Photometric Redshift Realism: A Technique for Reducing Catastrophic Outlier Redshift Estimates in Large-Scale Surveys

M. Wyatt* and J. Singal
Physics Department, University of Richmond, 138 UR Drive, Richmond, VA 23173, USA

13 November 2019

ABSTRACT
We present results of using individual galaxies’ effective redshift probability density information as a method of identifying potential catastrophic outliers in empirical photometric redshift estimation. In the course we develop a method of modification of the redshift distribution of training sets to improve both the baseline accuracy of high redshift (\(z > 1.5\)) estimation as well as catastrophic outlier mitigation. We demonstrate these using two real test data sets and one simulated test data set spanning a wide redshift range (0 < \(z\) < 4). We present these results in the context of “photometric redshift realism” and aim to show that the methods and results presented here can inform a ‘prescription’ that can be applied as a realistic photometric redshift estimation scenario for a hypothetical large-scale survey. We find that with appropriate optimization, we can identify a large percentage (>30%) of catastrophic outlier galaxies while simultaneously incorrectly flagging only a small percentage (<7% and in many cases <3%) of non-outlier galaxies as catastrophic outliers. We find also that our training set redshift distribution modification results in a significant decrease (>10%) in the percentage of outlier galaxies greater than \(z = 1.5\) with only a small increase (<3%) in the percentage of outlier galaxies less than \(z = 1.5\) compared to the unmodified training set. In addition, we find that this modification can in some cases decrease the percentage of incorrectly identified non-outlier galaxies by almost 20%, while in other cases cause only a small (<1%) increase in this metric.

Key words: techniques: photometric – galaxies: statistics – methods: miscellaneous

1 INTRODUCTION
Soon, large scale surveys such as those done by the Large Synoptic Survey Telescope (e.g. Ivezic et al. 2008) will observe up to hundreds of millions of individual galaxies in a limited number of photometric bands. One important quantity that must be determined from the data collected in these surveys is each galaxy’s redshift, which is a key measurement for practically every science goal in extragalactic astrophysics and cosmology.

The traditional spectroscopic method for determining redshifts is time consuming and therefore not practical for upcoming large-scale surveys. Redshift determinations for most galaxies in these surveys will instead rely on a quicker method of redshift estimation termed photometric redshift (abbreviated “photo-z” — see e.g. Salvato, Ilbert & Hoyle (2019) for a recent review). While faster, this method is less certain because it is based on a galaxy’s measured brightness in a smaller number of wider wavelength bands of light compared with the traditional method. A crucial goal of the cosmological survey community is to develop methods of accurate photo-z estimation with well understood errors (e.g. Mandelbaum et al. 2019).

Inaccuracies in photo-z estimation are often viewed in terms of different degrees of “outliers.” Here we follow convention (e.g. Hildebrandt et al. 2010) and define outliers as those galaxies where

\[
O : \left| \frac{z_{\text{phot}} - z_{\text{spec}}}{1 + z_{\text{spec}}} \right| > 0.15, \tag{1}
\]

where \(z_{\text{phot}}\) and \(z_{\text{spec}}\) are the estimated photo-z and actual (spectroscopically determined if available) redshift of the object. For the most inaccurate photo-z estimations the term “catastrophic outliers” (hereafter COs) is often used. Although there is not a standard, universal definition of catastrophic outliers, we use a definition that is typical (e.g. Bernstein & Huterer 2010):

\[
O_c : \left| z_{\text{phot}} - z_{\text{spec}} \right| > 1.0. \tag{2}
\]
Those galaxies whose photo-z estimates are sufficiently close to the actual redshift to not be outliers by the above definition are termed “non-outliers” (NOs).

Because COs can have detrimental effects on the science goals of large scale surveys (e.g. Graham et al. 2018; Hearin et al. 2010; Bernstein & Huterer 2010), mitigating them is therefore a crucial aim. One promising strategy could involve a method of identifying or ‘flagging’ potential catastrophic outliers, so that these galaxies could be excluded or de-weighted in statistical analyses.

