Exhausting the Information: Novel Bayesian Combination of Photometric Redshift PDFs

Matias Carrasco Kind* and Robert J. Brunner
Department of Astronomy, University of Illinois, Urbana, IL 61820 USA

5 June 2014

ABSTRACT

The estimation and utilization of photometric redshift probability density functions (photo-z PDFs) has become increasingly important over the last few years and currently there exist a wide variety of algorithms to compute photo-z's, each with their own strengths and weaknesses. In this paper, we present a novel and efficient Bayesian framework that combines the results from different photo-z techniques into a more powerful and robust estimate by maximizing the information from the photometric data. To demonstrate this we use a supervised machine learning technique based on random forest, an unsupervised method based on self-organizing maps, and a standard template fitting method but can be easily extend to other existing techniques. We use data from the DEEP2 and the SDSS surveys to explore different methods for combining the predictions from these techniques. By using different performance metrics, we demonstrate that we can improve the accuracy of our final photo-z estimate over the best input technique, that the fraction of outliers is reduced, and that the identification of outliers is significantly improved when we apply a Naïve Bayes Classifier to this combined information. Our more robust and accurate photo-z PDFs will allow even more precise cosmological constraints to be made by using current and future photometric surveys. These improvements are crucial as we move to analyze photometric data that push to or even past the limits of the available training data, which will be the case with the Large Synoptic Survey Telescope.

Key words: methods: data analysis – methods: statistical – surveys – galaxies: distances and redshifts – galaxies: statistics.

1 INTRODUCTION

Spectroscopic galaxy surveys have played an important role in understanding the origin, composition, and evolution of our Universe. Surveys like the Sloan Digital Sky Survey (SDSS; York et al. 2000), WiggleZ (Drinkwater et al. 2010), and BOSS (Dawson et al. 2013) have imposed important constraints on the allowed parameter values of the standard cosmological model (e.g., Percival et al. 2010; Blake et al. 2011; Sánchez et al. 2013). However, spectroscopic measurements are considerable more expensive to obtain than photometric data, they are more likely to suffer from selection effects, and they provide much smaller galaxy samples per unit telescope time. As a consequence, current ongoing and future galaxy surveys like the Dark Energy Survey (DES[1]) and the Large Synoptic Survey Telescope (LSST[2]) are pure photometric surveys. These surveys will enable cosmological measurements on galaxy samples that are currently at least a hundred times larger than comparable spectroscopic samples, that have relatively simple and uniform selection functions, that extend to fainter flux limits and larger angular scales, thereby probing much larger cosmic volumes and will photometrically detect galaxies that are too faint to be spectroscopically observed.

With the growth of these large photometric surveys, the estimation of galaxy redshifts by using multi band photometry has grown significantly over the last two decades. As a result, a variety of different algorithms for estimating photo-z 's based on statistical techniques have been developed (see, e.g., Hildebrandt et al. 2010; Abdalla et al. 2011; Sánchez et al. 2014, for a review of current photo-z techniques). Over the last several years, particular attention has been focused on techniques that compute a full probability density function (PDF) for each galaxy in the sample. A photo-z PDF contains more information than a single photo-z estimate, and the use of photo-z PDFs has been shown to improve the accuracy of cosmological measurements (e.g., Mandelbaum et al. 2008; Myers et al. 2009; Jee et al. 2013).

Photo-z techniques can be broadly divided into two categories: spectral energy distribution (SED) fitting, and training based algorithms. Template fitting approaches (see e.g., Benítez 2000; Bolzonella et al. 2000; Feldmann et al. 2006; Ilbert et al. 2006; Assef et al. 2010) estimate photo-zs by find-

* E-mail: mcarras2@illinois.edu
[1] http://www.darkenergysurvey.org/
[2] http://www.lsst.org/lsst/

ing the best match between the observed set of magnitudes or colors, and the synthetic magnitudes or colors taken from the suite of templates that are sampled across the expected redshift range of the photometric observations. This method is often preferred over empirical techniques as they can be applied without obtaining a high-quality spectroscopic training sample. However, these techniques do require a representative sample of template galaxy spectra, and they are not exempt from uncertainties due to measurement errors on the survey filter transmission curves, mismatches when fitting the observed magnitudes or colors to template SEDs, and color–redshift degeneracies. The use of training data that include known redshifts can also improve these predictions (e.g., Ilbert et al. 2006; Newman et al. 2013b). On the other hand, machine learning methods have been shown to have similar or even better performance (e.g., Collister & Lahav 2004; Carrasco Kind & Brunner 2013a) when the spectroscopic training sample is populated by representative galaxies from the photometric sample.

Machine learning methods have the advantage that it is easier to include extra information, such as galaxy profiles, concentrations, or different modeled magnitudes within the algorithm. However, they are only reliable within the limits of the training data, and one must exercise sufficient caution when extrapolating these algorithms. These techniques can be sub-categorized into supervised and unsupervised machine learning approaches. For supervised techniques (e.g., Connolly et al. 1995; Brunner et al. 1997; Collister & Lahav 2004; Wadadekar 2005; Ball et al. 2008; Lima et al. 2008; Freeman et al. 2009; Gerdes et al. 2010; Carrasco Kind & Brunner 2013a), the input attributes (e.g., magnitudes or colors) are provided along with the desired output (e.g., redshift). This training information is directly used by the algorithm during the learning process. In this case, the redshift information from the training set *supervises* the learning process and decisions are made by using this information. On the other hand, unsupervised machine learning photo-z techniques (e.g., Geach 2012; Way & Klose 2012; Carrasco Kind & Brunner 2014a) are less common as they do not use the desired output value (e.g., redshifts from the spectroscopic sample) during the training process. Only the input attributes are processed during the training, leaving aside the redshift information until the evaluation phase.

Given the importance of these photo-z PDFs, there is a present demand to compute them as efficiently and accurately as possible. Additional requirements include the need to understand the impact of systematics from the spectroscopic sample on the estimation of these PDFs (e.g., Oyaizu et al. 2008; Cunha et al. 2012a,b), and to maximally reduce the fraction of catastrophic outliers (e.g., Gorecki et al. 2014). Considerable effort has, therefore, been put into both the development of different techniques and the exploration of new approaches in order to maximize the efficacy of photo-z PDF estimation. Yet, the combination of multiple, independent photo-z PDF techniques has remained under explored (e.g., Carrasco Kind & Brunner 2013b; Dahlen et al. 2013).

In this paper we extend our previous exploratory work in combining machine learning techniques with template fitting methods (Carrasco Kind & Brunner 2013b) to explicitly address this issue by presenting a novel Bayesian framework to combine and fully exploit different photo-z PDF techniques. In particular, we show that the combination of a standard template fitting technique with both a supervised and an unsupervised machine learning method can improve the overall accuracy over any individual method. We also demonstrate how this combined approach can both reduce the number of outliers and improve the identification of catastrophic outliers when compared to the individual techniques. Finally, we show that this methodology can be easily extended to include additional, independent techniques and that we can maximize the complex information contained within a photometric galaxy sample.

This paper is organized as follows. In Section 2 we present the algorithms used in this work to generate the individual photo-z PDF estimates and we provide a brief description on their individual functionality. We describe, in Section 3, the different Bayesian approaches by which different photo-z techniques are combined. Section 4 introduces the data sets employed to test this Bayesian approach taken from the SDSS and DEEP2 surveys. In Section 5 we present the main results of our combination approach and compare these results to those from the individual photo-z PDF methods. In Section 6 we discuss the application of a Naïve Bayes combination technique for outlier detection. In Section 7 we conclude with a summary of our main points and a more general discussion of this new approach.

