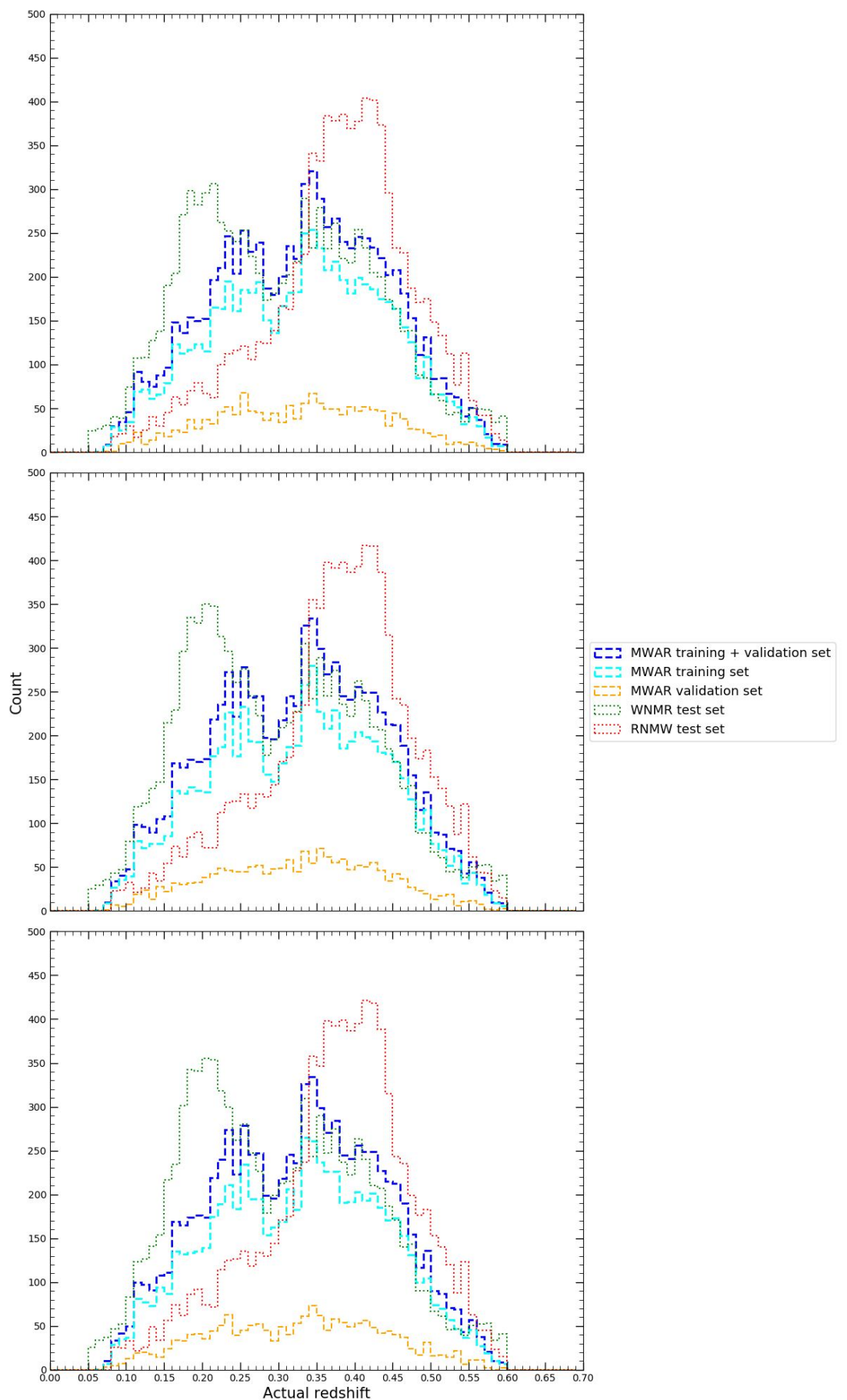


Figure SA2. Frequency histograms of the 'actual' redshift distributions, where photometric redshifts of clusters in the MWAR dataset (blue dashed line), MWAR training set (cyan dashed line), MWAR validation set (orange dashed line) and WNMR test set (green dotted line) are originally estimated by WHL12. Whilst photometric redshifts of clusters in the RNMW test set (red dotted line) are originally estimated by redMaPPer. The top row contains clusters within a 10 arcseconds search radius, the middle row contains clusters within a 21 arcseconds search radius and the bottom row contains clusters within a 32 arcseconds search radius.

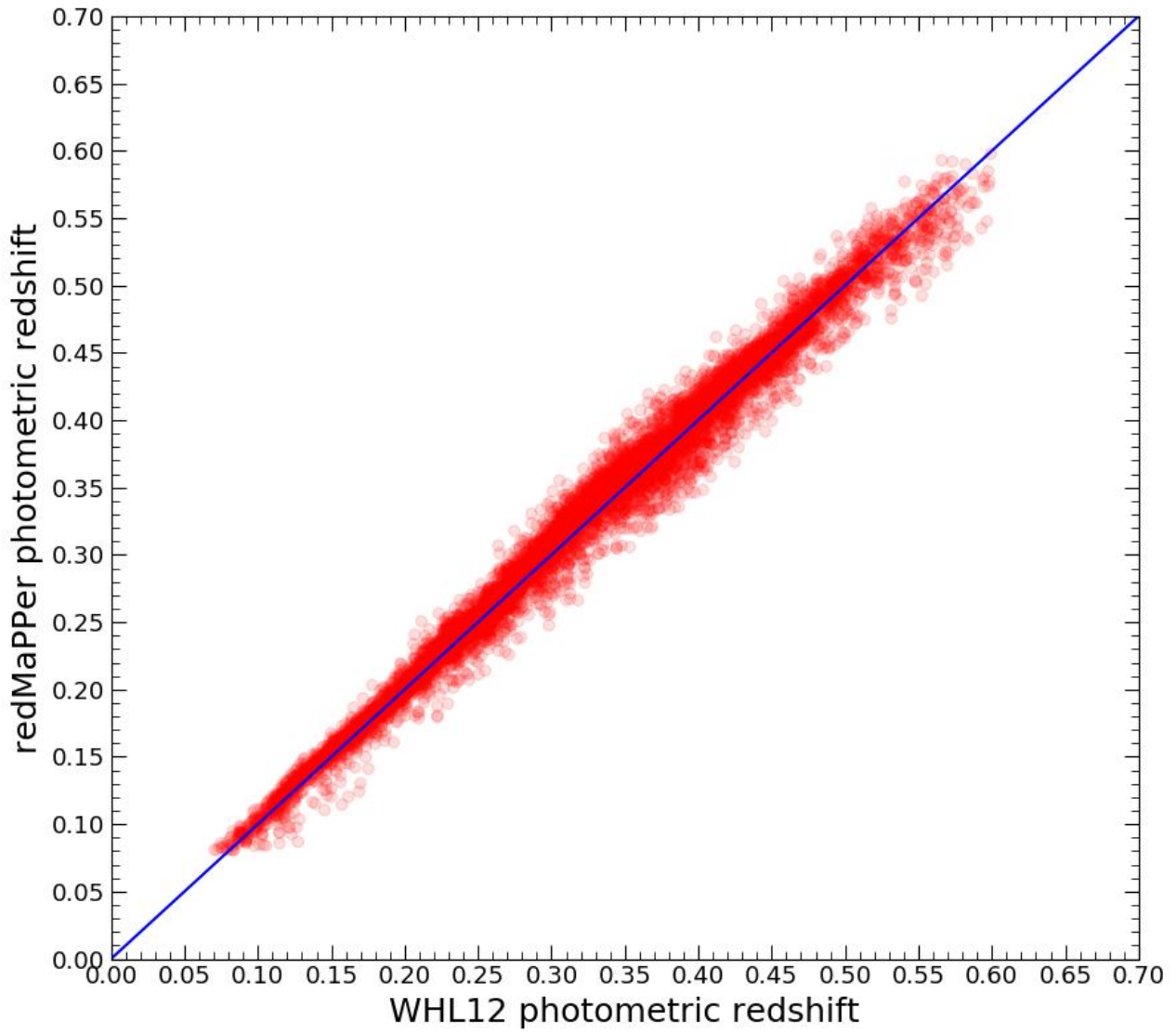


Figure SA3. Comparison of photometric redshifts for clusters in the MWAR dataset that are originally estimated by the WHL12 and redMaPPer cluster catalogues.