One traditional class of photo-z estimation techniques are the so-called “empirical” or “training set” methods, which work by developing a mapping from input parameters to redshift using a training set of galaxies whose redshifts are known, and then applying these mappings to an evaluation set (e.g. Hildebrandt et al. 2010). Empirical techniques encompass a variety of approaches including those involving machine learning. For this analysis we utilize SPIDERz (Support vector classification for IDENTifying redshifts), a custom support vector machine (SVM) type machine learning classification algorithm — an example of an empirical method — which is available to the community. As discussed in Jones & Singal (2017, 2019), SPIDERz has many customizable features, provides accurate redshift estimates on a variety of test data sets, and, crucially for this work, provides an effective redshift probability density function (EPDF) for each galaxy, which comes about naturally as a product of the SVM classification. The SVM works by comparing each possible redshift for a galaxy, divided into a user-inputted number of bins, and ‘voting’ for the more likely option for each pair, which naturally provides an EPDF for the photometric redshift of each galaxy consisting of the total votes for each bin. Because bins that are unlikely to be accurate are paired against each other this method artificially inflates some bins with lower probabilities of being correct. Because of this it is not a true probability density function, but an EPDF, and is sufficient for the methods presented here. Some example EPDFs for individual galaxies are shown in Figure 1. For a single-valued discrete photo-z estimate, SPIDERz chooses the most probable bin of the EPDF.

In a previous work (Jones & Singal 2019) SPIDERz’s naturally occurring EPDFs were noted as a method of potentially identifying and flagging potential COs, because galaxies with multiple widely separated peaks in redshift probability or without a clearly most probable redshift are more likely to be COs than those exhibiting a clear peak in probability, as can be visualized in Figure 1. Preliminary results obtained there were promising, indicating that a relatively large fraction of COs (~50%) could potentially be flagged while only flagging a relatively small fraction of the total NOs (<10%). In this work we will systematically explore the use of EPDFs as a method of flagging COs in photo-z determination using a range of test data sets, and use this in conjunction with a modification of the redshift distribution of training sets in order to provide a realistic potential ‘prescription’ for achieving robust, accurate, and well understood photo-z estimates in future large-scale surveys where the effect of COs — especially at high redshifts — is reduced. Although in a real survey scenario the redshifts of the evaluation sets would be unknown, here we use evaluation test sets with known redshifts in order to compare and test our results.

In §2 we present the test data sets used throughout this work. In §3 we discuss what we will term “photo-z realism:” what some photometric redshift studies have done that could be overly optimistic relative to future large-scale, several-band surveys with a large redshift range — such as LSST. In §4 we describe a novel method for flagging potential outliers

 available from http://spiderz.sourceforge.net with usage documentation provided there

---

**Figure 1.** Three examples of individual real galaxy EPDFs (effective probability vs. redshift) output by a SPIDERz photo-z determination using the COSMOS-reliable-z test data set described in §2 with 2000 training galaxies. The top panel EPDF manifests a single probability peak which indicates a likely reliable photo-z determination. The middle panel shows a doubly peaked EPDF where a redshift of \(z \sim 0.2\) and one of \(\sim 2.8\) are nearly equally probable, and the bottom panel shows an EPDF without clear peaks, both of which could indicate a possibly inaccurate photo-z estimate. EPDFs are discussed in §1.
in estimations and our method for optimizing this. In §5 we explain our novel method for modifying training sets in order to produce fewer outlier galaxies as well as better flagging results using our algorithm. Finally, in §6 we present our prescription for use in a realistic large-scale survey scenario.

## 2 TEST DATA

In order to systematically explore using EPDFs as a method of flagging COS in photo-z determination and the modification of training set distributions introduced above, we require test data sets with a large number of galaxies with photometric data as well as known redshifts over a wide range of redshifts and galaxy types. These sets could consist of either realistically simulated or real observed galaxy data.