2 PHOTO-Z METHODS

To develop and test our combination framework, we consider three, distinct photo-z PDF estimation techniques; we briefly discuss each one of them in this section. We make the reasonable assumption that these three techniques are independent in their nature where two of these methods implement machine learning algorithms. The first method is a supervised machine learning technique we have published called `TPZ` (Trees for Photo-Z, Carrasco Kind & Brunner 2013a, hereafter CB13), which uses prediction trees and a random forest to produce probability density functions. The second method is an unsupervised technique we have published called `SOM`z (Carrasco Kind & Brunner 2014a, hereafter CB14), which uses self organizing maps (SOM) and a random atlas to produce a probability density function. We have recently incorporated these two implementations into a new, publicly available and growing photo-z PDF prediction framework called `MLZ`[3] (Machine Learning for photo-Z).

The third method is a Bayesian template fitting technique based on `BPZ` (Bayesian Photometric Redshifts; Benítez 2000), which fits spectral energy density templates from a preselected library to an observed set of measured flux values. Taken together, these three methods span the three standard published approaches in computing photo-zs in the literature. Any new method would, very likely, be functionally similar to one of these three methods; therefore, any of these three methods could in principle be replaced by a similar method to avoid redundancy. This can be most easily demonstrated for template fitting methods, where an additional set of photo-z estimations can be utilized by adopting a different template library (e.g., Dahlen et al. 2013). In this particular case, the underlying code is essentially unchanged, but the photo-z results will change as different spectral libraries are adopted.

[3] http://lcdm.astro.illinois.edu/code/mlz.html

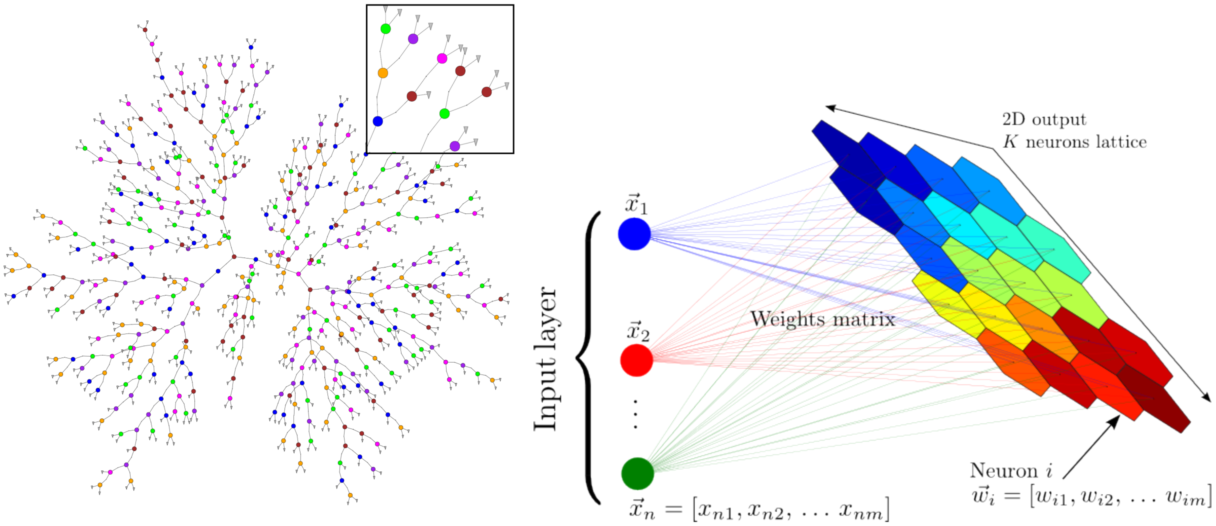

Figure 1. *Left*: A simplified example of a binary prediction tree plotted radially, taken from CB13. The initial node is close to the center of the figure; each node is subdivided and the splitting process terminates when a pre-defined stopping criterion is reached. Individual colors represent a unique variable (e.g., a magnitude like g or r, or a color like $g - r$) used to split an individual node. Each leaf node provides a specific prediction based on the information contained within that terminal node (gray triangles in the figure). The subpanel highlights a specific branch of the tree at higher resolution for additional clarity. *Right*: A schematic representation of a self organized map, taken from CB14. The training set of n galaxies is mapped onto a two-dimensional lattice of K neurons that are represented by vectors containing the weights for each input attribute. Note that the galaxies and the weight vectors are of the same dimension m, and that one neuron can represent more than one training galaxy. The colors used in the map encode the target property from the galaxies grouped within that cell.

2.1 TPZ

TPZ (CB13) is a parallel, supervised algorithm that uses prediction trees and random forest techniques (Breiman et al. 1984; Breiman 2001) to produce photo-z PDFs and ancillary information for a sample of galaxies. Among the different non-linear methods that are used to compute photometric redshifts, prediction trees and random forests are one of the simplest yet most accurate techniques. Furthermore, they have been shown to be one of the most accurate algorithms for low as well as high multi-dimensional data (Caruana et al. 2008).

Prediction trees are built by asking a sequence of questions that recursively split the data into two branches until a terminal leaf is created that meets a pre-defined stopping criterion (e.g., a minimum leaf size or a maximum rms within that leaf). The small region bounding the data in the terminal leaf node represents a specific subsample of the entire data that all share similar characteristics. A comprehensive predictive model is applied to the data within each leaf that enables predictions to be rapidly computed in situations where many variables might exist that possibly interact in a nonlinear manner, which is often the case with photo-z estimation. A visualization of an example tree generated by TPZ is shown in the left panel of Figure 1. In this figure, the plotting colors represent the magnitudes (or source colors) in which the data are recursively divided. In practice, however, the prediction trees are generally both denser and deeper than the sample tree shown in the Figure.

To compute photo-z PDFs in this study, we have used regression trees, which are a specific type of prediction trees. Regression trees are built by first starting with a single node that encompasses the entire data, and subsequently splitting the data within a node recursively into two branches along the dimension that provides the most information about the desired output. The procedure used to select the optimal split dimension is based on the minimization of the sum of the squared errors, which for a specific node is given by

$$S(\text{node}) = \sum_{m \in values(M)} \sum_{i \in m} (z_i - \hat{z}_m)^2 \qquad (1)$$

where m are the possible values (bins) of the dimension M, z_i are the values of the target variable on each branch, and $\hat{z}_m$ is the specific prediction model used. In the case of the *arithmetic mean*, for example, we would have that $\hat{z}_m = \frac{1}{n_m} \sum_{i \in m} z_i$, where n_m are the members on branch m. This allows us to rewrite Equation 1 as

$$S(\text{node}) = \sum_{m \in values(M)} n_m V_m \qquad (2)$$

where V_m is the variance of the estimator $\hat{z}_m$.

At each node in our tree, we scan all dimensions to identify the split point that minimizes the function $S(\text{node})$. We choose the dimension that minimizes $S(\text{node})$ as the splitting direction, and this process is recursively repeated until either a predefined threshold in $S(\text{node})$ is reached or any new child nodes would contain less than the predefined minimum leaf size. When constructed, each terminal leaf within the prediction tree *contains* spectroscopic data with different redshift values; the final prediction value for a given leaf node is determined from a regression model that covers these spectroscopic data. The simplest model is to simply return the mean value of the set of spectroscopic training redshifts contained within the leaf node, which provides a single estimate of a continuous variable. Alternatively, all of the spectroscopic training redshifts can be retained and subsequently combined with data from the matching leaf nodes in other prediction trees to form an aggregate, final prediction.