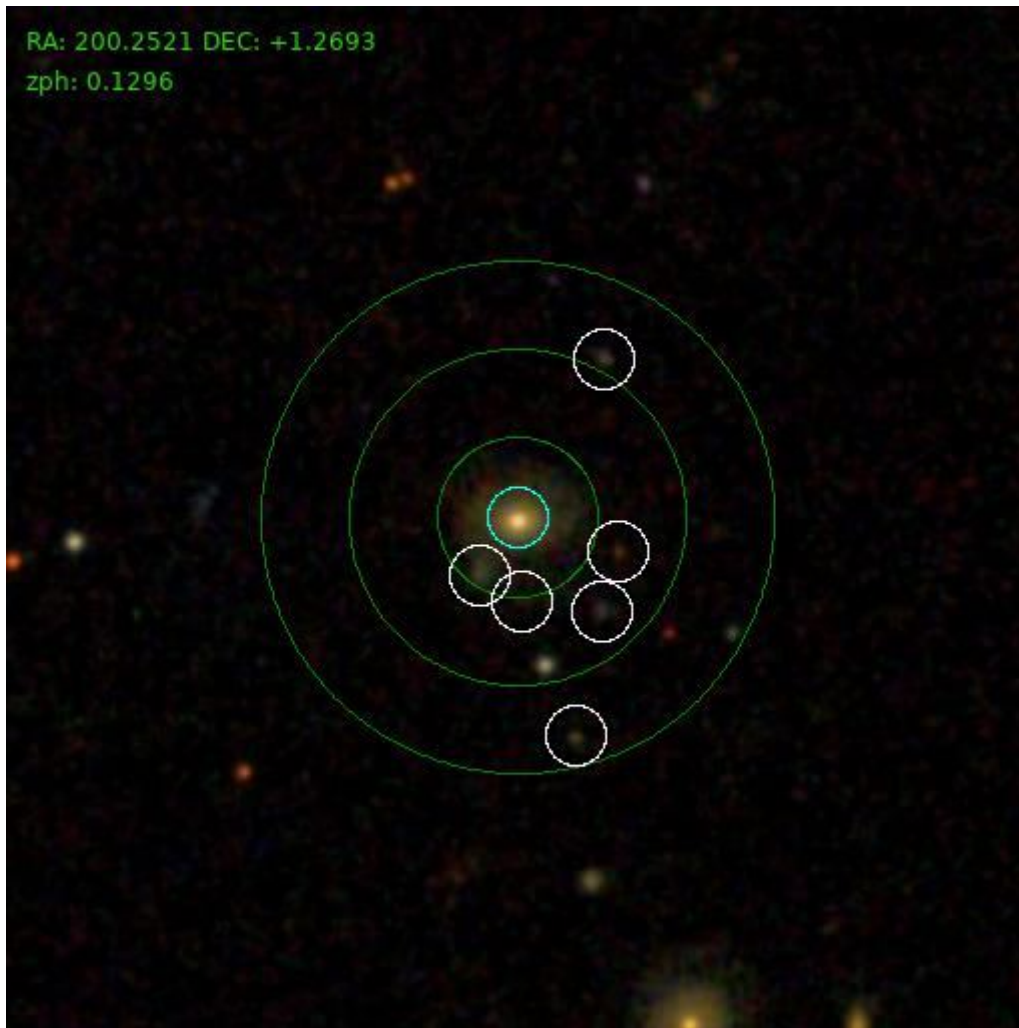


Figure SA4. An SDSS image of a cluster from the WNMR test set with J2000 coordinates of RA: 200.2521 and Dec: +1.2693. The cluster has a photometric redshift of z_{ph} : 0.1296 in the WHL12 cluster catalogue. The green circles represent the size of the 10, 21 and 32 arcseconds search radii respectively. Any objects with no circle surrounding it within the 32 arcseconds search radius is not considered to be a galaxy or is a galaxy with poor photometry. The white circles represent galaxies with 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that are removed if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. The cyan circles represent galaxies with also 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that remain if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. From which, we find that only the central galaxy remains within the 32 arcseconds search radius, such that the tuned model returns photometric redshift estimates of z_{ph} : 0.1269 with $CI_{95\%}=[0.0961, 0.3093]$ for the 10 arcseconds search radius, z_{ph} : 0.1092 with $CI_{95\%}=[0.0893, 0.4661]$ for the 21 arcseconds search radius and z_{ph} : 0.3976 with $CI_{95\%}=[0.1526, 0.4918]$ for the 32 arcseconds search radius. This example shows that the training set itself can contaminate model predictions when a larger search radius is applied, as we observe that no interloping galaxies are included when the 32 arcseconds search radius is used.

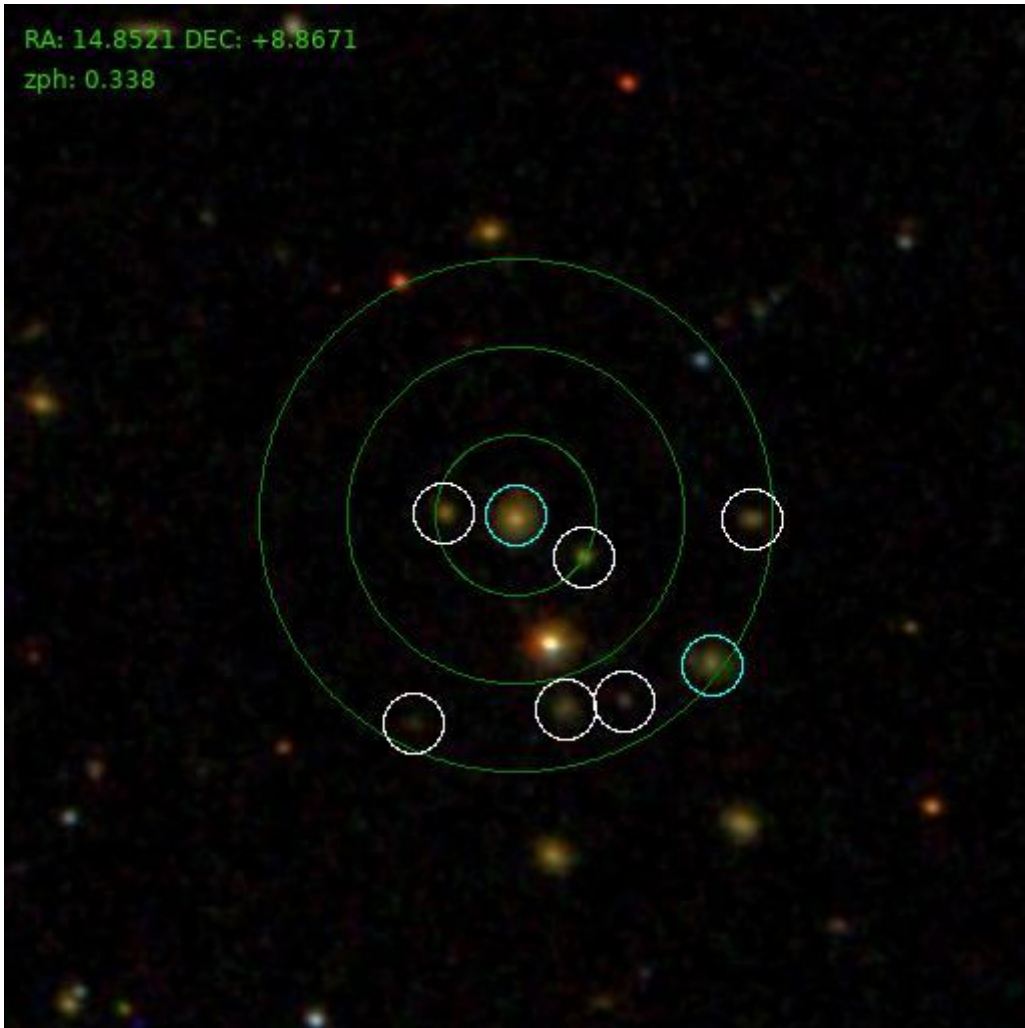


Figure SA5. An SDSS image of a cluster from the WNMR test set with J2000 coordinates of RA: 14.8521 and Dec: +8.8671. The cluster has a photometric redshift of $z_{ph}: 0.338$ in the WHL12 cluster catalogue. The green circles represent the size of the 10, 21 and 32 arcseconds search radii respectively. Any objects with no circle surrounding it within the 32 arcseconds search radius is not considered to be a galaxy or is a galaxy with poor photometry. The white circles represent galaxies with 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that are removed if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. The cyan circles represent galaxies with also 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that remain if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. From which, we find that the central galaxy and another galaxy remains within the 32 arcseconds search radius, such that the tuned model returns photometric redshift estimates of $z_{ph}: 0.1859$ with $CI_{95\%}=[0.1664, 0.2261]$ for the 10 arcseconds search radius, $z_{ph}: 0.1877$ with $CI_{95\%}=[0.1653, 0.2519]$ for the 21 arcseconds search radius and $z_{ph}: 0.1651$ with $CI_{95\%}=[0.1431, 0.2001]$ for the 32 arcseconds search radius. This example shows that the training set itself can contaminate model predictions at all search radii, as we observe that no obvious interloping galaxies are included for any search radii.