One test data set of real galaxies comes from the COSMOS2015 photometric catalog (Laigle et al. 2016) with a separate redshift estimation collection (Ilbert et al. 2009), which contains particularly reliable redshifts derived from a very large number of photometric bands — such a large number over such a large wavelength range that it approaches what we will term a “quasi-spectroscopic redshift.” The COSMOS2015 catalog provides photometry for some galaxies in up to 31 optical, infrared, and UV bands in addition to redshifts calculated from this photometry, so to maximize the reliability of the known quasi-spectroscopic redshifts for our photo-z test purposes, we restrict the use of galaxies to those meeting the following criteria: (i) non-error magnitude values for at least 30 bands of photometry, (ii) for which the \( \chi^2 \) for the Ilbert et al. (2009) redshift estimate is \(< 1\), and (iii) for which the photo-z value from the minimum \( \chi^2 \) estimate is less than 0.1 redshift away from the photo-z value from the peak of the pdf. These galaxies can be considered to have highly reliable previous quasi-spectroscopic redshift estimates. Applying these criteria results in a data set of 58,622 galaxies. We then limited the redshifts to \( z < 4 \) in order to prevent the occurrence of unoccupied bins, which resulted in a total of 58,619 galaxies. These galaxies span an \( i \)-band magnitude range from 27.11 to 19.00 with a median of 24.08. Their redshift distribution is shown in Figure 2.

We will refer to this set as “COSMOS-reliable-z”. For our test purposes, this test data set contains photometry for at least nine optical and infrared bands for each galaxy (\( u, B, V, r, i, z, Y, J, K \)), but for most purposes in this work we restrict photo-z estimation to the use of only five optical bands (\( u, B, r, i, \) and \( z \)).

The top panel of figure 3 shows a baseline photo-z estimate vs. actual redshift for the COSMOS-reliable-z test data set as evaluated with SPIDERz with five photometric bands along with a visualization of the density of galaxies in some of the regions of these plots, we also show the same results with a density visualization in Figure 4.

![Figure 2](image1.png)  
*Figure 2.* The redshift distributions \( [N(z)] \) for the three test data sets used in this analysis. The top graph shows the distribution for the COSMOS-reliable-z test data set (red) and the COSMOSxSpecs test data set (blue). The bottom graph shows the distribution for the BzK simulated test data set, where the lighter shade indicates the full data set and the darker shade shows the distribution after limiting all magnitudes at 28, as discussed in §2.

![Figure 3](image2.png)  
*Figure 3.* Estimated photo-z vs. actual redshift as evaluated with SPIDERz with five photometric bands using COSMOS-reliable-z (top) and COSMOSxSpecs (bottom) test data, with training sets as modified using the methods discussed in §5. Show in red are erroneously flagged NOs and shown in green are correctly flagged OS using the optimal flagging parameters found with the flagging method discussed in §4. Those points outside of the diagonal lines are outlier photo-z estimates as defined by equation 1. Given the density of galaxies in some of the regions of these plots, we also show the same results with a density visualization in Figure 4.
of which galaxies are flagged under the procedure discussed in §4.

In order to form another test data set of real galaxies with a somewhat different redshift distribution, we utilize a combination of several spectroscopic redshift collections (Hasinger et al. 2018; Salvato 2018; Momcheva et al. 2016) with the COSMOS catalog of magnitudes (Laigle et al. 2016), which together forms a data set complete with 9 photometric magnitude bands (of which we generally utilize five for photo-z estimation — $u, g, r, i,$ and $z$). For spectroscopic redshifts we use the reported "best available" value and eliminate galaxies which are flagged as having their redshift determined from photometry or as being stars. Limiting the redshifts to $z < 4$, this results in a test data set of 25,831 galaxies. These galaxies span an $i$-band magnitude range from 26.67 to 18.00 with a median of 22.41. We will refer to this set as "COSMOSxSpecs." Figure 2 shows the redshift distribution for this test data set, while the bottom panels of figures 3 and 4 show a baseline photo-z estimate vs. actual redshift for the as evaluated with SPIDERz with five photometric bands along with a visualization of which galaxies are flagged under the procedure discussed in §4.