We create bootstrap samples from the input training data by sampling repeatedly from the magnitude using the magnitude errors. We use these bootstrap samples to construct multiple, uncorrelated prediction trees whose individual predictions are aggregated to construct a photo-z PDF for each individual galaxy by using a technique called a random forest. We also use a cross validation technique called Out-of-Bag (Breiman et al. 1984, CB13) within TPZ to provide extra information about the galaxy sample. This information includes an unbiased estimation of the errors and a ranking of the relative importance of the individual input attributes used for the prediction. This extra information can prove extremely valuable when calibrating the algorithm, when deciding what attributes to incorporate in the construction of the forest, and when combining this approach with other techniques.

TPZ has been tested extensively on different datasets, including the SDSS, DEEP2, and DES. In all tests, TPZ has performed comparable to if not better than other machine learning approaches. When high quality training data are available, TPZ has been shown to actually outperform other comparable techniques, both training and template based. Carrasco Kind & Brunner (2013a) provides a more detailed discussion of the TPZ algorithm and its application to different datasets.

2.2 SOM*z*

A Self Organized Map (SOM): (Kohonen 1990, 2001) is an unsupervised, artificial neural network algorithm that is capable of projecting high-dimensional input data onto a low-dimensional map through a process of competitive learning. In our case, the high dimensional input data can be galaxy magnitudes, colors, or some other photometric attributes, and two dimensions are generally sufficient for the output map. A SOM differs from other neural network based-algorithms in that a SOM is unsupervised (the redshift information is not used during training), there are no hidden layers and therefore no extra parameters, and it produces a direct mapping between the training set and the output network. In fact, a SOM can be viewed as a non-linear generalization of a principal component analysis (PCA).

The key characteristic of the self organization is that it retains the *topology* of the input training set, revealing correlations between inputs that are not obvious. The method is unsupervised since the user is not required to specify the desired output during the creation of the low-dimensional map, as the *mapping* of the components from the input vectors is a natural outcome of the competitive learning process. Another important characteristic of a SOM when applied to photo-z estimation is the creation of a structured ordering of the spectroscopic training data, since similar galaxies in the training sample are mapped to neighboring neural nodes in the trained feature map (CB14).

We demonstrate the construction of a self-organizing map in the right-hand panel of Figure 1. During this phase, each node on the two-dimensional map is represented by weight vectors of the same dimension as the number of attributes used to create the map itself. In an iterative process, each galaxy in the input sample is individually used to correct these weight vectors. This correction is determined so that the specific neuron (or node), which at a given moment best represents the input galaxy, is modified along with the weight vectors of that node's neighboring neurons. As a result, this *sector* within the map becomes a better representation of the current input galaxy.

This process is repeated for every galaxy in the train sample, and this entire process is repeated for several iterations. Eventually the SOM converges to its final form where the training data is separated into *groups* of similar features, which is illustrated in Figure 1 by the different cell colors within the output map. The result of this direct mapping procedure is an approximation of the galaxy training probability density function, and the map itself can be considered a simplified representation of the full attribute space of the input galaxy sample.

Building on our experience in creating TPZ, we have developed a similar approach, named SOM*z* (CB14), where prediction trees are replaced by SOMs to create what we called a *random atlas*. The random atlas is constructed from multiple maps that are each constructed from different bootstrap samples selected from the input training data by perturbing the input attributes using their measured error, where each one of these maps are built using a random subsample of the attribute space. The multiple, uncorrelated maps are aggregated to generate a photo-z PDF, in a similar manner as described earlier for the random forest.

As described previously, our SOM implementation not only updates the best-matching node but also the topologically closest nodes to it. This functionality ensures that the entire region surrounding the best-matching node is identified as being similar to the current input galaxy. As a result, similar nodes within the map are co-located, which naturally mimics how the input galaxies that have similar properties tend to be co-located in the higher dimensional input parameter space. We apply this procedure iteratively to all input galaxies, which are processed randomly during each iteration to avoid any biases that might arise if galaxies are processed in a specific order.

When running SOM*z*, there are few different parameters that must be determined, including the map resolution (i.e., the number of pixels in the map), the number of iterations required to build the map, and, most importantly, the underlying two-dimensional topology used for the maps. In this paper we follow the guidelines we presented in CB14 for these parameters, and use a spherical topology for the map, which are constructed by using HEALPIX (Górski et al. 2005), where each pixel in our maps has the same area. This topology was shown to be more accurate in many cases when compared to other topologies like a rectangular or hexagonal grid. In addition, a spherical topology has natural periodic boundary conditions which avoids possible edge effects.

In analogy with TPZ, we use cross validation, or OOB data, to estimate unbiased errors and to determine the relative importance of the different input attributes for this technique. These are both key pieces of information that will be used during the combination process, as we need to ensure that the same process is uniformly applied to each photo-z estimation technique. By doing this, we will enable a robust analysis of the final results from the combination of the different techniques. Carrasco Kind & Brunner (2014a) (CB14) provides a complete description of the SOM*z* implementation, the performance of this technique when applied to real data, and an exploration of specific parameter configurations.

2.3 Template fitting approach

Using spectral templates to estimate galaxy photo-zs from broadband photometry has a long history (Baum 1962); and this approach is, not surprisingly, one of the most utilized techniques. A primary advantage of this technique is the fact that a training sample is not required, thus this approach can be considered unsupervised. On the other hand, this technique has

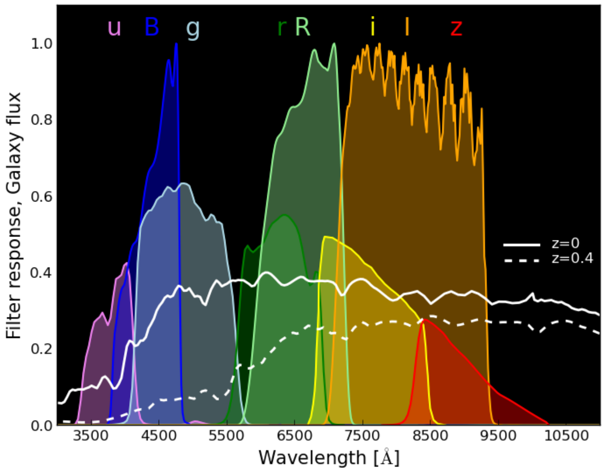

Figure 2. An Elliptical galaxy spectrum at z=0 and redshifted to z = 0.4 overlaid by the eight photometric filters from the DEEP2 galaxy survey (3 from the original survey and *ugriz* from a matched catalog (Matthews et al. 2013)).

the disadvantage that a complete and representative library of spectral energy distributions (SEDs) are required. Thus any incompleteness in our knowledge of the template SEDs that fully span the input galaxy photometry will lead to inaccuracies or misestimates in the computation of a galaxy photo-z.

A number of different groups have published template fitting photo-z estimation methods, all of which are roughly similar in nature. In this work, we have modified and parallelized one of the most popular, publicly available template fitting algorithms, `BPZ` (Benítez 2000). `BPZ` uses Bayesian inference to quantify the relative probability that each template matches the galaxy input photometry and determines a photo-z PDF by computing the posterior probability that a given galaxy is at a particular redshift. We can write this probability as $P(z \mid \mathbf{x})$ for a specific template t, where $\mathbf{x}$ represents a given set of magnitudes (or colors). If the identification of a specific template is not required, we can later marginalize over the entire set of templates $\mathbf{T}$.