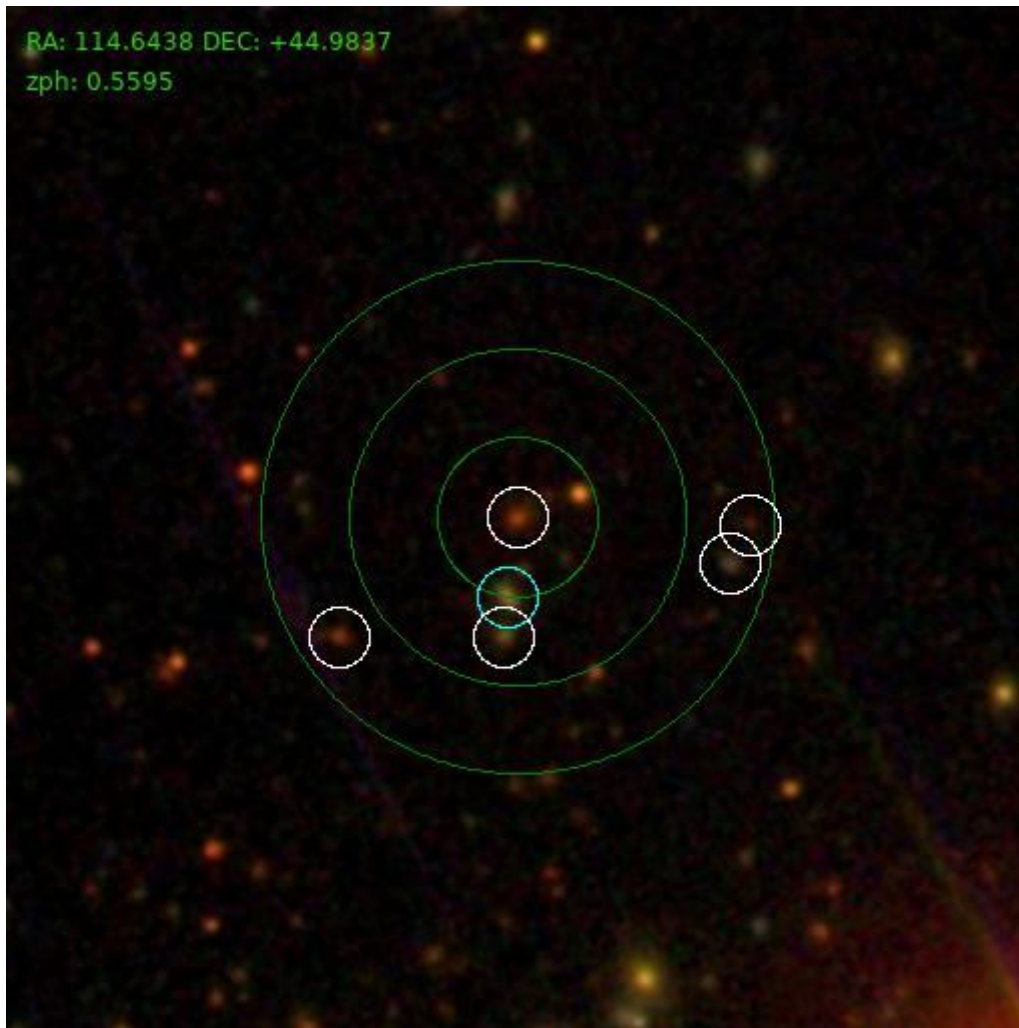


Figure SA6. An SDSS image of a cluster from the WNMR test set with J2000 coordinates of RA: 114.6438 and Dec: +44.9837. The cluster has a photometric redshift of $z_{ph}: 0.5595$ in the WHL12 cluster catalogue. The green circles represent the size of the 10, 21 and 32 arcseconds search radii respectively. Any objects with no circle surrounding it within the 32 arcseconds search radius is not considered to be a galaxy or is a galaxy with poor photometry. The white circles represent galaxies with 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that are removed if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. The cyan circles represent galaxies with also 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that remain if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. From which, we find that an interloping galaxy remains within the 32 arcseconds search radius, such that the tuned model returns photometric redshift estimates of $z_{ph}: 0.1730$ with $CI_{95\%}=[0.1153, 0.2350]$ for the 10 arcseconds search radius, $z_{ph}: 0.2399$ with $CI_{95\%}=[0.1960, 0.2710]$ for the 21 arcseconds search radius and $z_{ph}: 0.1637$ with $CI_{95\%}=[0.1152, 0.3672]$ for the 32 arcseconds search radius. This example shows that an interloping galaxy within the 10 arcseconds search radius, presumably at lower redshift than the cluster itself, can contaminate model predictions for all search radii.

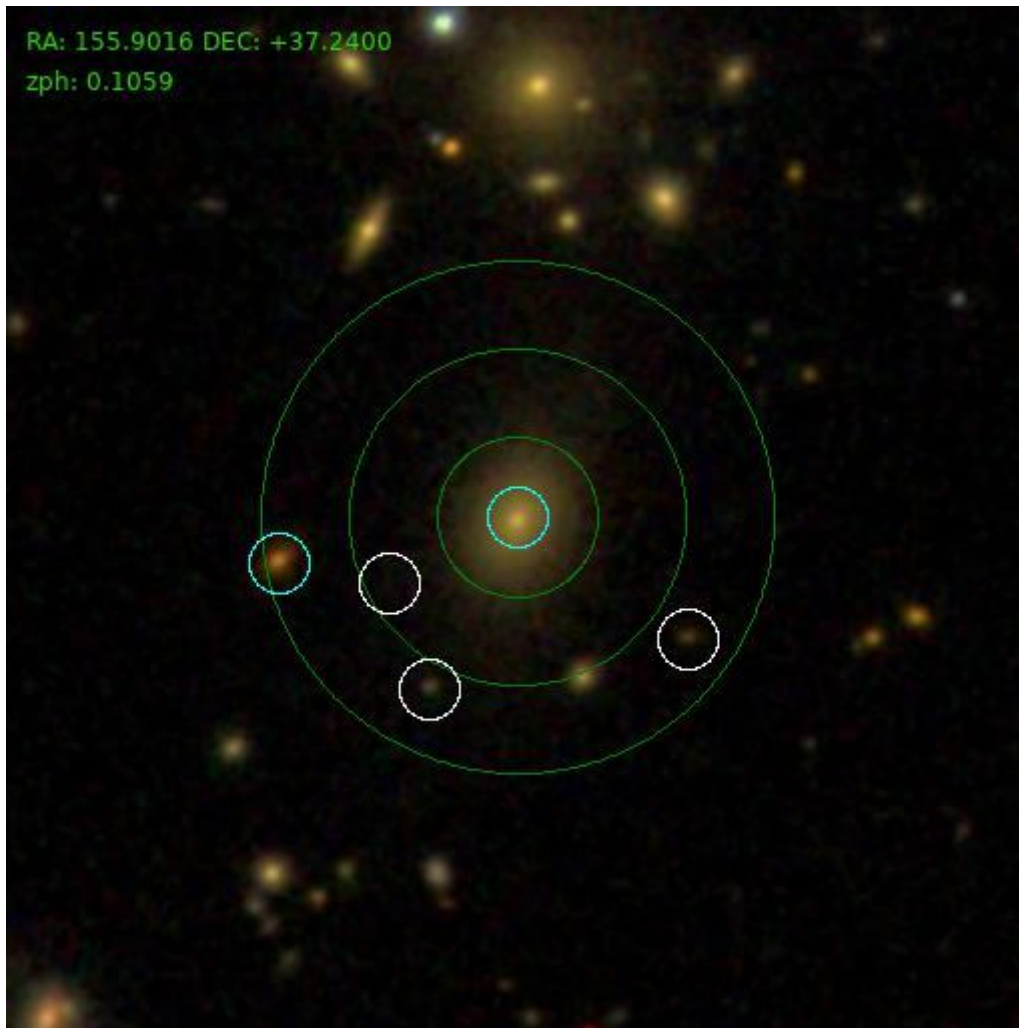


Figure SA7. An SDSS image of a cluster from the RNMW test set with J2000 coordinates of RA: 155.9016 and Dec: +37.2400. The cluster has a photometric redshift of $z_{ph}: 0.1059$ in the redMaPPer cluster catalogue. The green circles represent the size of the 10, 21 and 32 arcseconds search radii respectively. Any objects with no circle surrounding it within the 32 arcseconds search radius is not considered to be a galaxy or is a galaxy with poor photometry. The white circles represent galaxies with 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that are removed if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. The cyan circles represent galaxies with also 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that remain if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. From which, we find that the central galaxy and an interloping galaxy remains within the 32 arcseconds search radius, such that the tuned model returns photometric redshift estimates of $z_{ph}: 0.1115$ with $CI_{95\%}=[0.0914, 0.1239]$ for the 10 arcseconds search radius, $z_{ph}: 0.1121$ with $CI_{95\%}=[0.1024, 0.1198]$ for the 21 arcseconds search radius and $z_{ph}: 0.2709$ with $CI_{95\%}=[0.1972, 0.4003]$ for the 32 arcseconds search radius. This example shows that an interloping galaxy, presumably at higher redshift than the cluster itself, can contaminate model predictions when the search radius increases.