For a still different redshift and galaxy type distribution, we utilize simulated data from the BzK Deep lightcone mock catalog, with over 41 million simulated galaxies (Merson 2013), as discussed in §3. We utilize the simulated photometric observations in the $u, g, r, i,$ and $z$ bands for photometric redshift estimation tests. This data in its raw form resulted in almost no CO photo-z estimates, which is itself an indication of (i) the reliability of the SPIDERz photo-z determination, and (ii) the potential perils of photo-z studies using simulated data. In order to make the data set more usable for this investigation, which requires some CO estimates, it was necessary to add simulated noise. We altered each magnitude value by a random positive or negative amount, distributed evenly, between zero and 15%. We also removed all galaxies dimmer than magnitude 28 in any band in order to better reflect real large-scale survey data sets, resulting in a test data set of over 10 million galaxies. The redshift distribution before and after these alterations is shown in Figure 2.

3 PHOTO-Z REALISM

Many photometric redshift studies have been non-representative to varying degrees of the challenges that will arise from upcoming large-scale surveys with a limited number of photometric bands and for which redshift estimates will be needed for galaxies spanning a large range of redshifts. In a realistic scenario, it will be vital for e.g. weak lensing science goals to perform photo-z estimations using just the available measured magnitudes in several optical bands and to accurately estimate redshifts for galaxies extending to $z \sim 2$ and beyond. Importantly, we note that galaxies with redshifts higher than this, even if not directly the most relevant to weak lensing studies, are important because they can be mis-classified as lower redshift galaxies if COs, thus affecting other science goals. Because of this, we
believe that galaxies extending to redshifts $z \sim 4$ should be included in photometric redshift studies. In addition, in realistic scenarios the training set will be very small compared to the potential evaluation set, so we believe that photometric studies should utilize a training set that is significantly smaller in number than the evaluation set.

Here we aim to show the potential limits of optimistic photo-$z$ studies in order to motivate a prescription which can be applied to photo-$z$ estimations in future large-scale surveys in order to mitigate the effects of CO estimates.

While LSST will provide six bands of photometric data, some photometric redshift studies (e.g. Hildebrandt et al. 2010) have utilized more, in some cases many more, bands extending into infrared wavelengths. The presence or absence of additional infrared bands has a significant effect on photo-$z$ estimation accuracy. Figure 5 highlights an example of the degradation baseline photo-$z$ estimation when reducing from nine bands extending into the infrared to five optical-only bands.

There are also issues with generalizing photo-$z$ study results obtained with simulated data to real galaxy data scenarios. For example, using the BzK simulated test data set — even with the added noise and magnitude limit as discussed in §2 — with SPIDERz and five bands still results in an unrealistically accurate photo-$z$ estimation, as can be seen in Table 1.

Another potential scenario for large-survey photo-$z$ estimation — unless measures are taken specifically to avoid it — is one in which the set of galaxies with spectroscopic redshifts available — i.e. those which can comprise a training set in empirical photo-$z$ estimation — is not representative of the galaxies for which photo-$z$ estimations are desired. To test this, we use a training set comprised of galaxies from one test data set and then evaluate on a subset of galaxies from a different test data set. Then, in an attempt to improve the estimation in a way that would be possible if the redshifts of the evaluation set were truly unknown, we alter the training set to match the distribution of the $i$-band magnitudes by bins of 0.1 of the evaluation set. Despite the matching $i$-band distributions the photo-$z$ accuracy is poor in this situation, as shown in Table 1. These ‘mismatched’ training and evaluation sets perform very poorly even when using BzK test data as a training set, despite BzK data having a redshift distribution which covers a much larger range compared to the two real-galaxy COSMOS test data sets, and thus potentially contains more combinations of galaxy type and redshift than the COSMOS sets. This indicates that training and evaluation set mismatches in galaxy types and other characteristics creates problems for empirical photo-$z$ approaches. Because of this, we believe that it is essential that for future large scale surveys, the photo-$z$ training set be formed from a subset that is as representative as possible of the overall galaxy sample for which photo-$zs$ are desired. We return to this issue in §6.