By using Bayes theorem, we have:

$$P(z \mid \mathbf{x}) = \sum_{t \in \mathbf{T}} P(z, t \mid \mathbf{x}) \propto \sum_{t \in \mathbf{T}} \mathcal{L}(\mathbf{x} \mid z, t) P(z, t). \qquad (3)$$

$\mathcal{L}(\mathbf{x} \mid z, t)$ is the likelihood that, for a given redshift z and spectral template t, a specific galaxy has the set of magnitudes (or colors) $\mathbf{x}$. $P(z, t)$ is the prior probability of a specific galaxy is at redshift z and has spectral type t, this prior probability can be computed from a spectroscopic sample if one is available. The photo-z PDF is, therefore, either the posterior probability, if a prior is used, or the likelihood itself if no prior is used. This last point arises since the likelihood only depends on the collection of template SEDs; and, if this collection is representative of the overall galaxy sample, the likelihood can be used by itself as a photo-z PDF even without a spectroscopic training sample.

The use of a prior in a Bayesian analysis, however, is recommended. In this case, the prior probability can be computed directly from physical assumptions, from an empirical function calibrated by using a spectroscopic training sample (e.g., Benítez 2000), or from an empirical function calibrated by using machine learning techniques (see e.g., Carrasco Kind & Brunner 2013b, where we used Random Naïve Bayesian methods to compute the prior probabilities). For example, Benítez (2000) propose the following function for a single magnitude m_0:

$$\begin{aligned} P(z, t \mid m_0) &= P(t \mid m_0) P(z \mid t, m_0) \\ &\propto f_T e^{-k_t(m-m_0)} \times z^{\alpha_t} \exp\left(-\left[\frac{z}{z_{mt}(m)}\right]^{\alpha_t}\right). \end{aligned} \qquad (4)$$

where $z_{mt}(m) = z_0 t + k_{mt}(m - m_0)$. The five parameters of this function: f_T, m_0, α_t, z_{mt}, and k_{mt} can be constrained either by using direct fitting routines, or by using Markov Chain Monte Carlo methods to sample these parameters. These five parameters are dependent on the template t and can be quantified independently. For additional details on the underlying Bayesian approach, we refer the reader to the original paper by Benítez (2000).

As the goal of a template fitting method is to minimize the difference between observed and theoretical magnitudes (or colors), this approach is heavily dependent on both the library of galaxy SED templates that are used for the computation and the accuracy of the transmission functions for the filters used for particular survey. SED libraries are generally built from a base set of SED templates. These base templates broadly cover the Elliptical, Spiral, and Irregular categories, and a template library can be constructed by interpolating between the base spectral templates to create new spectra. One of the most widely used set of base templates are the four CWW spectra (Coleman et al. 1980), which include an Elliptical, an Sba, an Sbb, and an Irregular galaxy template. When extending an analysis to higher redshift, these temples are often augmented with two star bursting galaxy templates published by Kinney et al. (1996). One additional effect some template approaches consider is the presence of interstellar dust, which will introduce artificial reddening.

Once the library of galaxy SED templates has been constructed, the templates are convolved with the transmission functions for a particular survey to generate synthetic magnitudes as a function of redshift for each galaxy template. For the most accurate results, these transmission functions should include the effects of the Earth's atmosphere (if the observations are ground-based), as well as all telescope and instrument effects. This convolution process is demonstrated visually in Figure 2, which presents an example Elliptical galaxy spectral template at redshift zero and at a redshift 0.4. Overplotted on this figure is the filter set (B, R, and I) used by the DEEP2 survey, which is the data analyzed in this paper, along with the five extra filters: u, g, r, i, z presented in the DEEP2 photometry catalog compiled by Matthews et al. (2013).

3 PHOTO-Z PDF COMBINATION METHODS

We now turn our attention to the different methods with which we can combine distinct photo-z PDF estimation techniques (see e.g., Carrasco Kind & Brunner 2013b, where we first discussed combining Bayesian and machine learning predictions). In the statistics and machine learning communities, this topic is known as *ensemble learning* (Rokach 2010). Recently, Dahlen et al. (2013) have demonstrated that, on average, an improved photo-z estimate can be realized by combining the results from multiple template fitting methods. In this section, we build on this previous work to identify how Bayesian techniques can be used to construct a combined photo-z PDF estimator.

We can frame the problem mathematically by writing the set of photo-z PDFs for a given galaxy as a set of models $\mathbf{M}$, where each individual model M_k (e.g., TPZ, SOMz, or modified BPZ) provides a distinct photo-z PDF or posterior probability. A photo-z PDF can be written as $P(z \mid \mathbf{x}, \mathbf{D}, M_k)$, where $\mathbf{x}$ is the set of magnitudes or colors (note that without loss of generality we can use other attributes in this process) used to make the prediction and $\mathbf{D}$ corresponds to the training set which consists of N_d galaxies. We can also abbreviate this photo-z PDF as $P_k(z)$. These photo-z PDFs are each subject to the following constraint:

$$\int_{z_1}^{z_2} P_k(z)dz = 1 \tag{5}$$

for every model M_k, where z_1 and z_2 are the lower and upper limits, respectively, for the redshift range spanned by the galaxy sample. In the following subsections, we introduce different methods to aggregate these photo-z PDFs and show the results of these different methods in §5.

Given the variety of photo-z PDF estimation methods we are using (i.e., supervised, unsupervised, and model-based), we fully expect the relative performance of the individual techniques to vary across the parameter space spanned by the data. For example, supervised methods should perform the best in areas populated by high quality training data, while unsupervised or model-based methods should perform better where we have little or no training data. As a result, we can bin a specific subspace of our multi-dimensional parameter space and apply an individual combination method to each bin separately. This technique is demonstrated later in more detail with the Bayesian Model Averaging method (although it is more generally applicable).

3.1 Weighted Average

The simplest approach to combine different photo-z PDF techniques is to simply add the individual PDFs and renormalize the sum. In this case the final photo-z PDF is given by:

$$P(z \mid \mathbf{x}, \mathbf{M}) = \sum_k P(z \mid \mathbf{x}, M_k). \tag{6}$$

We can improve on this simple approach by including weights in the previous equation:

$$P(z \mid \mathbf{x}, \mathbf{M}) = \sum_k \omega_k P(z \mid \mathbf{x}, M_k). \tag{7}$$

These weights, ω_k, can be estimated for each input method by using the cross validation or OOB data, or from an intrinsic characteristic of the photo-z PDF, such as *zConf* that we introduced in CB13. In this work we use three weight schemes in addition to the uniform case:

PDF shape weights

In this case, ω_k is given by the the *zConf* parameter, which is similar to the *odds* parameter presented in Benítez (2000) *zConf* is defined as the integrated probability between $z_{\rm phot} \pm \sigma_k(1 + z_{\rm phot})$, where $z_{\rm phot}$ is a single estimated value for the photo-z PDF. This single photo-z estimate can be either the mean or the mode of the photo-z PDF. Likewise, we can estimate σ_k for each input method either by using the OOB data, by selecting a constant value across all input methods, or by selecting these values separately so that all photo-z PDFs have the same cumulative *zConf* distributions. *zConf* quantifies the sharpness of the PDF and can take values from zero to one. In CB13 and CB14, we demonstrated that there is a correlation between this value and the accuracy of the overall photo-z. Specifically, we observed that, on average, galaxies with higher *zConf* have more accurate photo-z PDFs than galaxies with lower *zConf* values.