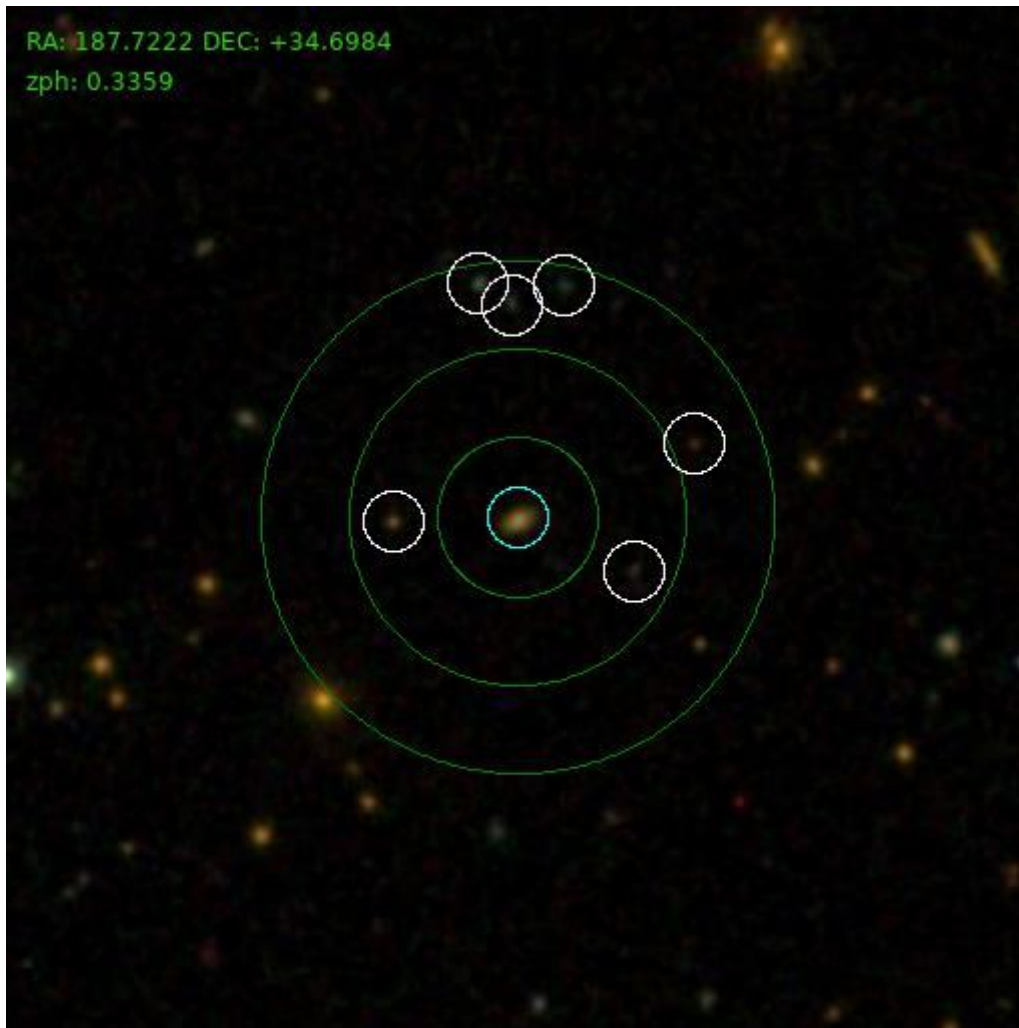


Figure SA8. An SDSS image of a cluster from the RNMW test set with J2000 coordinates of RA: 187.7222 and Dec: +34.6984. The cluster has a photometric redshift of $z_{ph}: 0.3359$ in the redMaPPer cluster catalogue. The green circles represent the size of the 10, 21 and 32 arcseconds search radii respectively. Any objects with no circle surrounding it within the 32 arcseconds search radius is not considered to be a galaxy or is a galaxy with poor photometry. The white circles represent galaxies with 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that are removed if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. The cyan circles represent galaxies with also 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that remain if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius are applied at the same time. From which, we find that only the central galaxy remains within the 32 arcseconds search radius, such that the tuned model returns photometric redshift estimates of $z_{ph}: 0.1664$ with $CI_{95\%}=[0.1189, 0.1820]$ for the 10 arcseconds search radius, $z_{ph}: 0.1617$ with $CI_{95\%}=[0.1264, 0.2041]$ for the 21 arcseconds search radius and $z_{ph}: 0.1474$ with $CI_{95\%}=[0.1191, 0.2131]$ for the 32 arcseconds search radius. This example also shows that the training set itself can contaminate model predictions at all search radii, as we observe that no interloping galaxies are included for any search radii.

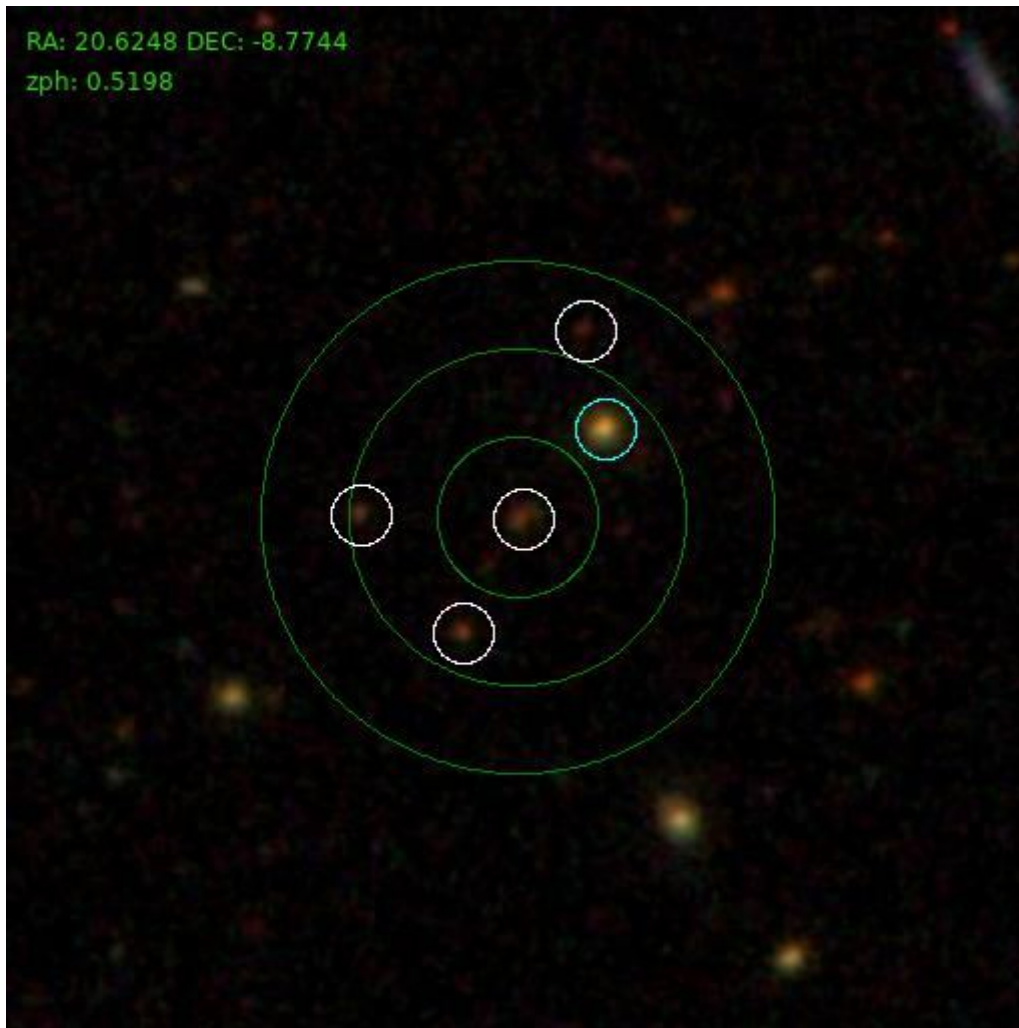


Figure SA9. An SDSS image of a cluster from the RNMW test set with J2000 coordinates of RA: 20.6248 and Dec: -8.7744. The cluster has a photometric redshift of $z_{ph}: 0.5198$ in the redMaPPer cluster catalogue. The green circles represent the size of the 10, 21 and 32 arcseconds search radii respectively. Any objects with no circle surrounding it within the 32 arcseconds search radius is not considered to be a galaxy or is a galaxy with poor photometry. The white circles represent galaxies with 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that are removed if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius were applied at the same time. The cyan circles represent galaxies with also 'clean' photometry, as determined by SDSS, found within the 32 arcseconds search radius that remain if all LM-0.5 filter magnitude-cuts (excluding the u filter) for the 32 arcseconds search radius were applied at the same time. From which, we find that an interloping galaxy remains within the 32 arcseconds search radius, such that the tuned model returns photometric redshift estimates of $z_{ph}: 0.5121$ with $CI_{95\%}=[0.4954, 0.5281]$ for the 10 arcseconds search radius, $z_{ph}: 0.4191$ with $CI_{95\%}=[0.2343, 0.4699]$ for the 21 arcseconds search radius and $z_{ph}: 0.1415$ with $CI_{95\%}=[0.1259, 0.2168]$ for the 32 arcseconds search radius. This example shows that an interloping galaxy, presumably at lower redshift than the cluster itself, can contaminate model predictions when the search radius increases.