4 Flagging and Flagging Optimization

Our method for using EPDFs to flag potential COs uses the heights and locations of peaks within the EPDFs, some examples of which are shown in Figure 1. The flagging procedure first identifies the maximum peak of a galaxy’s EPDF, which corresponds to the most likely redshift and therefore the discrete single-valued photo-$z$ estimate. If a second peak in redshift is above a predetermined minimum fraction of the height of the primary peak ($p_{f, \text{min}}$) and farther than a predetermined minimum distance away in redshift from the maximum peak ($\Delta z_{\text{peak,min}}$), the galaxy is flagged as a potential outlier. This idea was explored preliminarily in a previous work (Jones & Singal 2019) and was found to be a promising strategy. Here, rather than attempting to find one optimal combination of these parameters for all data configurations, we instead optimize these parameters for each data configuration individually. The ultimate goal...
of the optimization of these parameters is to flag more COs while simultaneously flagging fewer NOs. However, simply increasing the ratio of flagged COs to flagged NOs generally results in an unacceptable number of flagged NOs in high redshift ($z > 1.5$) bins. Due to the importance of galaxies of high redshift in many science goals, we wished to avoid this. Keeping this in mind, we utilize the following penalty function:

$$f_P = \frac{O_{\text{flagged}}^{\text{non}} > 1.5}{O_{\text{total}}^{\text{non}} > 1.5}$$

This penalty function is then applied to a “goodness function:"

$$f_G = \frac{O_{\text{flagged}}}{O_{\text{total}}},$$

which is simply the ratio of the total percentage of flagged COs to the penalty function. Each combination of parameters $p_f, min$ and $\Delta z_{\text{peak}, min}$ is then checked to maximize the goodness function, with the addition of a minimum cutoff of 30% flagged COs. Some visualizations of the effectiveness of flagging in this way can be seen in Figure 6 and some metrics are presented in Table 1. We see that this method of flagging according to optimized $p_f, min$ and $\Delta z_{\text{peak}, min}$ values can effectivly flag a large portion of COs while simultaneously flagging a much smaller percentage of NOs, and relatively small percentages of NOs even in high redshift bins, depending on the test data set.

5 TRAINING SET MODIFICATION

Here we explore modifying the redshift distribution of training sets in order to reduce the number of COs and to increase the efficacy of flagging. An example of one modification for each data set can be seen in Figure 7.

The modification of training sets involves creating a distribution of galaxies in redshift across the entire range of redshifts that is less heavily skewed toward lower redshifts to create what we will label an “evened” training set. In other words, we seek to increase the fraction of galaxies in the training set which are in the highest (usually under-populated) redshift bins. With an eye toward formulating a useful prescription for large-scale surveys as discussed in §6.
we do this in the following manner: we first randomly select a separate 8,000 and 10,000 galaxies from the full test data set. 2,000 galaxies are also randomly selected from the 8,000 galaxy subset to be used as a fiducial sample only for its redshift distribution. Because of the different redshift distributions of the two test data sets, the “evened” modified training set is then created slightly differently in the two cases. For the COSMOS-reliable-z test data set, the distribution of the 2,000 galaxy set is used from \( z = 0 \) to \( z = 1 \), then the number of galaxies in the \( z = 0.9 - 1.0 \) bin is used as the value for each successive bin until a lower number of galaxies in a bin is found, usually around \( z = 2 \). The 8,000 galaxy distribution is used for all redshift bins above this.

In the case of the COSMOSxSpecs test data set, 2,000 galaxy training set distribution is used from \( z = 0 \) to \( z = 1 \) and the 8,000 galaxy training set distribution is used for the remaining bins. These modifications are visualized in Figure 7. We see improvements by purposefully altering the training set to have a different redshift distribution than that of the evaluation set.

These modifications result in improvements in both the general reliability of the photo-z determinations (e.g. percentages of outliers and COs), especially at high \( (z > 1.5) \) redshifts, and improvements in the reliability of flagging. We see improvements in the reliability of flagging (visualized in Figure 6) in the form of significant reductions in the numbers of flagged NOs, especially at higher redshifts, as well as increases in the percentages of flagged COs. Improvements in the general photo-z reliability are visualized in Figure 8. We see a significant increase in the numbers of NOs in the range \( z > 1 \) as well as decrease the number of COs in the range \( z > 1 \). The latter is particularly interesting because it seems to go against the convention of machine learning in which it is standard to desire a training set which is as closely representative to the evaluation set as possible. As shown in Figure 6, although in the case of the COSMOS-reliable-z test data set the percentage of flagged NOs resulting from the training set modification does not decrease like in the case of the COSMOSxSpecs test data set but rather remains very similar, it is important to note that the percentage of non-flagged NOs above redshift \( z = 1 \) of the total number of galaxies in each bin is increasing. This is because the modification causes more NOs at higher redshifts, although in the case of COSMOS reliable-z, the percentage...
of flagged NOs remains similar, the number of non-flagged NOs is increased. Several metrics relevant to both of these categories of improvements are also shown in Table 1.