Best fit weights

An alternative method to compute the values of ω_k is to use the cross-validation data to first determine the weight values that minimize the difference between $z_{\rm phot}$ and $z_{\rm spec}$; and, second to apply these best fit values to the test data. This method seeks the optimal linear combination of each individual PDF, thus it allows the values of ω_k to be negative. After the combination is completed, we renormalize according to Equation 5. This method can be applied to a binned sub-sample to take advantages of the performance of each method in different areas of the attribute space.

Oracle scheme

As mentioned, when the input, multi-dimensional data have been binned (c.f. Figure 9), we can use the cross-validation data to select only one model from among all available input models to only be used with the test data located within that specific bin. Since we are allowed to only select one input model, this will result in an assigned weight value of one for the chosen model and zero otherwise, however the chosen model is allowed to vary between bins.

The primary disadvantage of these simple, additive models is that incorrect estimates for the errors for the selected input model can bias the final result. On the one hand, if a technique has underestimated errors, the final result will be biased towards this one input method. On the other hand, overestimation of the errors will bias the final result away from this particular method. One approach to address this issue, as discussed by Dahlen et al. (2013), is to either smooth or sharpen the photo-z PDFs estimated by each method by using the OOB data until their error distributions are approximately Gaussian with unit variance. We can generalize this approach to transform a photo-z PDF as $P_k(z) = P_k(z)^{\alpha_k}$, where we adjust the value of α_k by using either the cross validation data when errors are over estimated or use a Gaussian smoothing filter when they are under estimated.

3.2 Bayesian Model Averaging

Bayesian Model Averaging (BMA) is an ensemble technique that combines different models within a Bayesian framework. BMA accounts for any uncertainty in the correctness of a given model by integrating over the model space and weighting each model by the estimated probability of being the *correct* model. As a result, BMA acts as a model selection procedure that handles the uncertainty in selecting the best model by using a combination of models instead. This is because BMA considers the uncertainty in selecting the best model while working under the assumption that only one model is actually the best (Monteith et al. 2011). BMA has been used for astrophysical problems (see e.g., Gregory & Loredo 1992; Trotta 2007; Debosscher et al. 2007) in, for example, the determination of cosmological parameters and variable star classification (see, Parkinson & Liddle 2013, for a review on using BMA in astronomy).

When using BMA, the training data are used to characterize each of the models that will be combined. For each galaxy, the final PDF, $P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M})$, is given by:

$$P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M}) = \sum_k P(z \mid \mathbf{x}, M_k) P(M_k \mid \mathbf{D}). \tag{8}$$

$P(M_k \mid \mathbf{D})$ is the probability of the model M_k given the training data $\mathbf{D}$, which can be viewed as a simple, model dependent weighting scheme. This probability can be computed by using Bayes' Theorem:

$$\begin{aligned} P(M_k \mid \mathbf{D}) &= \frac{P(M_k)}{P(\mathbf{D})} P(\mathbf{D} \mid M_k) \\ &\propto P(M_k) \prod_{i=1}^{N_d} P(d_i \mid M_k). \end{aligned} \tag{9}$$

We have omitted the $P(\mathbf{D})$ term as it is merely a normalization factor and we use the same data for all models. d_i is the i^{th} element from the training data $\mathbf{D}$, which are assumed to be independent.

For each model, we assign the value ϵ_k as an average error for the estimation process. ϵ_k can be computed as the fraction $N_k^{(b)}/N_d$, where $N_k^{(b)}$ is the number of galaxies considered to be misestimated or *bad* for the particular photo-z PDF method k. To quantify when a specific galaxy is a bad prediction we compute

$$N_{k,i}^{(b)} = \begin{cases} 1 & \text{if } \int_{z_s-\delta_z}^{z_s+\delta_z} P(z \mid \mathbf{x}, d_i) dz \leqslant \pi_z, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

In this equation, z_s is the spectroscopic redshift for the i^{th} training set galaxy. The first parameter, δ_z, controls the width of a window centered on z_s within which we accumulate photo-z probability for the i^{th} training galaxy around the true redshift. The second parameter, π_z, is the minimum probability within this window for which we consider the model prediction to be good. We find that $\pi_z = 0.5$ and $\delta_z = 0.05$ provides a good discriminant between good and bad photo-z model estimates.

Given the individual good/bad predictions for each training set galaxy, we can compute the total number of bad predictions, $N_k^{(b)}$, by summing over the individual predictions, $N_{k,i}^{(b)}$, for the entire training data, $\mathbf{D}$. The total number of good prediction will naturally be $N_d - N_k^{(b)}$. As a result, we can rewrite Equation 9:

$$P(M_k \mid \mathbf{D}) \propto P(M_k) (1-\epsilon_k)^{N_d - N_k^{(b)}} (\epsilon_k)^{N_k^{(b)}}, \tag{11}$$

where $P(M_k)$ is the probability of each model k, which we can assume to be unity for all models. Therefore, the final PDF for each galaxy is given by

$$\begin{aligned} P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M}) \propto \sum_k P(z \mid \mathbf{x}, M_k) P(M_k) \times \\ (1-\epsilon_k)^{N_d - N_k^{(b)}} (\epsilon_k)^{N_k^{(b)}}. \end{aligned} \tag{12}$$

We applied the BMA technique to individual bins within the multi-dimensional parameter space occupied by a given data set. We demonstrate this binned BMA technique in Figure 9, where we use a Self Organized Map to project our entire input parameter space to a two-dimensional map. In this manner, all magnitudes or colors are used to form the binned regions within which the parameters of the ensemble learning approach can vary. After computing photo-z PDFs for all galaxies with each method, we use BMA to determine the relative weights for these input techniques within each bin; we can visualize these weights as different colors across the two-dimensional map, as shown in Figure 9. This figure graphically displays how the *accuracy* of each photo-z PDF estimation varies across the parameter space, and thus how the different weights themselves vary.

3.3 Bayesian Model Combination

As discussed, Bayesian Model Averaging tries to select the best model among the ones introduced to the algorithm. Alternatively, we can modify BMA to produce an more optimal model combination technique (Monteith et al. 2011) known as Bayesian Model Combination (BMC). With BMC, instead of directly combining the three different photo-z PDF estimates as was the case with BMA, the Bayesian process is used to explore different combinations of the individual photo-z PDF techniques. Thus, an ensemble of different photo-z PDF combinations are generated and we directly compare different model combinations.

As a simple example, we could first generate hundreds different random weights for all three of our photo-z PDF estimation techniques, and second use these to compute hundreds of new *sets* of PDFs by computing a simple weighted average by using Equation 7. Finally, we could apply BMA to this PDF ensemble to determine the final PDF. In this case, we could write Equation 8:

$$P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M}, \mathbf{E}) = \sum_{e \in \mathbf{E}} P(z \mid \mathbf{x}, \mathbf{M}, e) P(e \mid \mathbf{D}), \tag{13}$$

where e is an element from the set $\mathbf{E}$ of these hundreds combined models. Here we need to compute the performance of each combination e and apply the BMA formulation, shown in Equations 9 and 10, to those models by using the model e instead of M_k, i.e.,

$$P(e \mid \mathbf{D}) \propto P(e) \prod_{i=1}^{N_d} P(d_i \mid e). \tag{14}$$

Fundamentally, with BMC we are marginalizing over the uncertainty in the correct model combination, where in BMA we marginalized over the uncertainty in identifying the correct model from the entire ensemble.