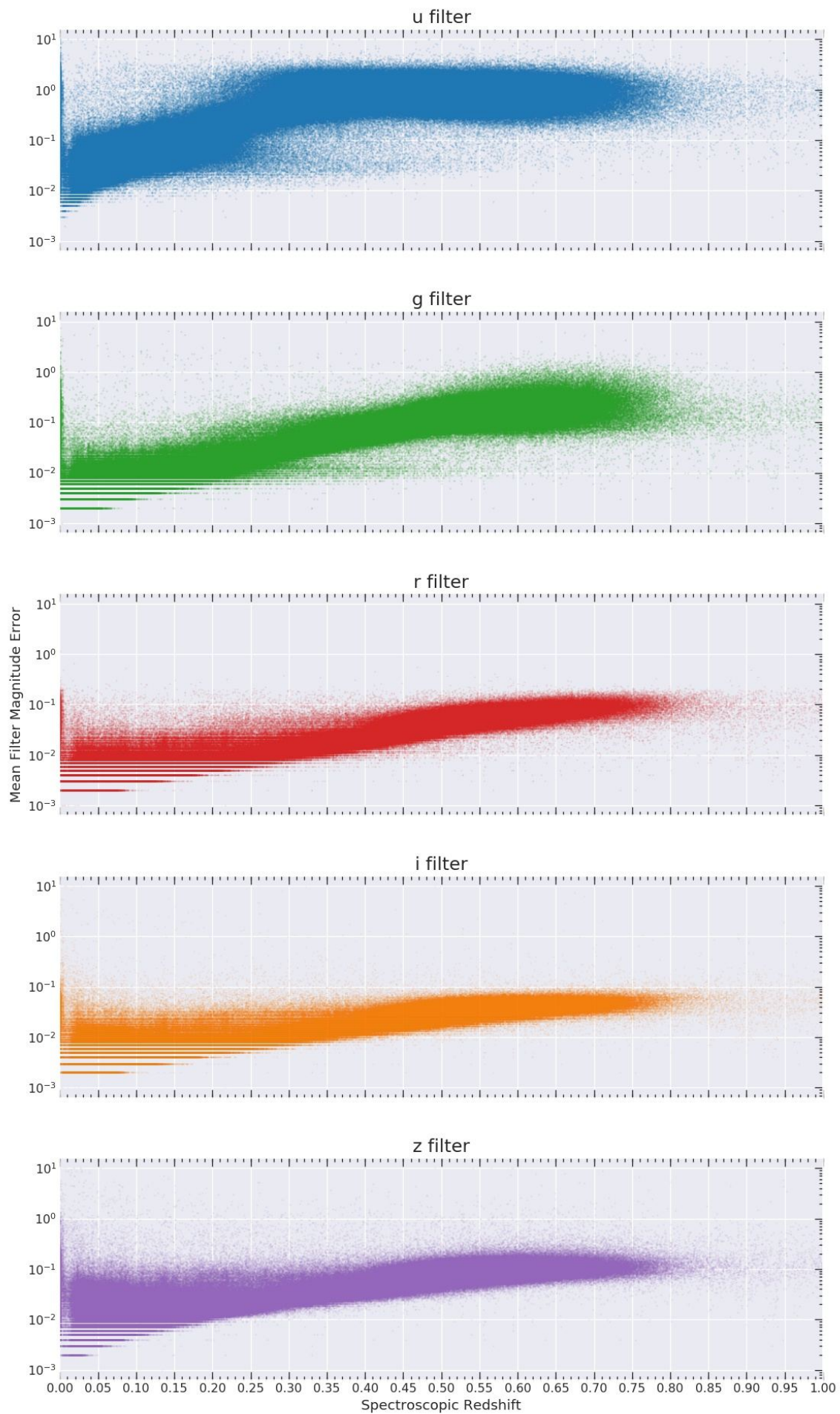
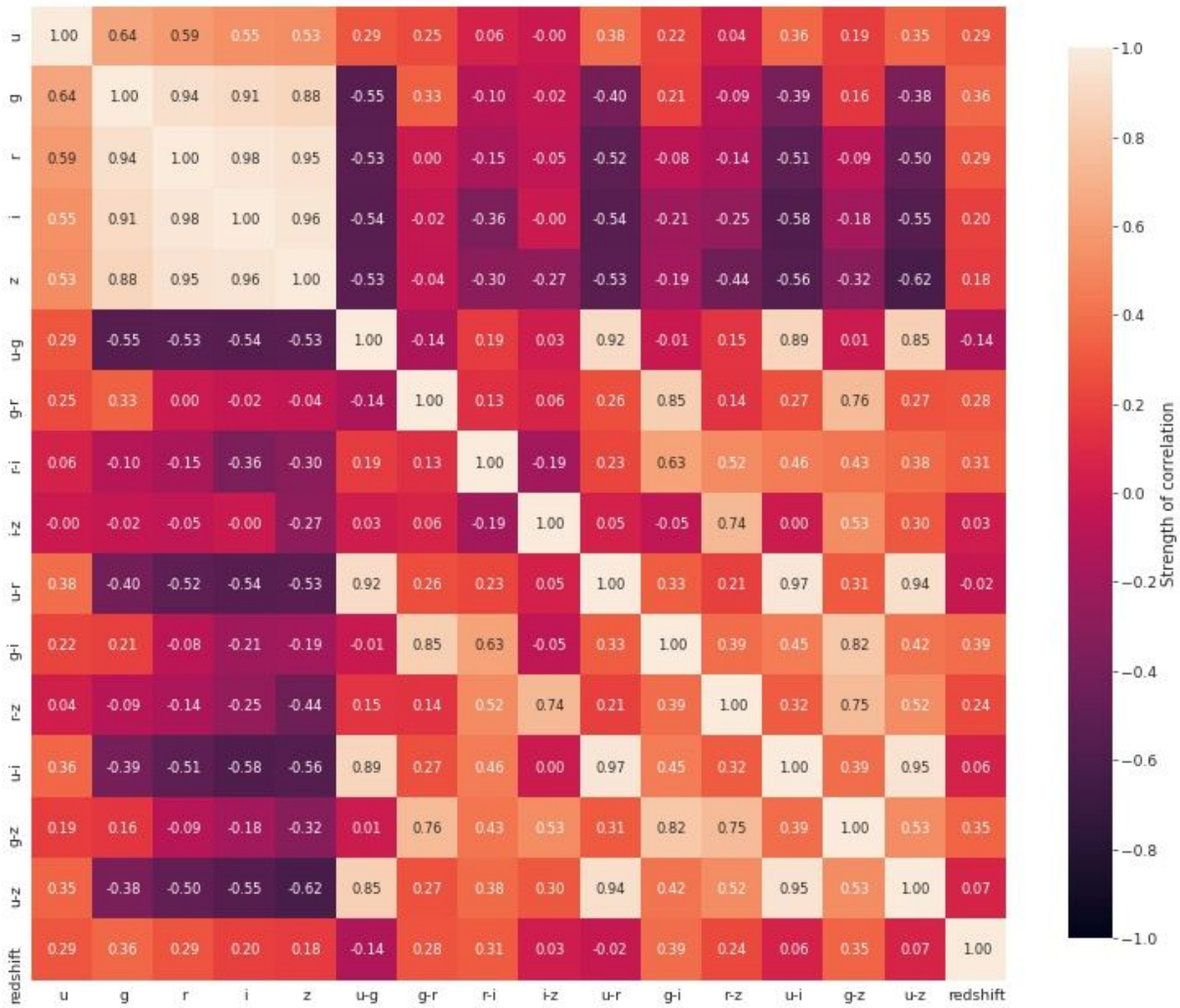
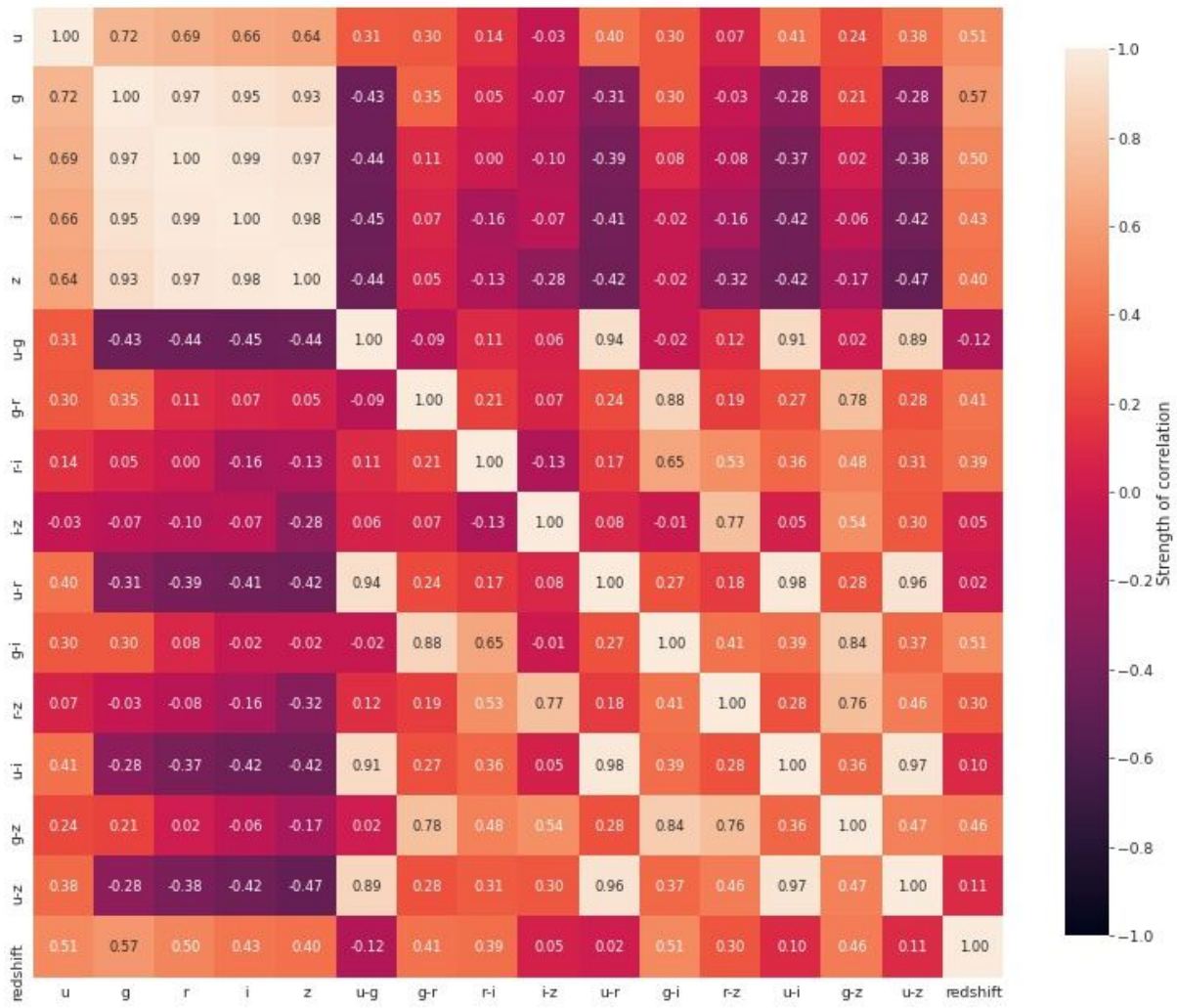


Figure SA10. Plots displaying the mean filter magnitude errors versus spectroscopic redshift of observed galaxies in SDSS for the u (top row), g (second from top row), r (middle row), I (second from bottom row) and z (bottom row) filters.