It is important to note that these improvements do not come about simply by virtue of having a higher number of high redshift galaxies, but by having a higher fraction of high redshift galaxies. This modification does not involve increasing the number of high redshift galaxies compared to the baseline training set distribution, but rather removes some lower redshift galaxies as seen in Figure 7 resulting in a training set that is actually smaller in number (although to compose this training set an equal number of galaxies is needed to begin with, as discussed in §6).

6 PRESCRIPTION

In section §3 we demonstrated that empirical photo-z results are poor when a training set is not drawn from the same survey data as the evaluation set, while in §4 and §5 we showed the possible utility of flagging potential COs and modifying training set redshift distributions when those training sets are drawn from the same survey as the evaluation set. Based on these conclusions and the results discussed in those sections and as seen e.g. in Table 1, here we present a prescription for achieving useful photo-z estimation with reduced high redshift COs, efficiently flagged remaining COs, and increased high redshift non-flagged NOs in one realistic photometric estimation scenario for a hypothetical large-scale survey with a limited number of photometric bands, starting from a large set (potentially millions) of observed galaxies for which redshifts are unknown and desired to be estimated:

(i) Obtain spectroscopic redshifts for a random subset of 18,000 galaxies

(ii) Set aside a randomly selected 10,000 of these galaxies, to be used as a testing set for flagging parameter determination.

(iii) Modify the remaining 8,000 galaxies to have a more even redshift distribution by removing some lower redshift galaxies, as described in §5.

(iv) Train on the modified training set and evaluate on the 10,000 galaxy subset selected earlier to find the optimal flagging parameters which maximize the goodness function as discussed in §4.

(v) Estimate the redshifts of the millions of galaxies for which photometric redshifts are desired using the modified training set and flag potential COs using the found optimal parameters from step (iv).

This prescription is operable in principle for any campaign in which photo-zs are determined with any empirical method which provides probability distribution over redshift information for each galaxy.

Steps (iii) and (v) are necessary because the optimization routine works by using known redshifts in order to calculate the percentages of flagged NOs and COs. Because in real situations redshifts are not known for the evaluation sets, it is necessary to create a smaller evaluation set with spectroscopic redshifts in order to optimize the parameters for the full evaluation. In the case of the COSMOS-reliable-z data set, we find that 10,000 galaxies is the smallest number that produces consistent results in terms of finding the same optimized flagging parameters $p_{f,\text{min}}$ and $\Delta z_{\text{peak,\min}}$, while in the case of the COSMOSxSpecs test data set we find that even 10,000 galaxy evaluation sets do not find consistent optimal parameters across multiple randomizations. We have identified the cause of this as being certain galaxies which have a difference in a very small number (in some cases 1) of “votes” in certain bins of their EPDFs. This is ultimately based on the difference in scaling of the magnitudes that occurs when galaxies which are extremal in magnitude do or do not make it into the randomly selected evaluation set due to the small number of galaxies in high redshift bins in the COSMOSxSpecs test data set.

The question could be raised as to whether it is realistic to expect that spectroscopic redshifts could be obtained for all 18,000 randomly selected galaxies. In order to explore this complication, we will assume (here for simplicity) that spectroscopic redshifts for the dimmest 10% (in $i$-band) of the 18,000 selected galaxies cannot be determined. As examples, the resulting effect of this on the redshift distributions of our real galaxy test data sets are shown in Figure 9. Figures 10 and 11 compare the baseline and modified “evened” cases for both COSMOS test data sets with the dimmest 10% (in $i$-band) galaxies were removed from the training sets. Table
we show results for determinations with 9 photometric bands, with we show results from determinations where some training set's have been modified to have a more shows several metrics for various cases of determinations we dis-

we show results from determinations where the dimmest 10% of galaxies (in () via analyses with SPIDERz, a custom sup-