The number of model combinations $\mathbf{E}$ is, in principle, infinite, and in practice can be very large. To overcome this, we can use sampling techniques over a reasonable, finite number of models. Naively we might use randomly generated weights, however, this approach can be costly to fully span the allowed range of weights and convergence towards a satisfactory solution might be slow. Thus, instead of assigning weights randomly or using incremental steps within a regular grid, we sample the weights from a Dirichlet distribution where the *concentration* parameters are modified until they converge to stable values. We require that the set of weights, w_k, for each of the three models, M_k, satisfy $\sum w_k = 1$ and also $w_k > 0$.

For a concentration parameter $\boldsymbol{\alpha}$ of the same dimension as $\mathbf{w}$, we have that the probability distribution for $\mathbf{w}$ is given by:

$$P(\mathbf{w}) \sim \mathcal{D}\text{ir}(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k w_k^{\alpha_k - 1}, \tag{15}$$

where $\mathcal{D}\text{ir}(\boldsymbol{\alpha})$ is the *Dirichlet* distribution, $\Gamma(\alpha_k)$ is the *gamma function* and k are the base models, which in this paper are TPZ, SOMz, and our modified BPZ. In order to generate a set $\mathbf{E}$ of combined models, we first set α_k to unity for all values of k. Second, we sample from this distribution n_s times (n_s is a fixed number, generally between 2 and 5, which we fixed at 3) to get a set of n_s weights and n_s new model combinations. Next, we compute $P(e \mid D)$ by using Equations 9 and 10 for each model in the set of n_s models. We, temporarily, select the best model among the set n_s, i.e, the one with highest $P(e \mid \mathbf{D})$, and update the α_k parameters by simply adding the weights from the corresponding model to the current values of $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{t} + \max_{\mathbf{w}_e \in n_s} P(e \mid \mathbf{D}) \tag{16}$$

where t is just a symbolic reference to the fact that $\boldsymbol{\alpha}$ is being updated every 3 steps.

We use the latest values for $\boldsymbol{\alpha}$ to continue the sampling process to obtain the next set n_s of model combinations. As a result, we continually (by adding n_s new models at each step) extend our set of model combinations $\mathbf{E}$. As the chain of models in this set is constructed iteratively, the process can be terminated either when a predefined number of model combinations has been reached or when new model combinations have started to converge. This process behaves similarly to a Markov Chain Monte Carlo process, and we have an analogous phase to the *burn in* step, where we can omit some number of model combinations at the start of our set $\mathbf{E}$ of model combinations. Thus, our final photo-z PDF prediction is the application of BMA over the remaining elements in $\mathbf{E}$, we have set for this work the size of E to be 800. Finally, we note that, as was the case with BMA, we can develop a binned version of our BMC technique, where we develop different model combinations for different region of the magnitude (color) space by using a SOM.

3.4 Hierarchical Bayes

A Hierarchical Bayesian (HB) method provides a different approach to combine the individual photo-z PDFs. In a manner similar to BMA, we include the uncertainty that a given photo-z PDF for a specific galaxy might be incorrectly predicted as a set of nuisance parameters over which we later marginalize.

Adopting our previous notation, we follow a similar approach to Fadely et al. (2012) and Dahlen et al. (2013), and we write the photo-z PDF for an individual galaxy for each base method k:

$$P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k) = \sum_j P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_{kj}) \times P(\theta_{kj} \mid \mathbf{D}, M_k), \tag{17}$$

where we have introduced the *hyperparameter* θ_k, a nuisance parameter that characterizes our uncertainty in the prior distribution of model k. The parameter θ_k can be quantified in different forms, but essentially is the misclassification probability of the k^{th} method. Thus, we quantify this mis-prediction probability with $P(\theta_k)$; and we drop the dependence on $\mathbf{x}$, the measured galaxy attributes, as it does not directly affect the parameter θ_k. Since we will marginalize over θ, we keep the term $\mathbf{D}$ as we can use the training data to place limits on θ_k by using the cross-validation data. We note that these probabilities are subject to:

$$\sum_j P(\theta_{kj} \mid \mathbf{D}, M_k) = 1. \tag{18}$$

If we consider the case where galaxies are predicted correctly or are outliers, j is a binary state. In this model, if we assume that γ_k is the fraction of galaxies that are mis-predictions or are labeled as outliers for method k, we have: $P(\theta_{k0} \mid \mathbf{D}, M_k) = \gamma_k$ and $P(\theta_{k1} \mid \mathbf{D}, M_k) = (1 - \gamma_k)$. In this case, Equation 17 becomes:

$$P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k) = P_{def}(z \mid M_k, \theta_k)\gamma_k + P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k)(1 - \gamma_k), \tag{19}$$

where $P_{def}(z \mid M_k, \theta_k)$ is the default PDF that should be used for the k^{th} method when the original PDF for that method has been determined to be mis-predicted or wrong. In the second term, we use the original PDF for the method k, which is multiplied by the fraction of well predicted objects $1 - \gamma_k$.

The final PDF after we combine the different photo-z PDFs from our base methods in the HB approach is given by:

$$P(z \mid \mathbf{x}, \mathbf{D}, \theta) = \prod_k P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k)^{1/\beta}. \tag{20}$$

Here, following Dahlen et al. (2013), we have introduced an extra parameter β, which is a constant value that quantifies the degree of covariance between the different base methods. $\beta = 1$ corresponds to complete independence between the base methods, while $\beta = 3$ (or, more generally, the total number of methods) would correspond to full covariance between them. We can compute β from the OOB sample in such way the final error distribution follows a normal distribution with zero mean and unit variance, as we have done in this paper. Alternatively, we can marginalize over all possibles values of β when no cross validation data is available and we can integrate over the uncertainty of this parameter.

Finally, by marginalizing over θ we have our final PDF: $P(z \mid \mathbf{x}, \mathbf{D})$, or simply $P(z)$ given by:

$$P(z) = \int_0^1 P(z \mid \mathbf{x}, \mathbf{D}, \theta) P(\theta) d\theta, \tag{21}$$

where $P(\theta)$ is a constant which in the simple case is equal to unity. If OOB data is available, we can narrow down the range of allowed values for θ (or effectively γ_k), so we can set up a limited range for γ_k based on the performance of each method k on this data. In this case, $P(\theta)$ will act as a top-hat window function. In any case, the final $P(z)$ is subject to Equation 5. As discussed before, we can either apply the HB approach to the entire data set, or we can partition the input space and apply the HB approach independently to the binned regions of the parameter space.

4 DATA

To explore different configurations and to demonstrate the capabilities and the efficacy of these photo-z combination techniques, we follow the approach we presented in CB13 and CB14, but in this paper we restrict our analysis to data obtained by the Deep Extragalactic Evolutionary Probe (DEEP) survey and the Sloan Digital Sky Survey (SDSS). In the rest of this section we provide a summary of these data and detail how we extracted the data sets from these surveys that we use in the analysis presented in §5.

Table 1. The photo-z PDF combination methods, their weights and abbreviations presented in this paper.