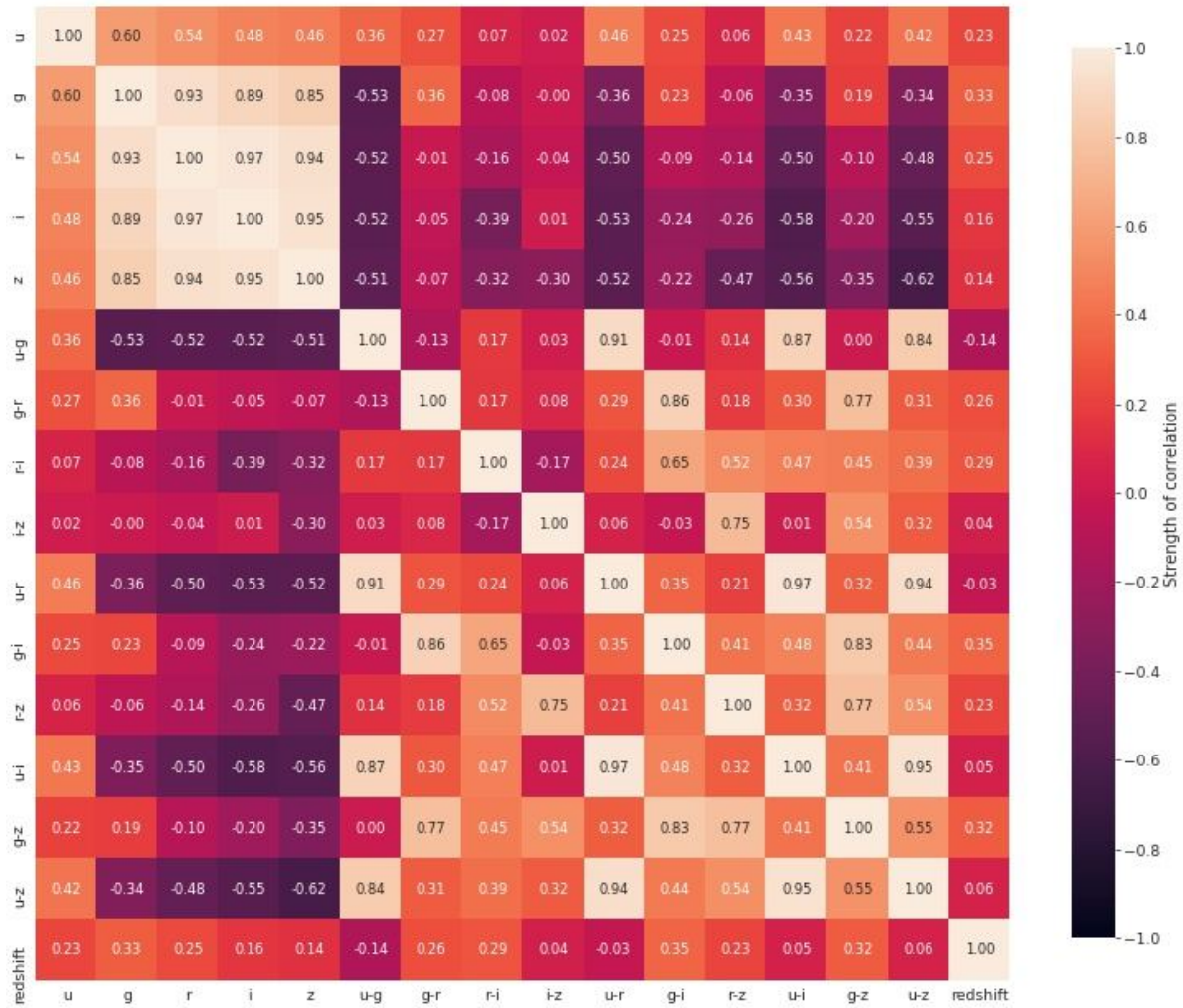


Figure SA11. Correlation matrix heatmaps of all features with raw photometry data from the MWAR training set for the 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius. The colourbar represents the strength of correlation between the features.

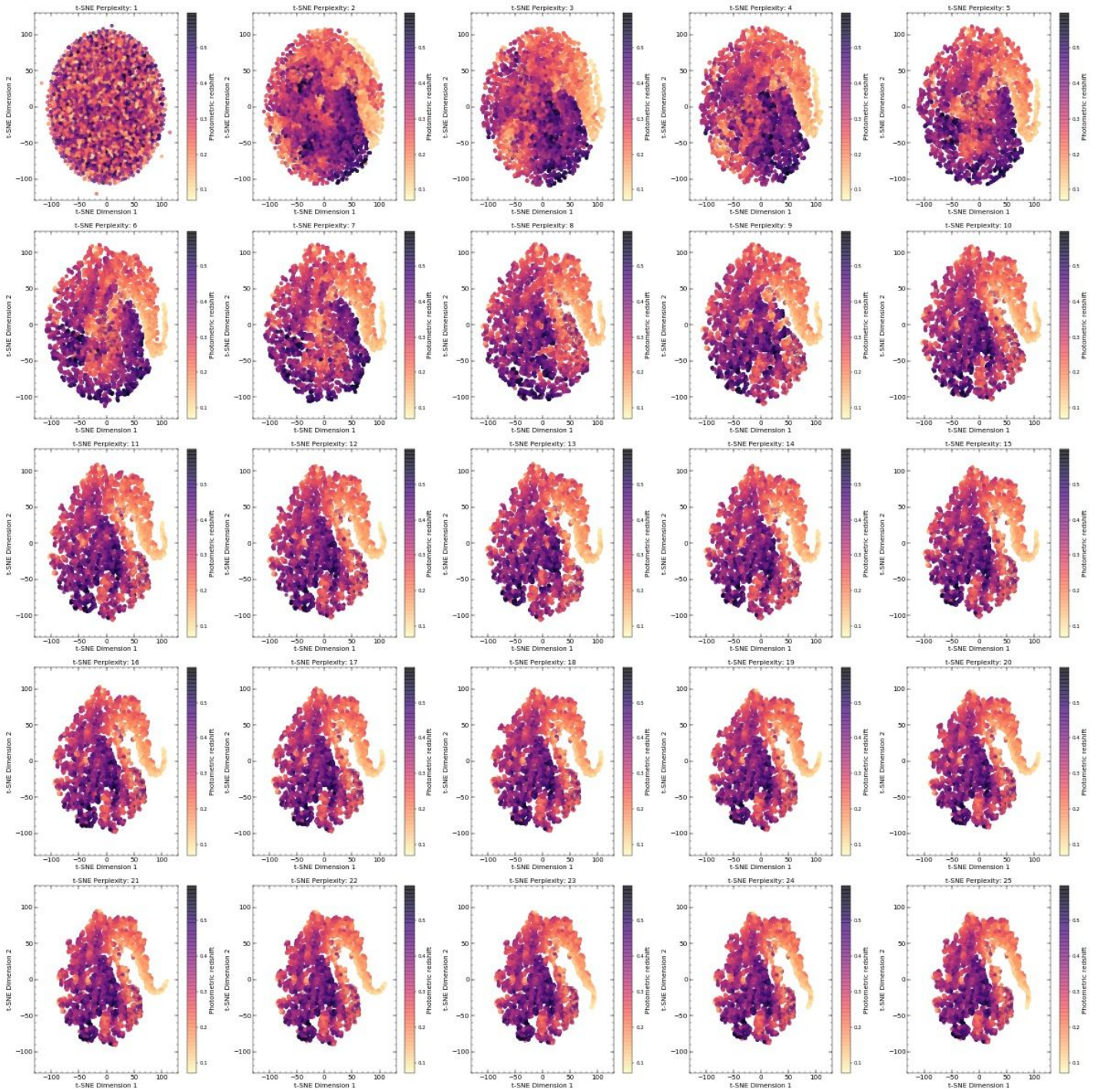


Figure SA13. Two dimensional representation of the feature space in the MWAR training set with no feature scaling applied for a 10 arcseconds search radius with LM filter magnitude-cuts applied using the t-SNE algorithm. The colourbar represents the photometric redshift of galaxies found within the search radius of clusters originally estimated by WHL12. The t-SNE perplexity value relates to the number of nearest neighbours used to compress the dimensionality of the dataset.