Table 1.

| Test Data | O% | O%>1.5 | O%>1.5 | NO% | NO%>1.5 | NO%>2.0 | %NO%>1.5 | pf_min | Szpeak_min |
|-----------|----|--------|--------|------|---------|---------|----------|--------|------------|
| xSpecs    | 7.61 | 1.60 | 30.35 | 31.72 | 1.09 | 4.90 | 10.67 | 66.24 | 0.92 | 1.20 |
| reliable-z | 4.10 | 1.12 | 5.97 | 34.08 | 0.47 | 0.78 | 2.03 | 93.29 | 0.95 | 1.20 |

5 BAND:

Mismatched

| Train. Set | Eval. Set |
|-----------|-----------|
| xSpecs    | reliable-z | 15.13 | 5.94 | 72.32 | 36.71 | 6.60 | 31.68 | 52.94 | 18.91 | 0.94 | 1.10 |
| xSpecs    | BalK      | 15.18 | 4.66 | 39.22 | 48.84 | 5.12 | 10.84 | 17.27 | 54.19 | 0.95 | 1.20 |
| reliable-z | xSpecs    | 35.57 | 10.32 | 23.53 | 31.14 | 7.47 | 5.67 | 11.05 | 72.13 | 0.90 | 1.10 |
| reliable-z | BalK      | 35.90 | 13.33 | 43.94 | 38.35 | 7.21 | 8.95 | 12.89 | 51.05 | 0.91 | 1.10 |
| xSpecs    | 19.21 | 5.38 | 58.24 | 32.52 | 1.73 | 9.68 | 32.74 | 37.72 | 0.90 | 1.20 |
| BalK      | reliable-z | 21.11 | 5.54 | 75.72 | 30.18 | 0.93 | 24.48 | 74.36 | 18.33 | 0.80 | 1.10 |

BzK Runs

| Train. Set | Eval. Set |
|------------|-----------|
| xSpecs    | Baseline | 7.34 | 0.73 | 8.41 | 38.18 | 0.47 | 0.57 | 0.64 | 91.07 | 0.94 | 1.20 |
| xSpecs    | Modified | 7.08 | 0.64 | 8.98 | 39.21 | 0.44 | 0.61 | 0.70 | 82.25 | 0.91 | 1.20 |

Train. Mods.

|i-band Lim.|
| Test Data | Mod. |
|-----------|------|
| xSpecs    | Baseline | 11.26 | 3.12 | 53.85 | 30.13 | 3.08 | 31.12 | 28.35 | 40.57 | 0.94 | 0.90 |
| xSpecs    | “Evened” | 11.55 | 3.05 | 44.78 | 32.69 | 4.04 | 14.11 | 24.31 | 49.08 | 0.91 | 0.90 |
| reliable-z | Baseline | 10.18 | 2.78 | 27.61 | 31.62 | 1.09 | 3.01 | 6.62 | 70.22 | 0.95 | 1.10 |
| reliable-z | “Evened” | 10.52 | 2.84 | 18.47 | 31.13 | 2.33 | 2.74 | 5.36 | 79.29 | 0.94 | 1.20 |

Tallied for each determination are the percentage of galaxies which are outliers (O%), the percentage of galaxies which are catastrophic outliers (O%>1.5), the percentage of galaxies above z = 1.5 which are outliers (O%>1.5), the percentage of catastrophic outliers which are flagged (O%>1.5), the percentage of non-outliers which are flagged (NO%>1.5), the percentage of non-outliers above z = 1.5 which are flagged (NO%>1.5), the percentage of galaxies above z = 2.0 which are flagged (NO%>2.0), the percentage of galaxies above z = 1.5 which are non-flagged non-outliers (%NO%>1.5), and the optimal parameters used for flagging (pF_min and S_zpeak_min) determined using the methods discussed in §4. Relevant to §3 we show results for determinations with 9 photometric bands, with the simulated BzK test data set, and “mismatched” runs which use training sets drawn from different test data sets than the paired evaluation set. Relevant to §5 we show results from determinations where some training sets have been modified to have a more balanced “evened” redshift distribution. Relevant to §6 we show results from determinations where the dimmest 10% of galaxies (in i-band flux) have been eliminated from the training set. Further details about these determinations are presented in the associated text section.