Method	Weights^a	Abbreviation
Weighted Average	Uniform	$\mathrm{WA}_{\mathrm{flat}}$
Weighted Average	$zConf$	$\mathrm{WA}_{\mathrm{shape}}$
Weighted Average	best fit	$\mathrm{WA}_{\mathrm{fit}}$
Weighted Average	oracle predictor	$\mathrm{WA}_{\mathrm{oracle}}$
Bayesian Model Averaging		BMA
Bayesian Model Combination		BMC
Hierarchical Bayes		HB

^{a} if applicable

4.1 Deep Extragalactic Evolutionary Probe

The DEEP survey is a multi-phase, deep spectroscopic survey that was performed with the Keck telescope. Phase I used the Low Resolution Imaging Spectrometer (LIRS) instrument (Oke et al. 1995), while phase II used the DEep Imaging Multi-Object Spectrograph (DEIMOS) (Faber et al. 2003). The DEEP2 Galaxy Redshift Survey is a magnitude limited spectroscopic survey of objects with $R_{AB} < 24.1$ (Davis et al. 2003; Newman et al. 2013a). The survey includes photometry in three bands from the Canada-France-Hawaii Telescope (CFHT) 12K: B, R, and I and it was recently extended by cross-matching the data to other photometric data sets. In this work, we use Data Release 4 (Matthews et al. 2013), the latest DEEP2 release that includes secure and accurate spectroscopy for over 38,000 sources. The original input photometry for the sources in this catalog was supplemented by using two u, g, r, i, and z surveys: the Canada-France-Hawaii Legacy Survey (CFHTLS; Gwyn 2012), and the SDSS. For additional details about the photometric extension of the DEEP2 catalog, see Matthews et al. (2013).

To use the DEEP2 data with our implementation, we have selected sources with secure redshifts (ZQUALITY$\geqslant$ 3), which were securely classified as galaxies, have no bad flags, and have full photometry. Even though the filter responses are similar, the u, g, r, i, and z photometry originates from two different surveys and are thus not identical. We therefore only present the results from those galaxies that lie within field 1 that have CFHTLS photometry. Furthermore, we have corrected these observed magnitudes by using the extinction maps from Schlegel et al. (1998). In the end, this leaves us with a total of 10,210 galaxies each with eight band photometry and redshifts. From this data set, we randomly select 5,000 galaxies for training and hold the remainder out for testing. The computation of photo-z PDFs was completed by using the magnitudes in the bands B, R, I, u, g, r, i, and z and their corresponding colors $B-R$, $R-I$, $u-g$, $g-r$, $r-i$, and $i-z$, providing a total of fourteen dimensions.

4.2 Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS; York et al. 2000) phases I, II and III conducted a photometric survey in the optical bands: u, g, r, i, z that covered more than 14,000 square degrees, more then one-quarter of the entire sky. The resultant photometric catalog contains photometry for over 10^8 galaxies, making the SDSS one of the largest sky surveys ever completed. The SDSS also conducted a spectroscopic survey of targets selected from the SDSS photometric catalog. In this paper, we use a subset of the spectroscopic data contained within the Data Release 10 catalog (Ahn et al. 2013, SDSS-DR10), which includes over two million spectra of galaxies and quasars which include those taken as apart as the Baryonic Oscillation Spectroscopic Survey (BOSS) program (Dawson et al. 2013).

Specifically, we selected galaxies by using the online CasJobs website[4] and the following query from the DR10 data base:

```
SELECT spec.specObjID,
    gal.dered_u, gal.dered_g, gal.dered_r,
    gal.dered_i, gal.dered_z,
    gal.err_u, gal.err_g, gal.err_r,
    gal.err_i, gal.err_z,
    spec.z AS zs
INTO mydb.DR10_spec_clean_phot
FROM SpecObj AS spec
JOIN Galaxy AS gal
ON spec.specobjid = gal.specobjid,
    PhotoObj AS phot
WHERE spec.class = ‘GALAXY’ -- Spectroscopic class
                            -- (GALAXY, QSO, or STAR)
AND gal.objId = phot.ObjID
AND phot.CLEAN=1            -- Clean photometry flag
                            -- (1=clean, 0=unclean)
AND spec.zWarning = 0       -- Bitmask of warning
                            -- vaules; 0 means all
                            -- is well
```

We also removed some additional bad photometric observations, ensured the redshift values were positive, and compute colors for the final catalog, which contains 1,147,397 galaxies. The spectroscopic data range from $z \approx 0.02$ up to $z \approx 0.8$; the full spectroscopic redshift distribution of these galaxies is shown in the gray shaded histogram presented in Figure 15. These data are dominated by the Main Galaxy Sample (MGS) at low redshifts, with mean redshift of $z \sim 0.1$, and by luminous red galaxies (LRG) at higher redshifts, with mean redshift of $z \sim 0.5$.

From this sample, we randomly selected 50,000 galaxies for training and hold the remaining 1,097,397 for testing. This training set corresponds to approximately 4.5% of the test set. We note that this is a blind test, as the testing data are not used in any way to train or calibrate the algorithms. Of all the measured attributes in the SDSS photometric catalog, we have only used the nine dimensions corresponding to the five galaxy, extinction corrected, model magnitudes and the four colors derived from these five magnitudes: u, g, r, i, z, $u-g$, $g-r$, $r-i$, and $i-z$.

5 RESULTS/DISCUSSION

We now turn to the actual application of the ensemble learning approaches described in §3 to the data introduced in §4. We present the seven combination methodologies we use in this section in Table 1, which also includes an abbreviated name that we will use to refer to a specific technique. We follow a similar approach to CB14 in order to compare different combination methods, and define the bias to be $\Delta z' = |z_{\mathrm{phot}} - z_{\mathrm{spec}}|/(1 + z_{\mathrm{spec}})$. We also present the standard

[4] http://skyserver.sdss3.org/CasJobs/

metrics we use to compare the performance of the different combination techniques in Table 2. As shown in this table, we define five metrics to address the bias and the variance of the results (the first five rows) and we present three values to characterize the outlier fraction.

We also use the KS metric, which represents the results of a Kolmogorov–Smirnov test that quantifies the likelihood that the predicted photo-z distribution and the spectroscopic redshift distribution $N(z)$ are drawn from the same underlying population. This metric provides a single, robust value to compare both distributions that does not depend on how the results are binned by redshift, and it is defined as the maximum distance between both empirical distributions.

To determine this statistic, we compute the empirical cumulative distribution function (ECDF) for both distributions. For the spectroscopic sample, the ECDF is defined as:

$$F_{\rm spec}(z) = \sum_{i=1}^{N} \Omega_{z^i_{\rm spec}<z} \qquad (22)$$

where N is the number of galaxies in the redshift sample, and

$$\Omega_{z^i_{\rm spec}<z} = \begin{cases} 1, & \text{if } z_{{\rm spec},i} < z \\ 0, & \text{otherwise} \end{cases} \qquad (23)$$

The ECDF for the photo-z distribution is simply the accumulation of the probability presented in the photo-z PDF. The summation is carried out over all galaxies in the sample. Given the ECDF for both the photo-z and spectroscopic distributions, we compute the KS statistic as:

$$\mathrm{KS} = \max_z \left(||F_{\rm phot}(z) - F_{\rm spec}(z)|| \right) \qquad (24)$$

Thus, as the KS statistic decreases, the two distributions become more similar.

All of the metrics listed in Table 2 are positive and characterized by the fact that lower metric values indicate a more accurate photo-z PDF. In CB14 we defined a new, meta-statistic called I-score (symbolically represented by $I_{\Delta z'}$) that provides a single statistic to simplify the comparison of different photo-z techniques. To compute this metric, we first normalize each set of metrics across all different photo-z estimation techniques so that we are not biased by different dynamic ranges. Thus, for example, we first compute the mean and standard deviation for $<\Delta z'>$ for each combination technique, and subsequently rescale all individual $<\Delta z'>$ values so that this set of values has zero mean and unit variance.