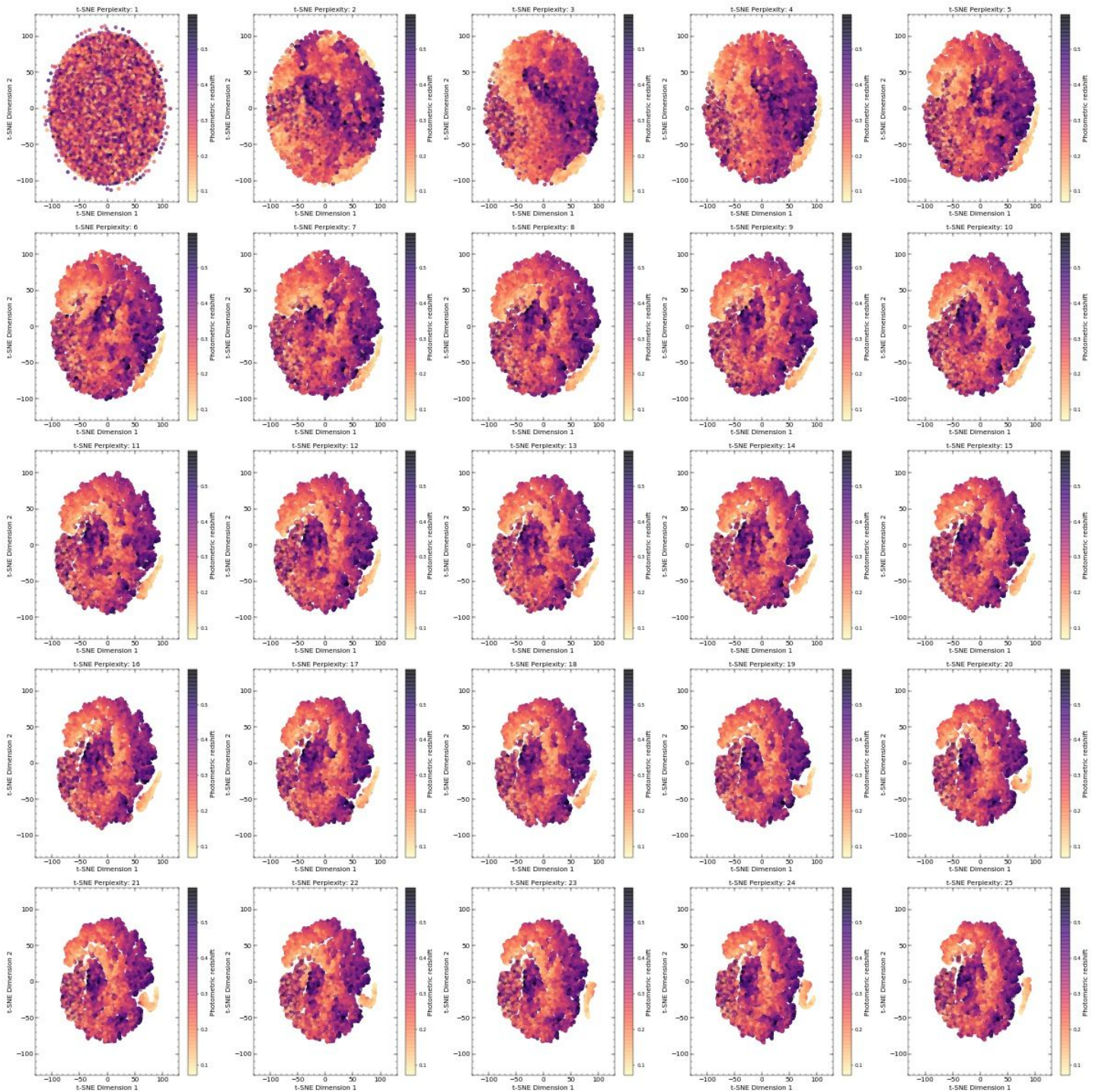


Figure SA14. Two dimensional representation of the feature space in the MWAR training set with no feature scaling applied for a 21 arcseconds search radius with LM filter magnitude-cuts applied using the t-SNE algorithm. The colourbar represents the photometric redshift of galaxies found within the search radius of clusters originally estimated by WHL12. The t-SNE perplexity value relates to the number of nearest neighbours used to compress the dimensionality of the dataset.

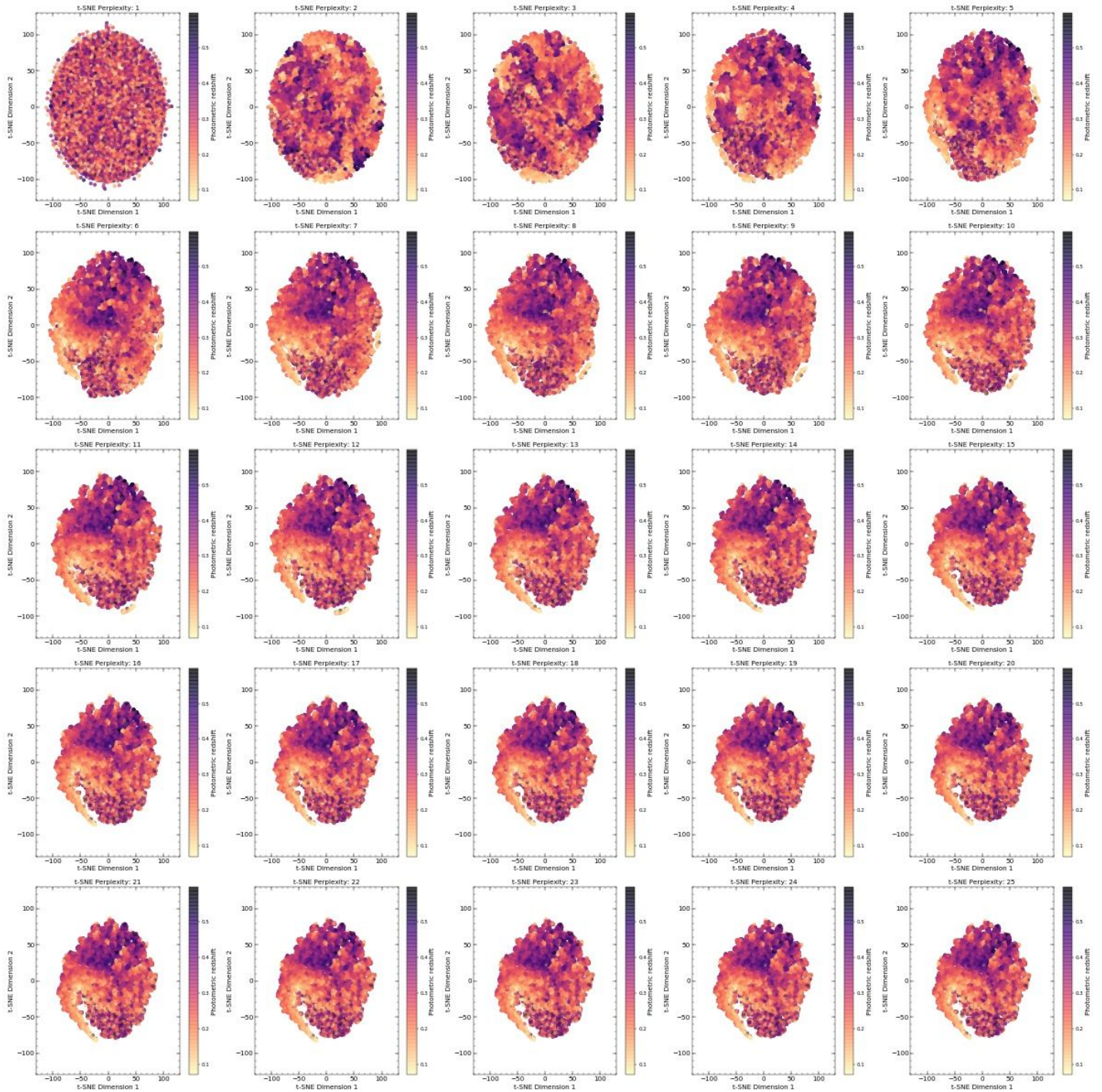


Figure SA15. Two dimensional representation of the feature space in the MWAR training set with no feature scaling applied for a 32 arcseconds search radius with LM-0.5 filter magnitude-cuts applied using the t-SNE algorithm. The colourbar represents the photometric redshift of galaxies found within the search radius of clusters originally estimated by WHL12. The t-SNE perplexity value relates to the number of nearest neighbours used to compress the dimensionality of the dataset.

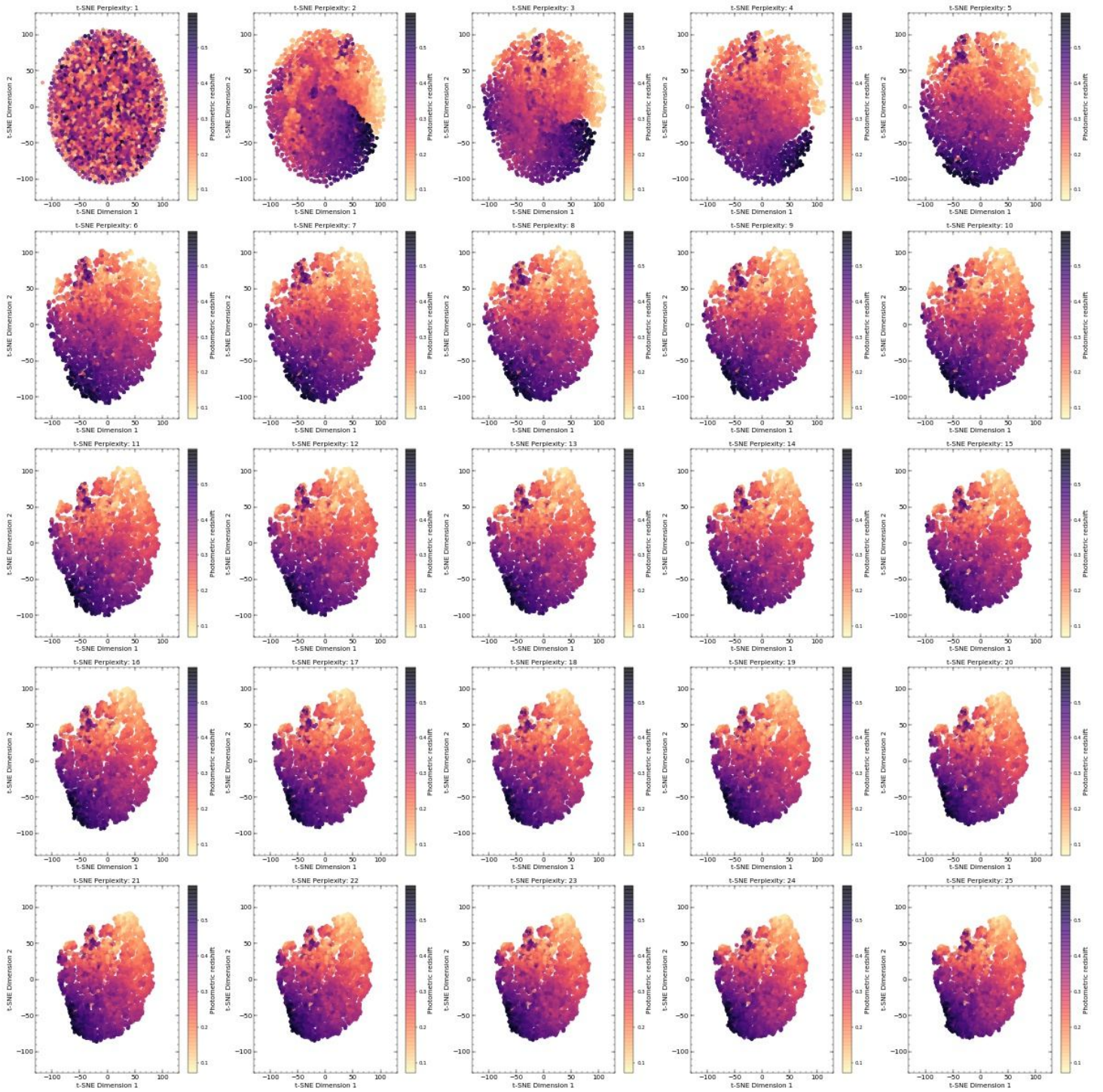


Figure SA16. Two dimensional representation of the feature space in the MWAR training set with feature scaling applied for a 10 arcseconds search radius with LM filter magnitude-cuts applied using the t-SNE algorithm. The colourbar represents the photometric redshift of galaxies found within the search radius of clusters originally estimated by WHL12. The t-SNE perplexity value relates to the number of nearest neighbours used to compress the dimensionality of the dataset.