As in the non-spectro-magnitude-limited case, we find that using the COSMOSxSpecs test data set even with a 10,000 galaxy evaluation set, we are not able to find consistently the same optimal parameters as with the full evaluation set. However, in the case of COSMOS-reliable-z we are able to consistently find optimal parameters which are close to those of the full evaluation set, but not as close as with the non-magnitude-limited case.

7 CONCLUSIONS

This work presented two intertwined strategies for mitigating the effects of CO photo-z estimates for large-scale surveys with a small number of photometric bands available: (i) using EPDFs to flag potential COs and (ii) modifying training sets to reduce high-redshift COs and flagged NOs. We evaluated these strategies using galaxy test data sets described in §2 via analyses with SPIDERz, a custom support vector machine package for photo-z estimation which naturally outputs an EPDF for each galaxy. In §3 we discussed the challenges that result from having only a limited number of optical bands with which to carry out photo-z estimates, as well as those arising when a training set is not drawn from the same galaxy sample as the evaluation set for which photo-zes are desired. In this analysis we evaluate assuming five photometric bands (u, g or B as available, r, i and z) available for a hypothetical large-scale survey. While
this may be overly pessimistic for LSST in particular which will feature six photometric bands (roughly the five preceding and a near-infrared band in addition), in reality many galaxies will not have photometry in all six bands and/or will report large photometric errors in one or more bands (e.g. Soo et al. 2018). Therefore we view five photometric bands as neither overly optimistic or overly pessimistic. In §4 we presented the strategy for flagging potential COs and showed some metrics of performance, indicating that a high percentage of COs can be flagged while simultaneously flagging a much lower percentage of NOs. In §5 we showed that relatively simple modifications to the redshift distributions of training sets can improve both high redshift photo-$z$ accuracy and flagging accuracy. Based on these results, we present a prescription in §6 for applying these strategies to photo-$z$ estimation in a large-scale extragalactic survey to achieve more accurate photo-$z$ estimates and to mitigate the effects of CO estimates.

ACKNOWLEDGEMENTS

Based in part on observations made with the NASA/ESA Hubble Space Telescope, obtained from the Data Archive at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555. Funding for the DEEP2 survey has been provided by NSF grants AST-0071048, AST-0071198, AST-0507428, and AST-0507483. Some of the data presented herein were obtained at the W. M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California and the National Aeronautics and Space Administration. The Observatory was made possible by the generous financial support of the W. M. Keck Foundation. The DEEP2 team and Keck Observatory acknowledge the very significant cultural role and reverence that the summit of Mauna Kea has always had within the indigenous Hawaiian community and appreciate the opportunity to conduct observations from this mountain.

REFERENCES

Bernstein, G. & Huterer, D. 2010, MNRAS, 401, 1399
Graham, M. et al. 2010, AJ, 155, 1
Hasinger, G. et al. 2018, ApJ, 858, 77
Hearin, A. et al. 2010, ApJ, 720, 1351
Hildebrandt, H. et al. 2010, A&A, 523, 832
Ilbert, O. et al. 2009, ApJ, 690, 1236
Ivezic, Z. et al. 2008 (arXiv:0805.2366)
Jones, E. & Singal, J. 2017, A&A, 600, A113
Jones, E. & Singal, J. 2019, PASP, In Press (arXiv:1709.01576)
Laigle, C. et al. 2016, ApJS, 224, 24
Mandelbaum, I. et al. 2019, White Paper: “The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document” (arXiv:1809.01669)
Merson, A. et al. 2013, MNRAS, 429, 556
Momcheva, I. et al. 2016, ApJS, 225, 27
Salvato, M. 2018, priv. comm.
Salvato, M., Ilbert, O. & Hoyle, B. 2019, Nat. Astron., 3, 212
Spergel, D. et al. 2015 (arXiv:1503.03757)
Soo, J. et al. 2018, MNRAS, 475, 3613

This paper has been typeset from a TeX/LaTeX file prepared by the author.