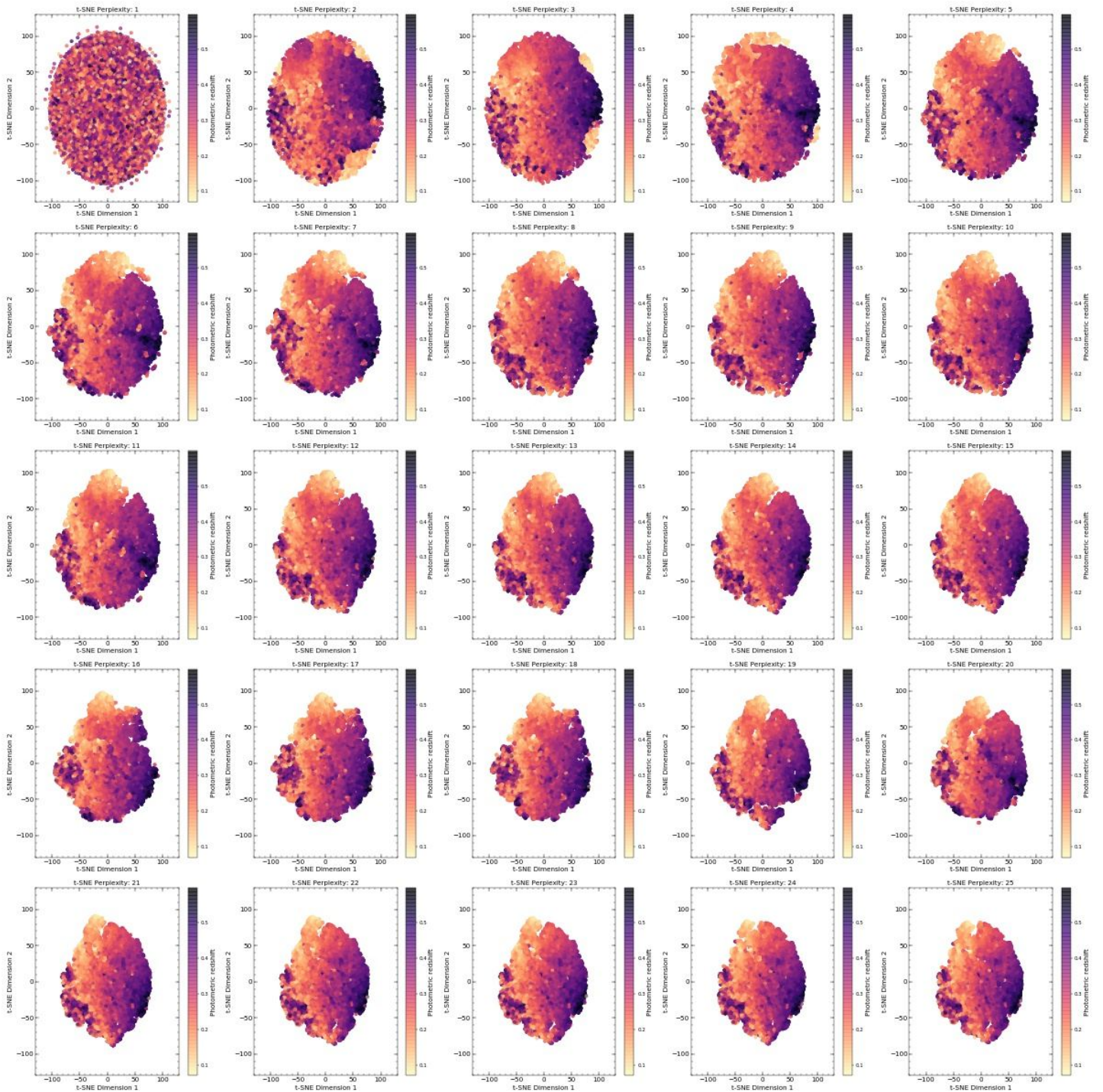


Figure SA17. Two dimensional representation of the feature space in the MWAR training set with feature scaling applied for a 21 arcseconds search radius with LM filter magnitude-cuts applied using the t-SNE algorithm. The colourbar represents the photometric redshift of galaxies found within the search radius of clusters originally estimated by WHL12. The t-SNE perplexity value relates to the number of nearest neighbours used to compress the dimensionality of the dataset.

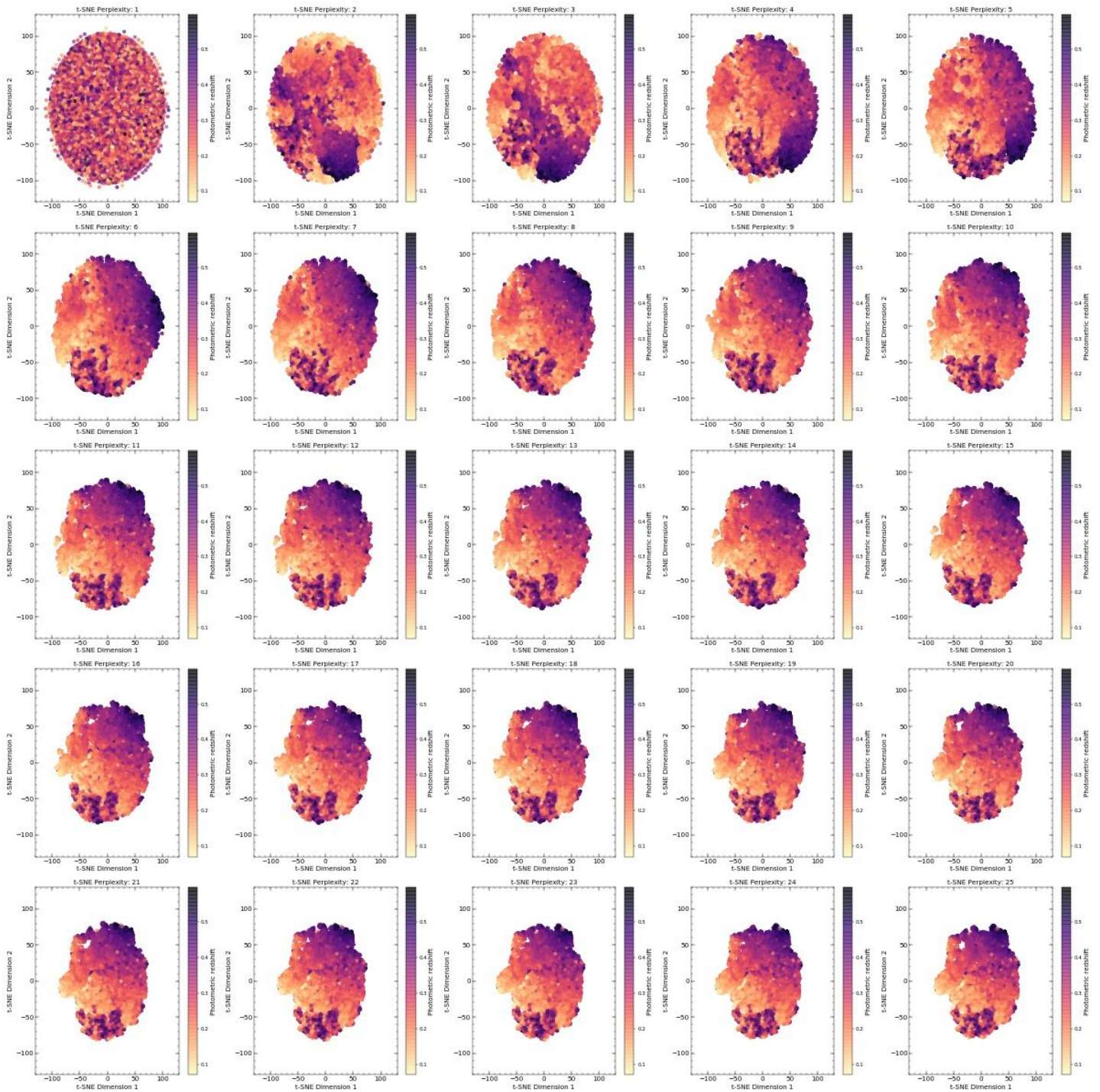


Figure SA18. Two dimensional representation of the feature space in the MWAR training set with feature scaling applied for a 32 arcseconds search radius with LM-0.5 filter magnitude-cuts applied using the t-SNE algorithm. The colourbar represents the photometric redshift of galaxies found within the search radius of clusters originally estimated by WHL12. The t-SNE perplexity value relates to the number of nearest neighbours used to compress the dimensionality of the dataset.

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