We continue this process for all nine statistics listed in Table 2, and compute their weighted sum to obtain the total I-score:

$$I_{\Delta z'} = \sum w_i M_i, \qquad (25)$$

where M_i is the rescaled metric and weight value for metric i out of the nine available. For simplicity, we use equal weights in the remainder of this paper (and thus the I-score is simply the average of the nine rescaled metrics for each technique). As a result, the photo-z method (or parameter configuration) with the lowest I-score will be the optimal estimation technique. On the other hand, if we were looking for the technique or the specific parameter configuration with, for instance, the lower outlier fraction, we could assign higher weights accordingly to select the best technique. In this way, we can efficiently select the best method or configuration for specific research requirement.

Table 2. The definition of the metrics used to compare different photo-z combination methods.

Metric	Meaning
$<\Delta z'>$	mean of $\Delta z'$
$\vert\Delta z'\vert_{50}$	median of $\Delta z'$
$\sigma_{\Delta z'}$	Standard deviation of $\Delta z'$
σ_{68}	Sigma value at which 68% of $\Delta z'$ is enclosed
$\sigma_{\rm MAD}$	Median absolute deviation = median($\vert\vert\Delta z' - \vert\Delta z'\vert_{50}\vert\vert$)
KS	Kolmogorov - Smirnov statistic for $N(z)$
$\mathrm{out}_{0.1}$	Fraction of outliers where $\Delta z' > 0.1$
$\mathrm{out}_{2\sigma}$	Fraction of outliers where $\vert\Delta z' - <\Delta z'>\vert > 2\sigma_{\Delta z'}$
$\mathrm{out}_{3\sigma}$	Fraction of outliers where $\vert\Delta z' - <\Delta z'>\vert > 3\sigma_{\Delta z'}$
$I_{\Delta z'}$	I-score, a weighted combination of all other metrics.

5.1 Cross validation data

In CB13, we introduced OOB data and demonstrated its use as a cross-validation data set that provided error quantification and overall performance similar to what could be expected when applying an algorithm directly to the test data set. When building a tree with TPZ or a map with SOMz, a fraction of the overall training data, usually one-third, is extracted and not used during the tree/map construction process. The resultant tree/map is subsequently applied to this unused data to make a photo-z prediction, and this process is repeated for every tree/map. These photo-z predications are aggregated for each galaxy to make a photo-z PDF; and by construction a galaxy can never be used to train any tree/map that is subsequently used to predict that galaxy's photo-z. Thus, as long as the OOB data remains similar to the final testing data, the OOB data provide results that will be similar to the final test data results and can be used to guide expectations when applied blindly to other data.

As an illustration of this process, Figure 3 compares the photometric (as computed by using SOMz) and spectroscopic redshifts for galaxies in the training (5,000 in total) and testing (5,210) samples as selected from field 1 of the DEEP2 data set. As shown in this Figure, the performance on both the OOB and the testing data are visually similar and there is no indication of overfitting. In addition, general features in the result, like the spread of the data or the slight tilt of the distribution of points relative to the diagonal, are observed in both samples.

A similar conclusion is observed with the SDSS data, as shown in Figure 4 where the photometric (as computed by using TPZ) and spectroscopic redshifts for 50,000 galaxies from the training set are compared to 50,000 randomly selected galaxies from the test set. Both distributions show similar behavior and global trends, thus we conclude that, as expected, the OOB data can be used to predict the performance of an PDF combination algorithm on real data.

Another method to contrast the results from these data is to compute the correlation between each of the three photo-z estimation techniques discussed earlier as a function of redshift. For this, we use the photo-z PDFs for all galaxies, and we calculate the Pearson correlation coefficient R_{ik} within each redshift bin. Even if the three input methods are completely independent, we should expect a positive correlation between them if their predictions are similar. In fact, we desire a positive correlation (but not necessarily a perfect correlation) between the techniques as this will indicate the different techniques are all performing well.

We present the Pearson correlation coefficient for the

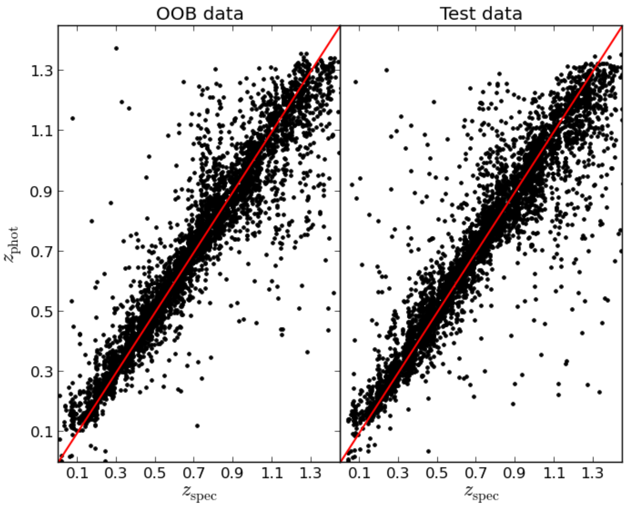

Figure 3. A comparison of the photometric (computed by using `SOMz`) and spectroscopic redshifts for training set (left) and test set (right) galaxies from field 1 of the DEEP2 survey.

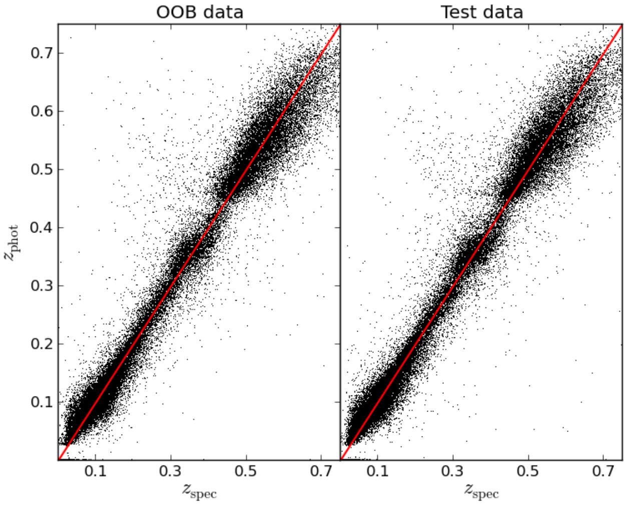

Figure 4. A comparison of the photometric (computed by using `TPZ`) and the spectroscopic redshift from the SDSS-DR10 for the 50,000 training set galaxies (left) and 50,000 galaxies randomly subsampled from the 1,097,397 galaxies in the test set (right).

three photo-z PDF estimation techniques for the DEEP2 data (top panel) and the SDSS data (bottom panel) in Figure 5. In this figure we display these correlation coefficient computed from the cross-validation (OOB) data (dashed line) and the test data (solid line). The global agreement between these lines further demonstrates the importance of the OOB data as a predictor of the performance of a given technique. This figure also demonstrates a tighter correlation between the two machine learning algorithms than between any machine learning algorithm and the template technique, which is not surprising given the similarities in the methods. While not shown, the shape of the covariance matrices resemble the spectroscopic $N(z)$ distributions presented in Figures 11 and 15. We conclude that this is expected since a larger number of galaxies can naturally produce a greater chance for divergent photo-z estimates.

As mentioned previously, a concern when combining photo-z PDFs from different methods is to reduce the likelihood of being biased by methods that might under- or overestimate their errors. To further demonstrate the importance of the cross-validation data, we compare the normalized error distribution between the cross-validation (OOB) and test data in Figure 6 for both DEEP2 (top panel) and SDSS (bottom panel) data, where the photo-z PDFs were generated by `TPZ` . In both cases, the two curves are nearly identical, and we confirmed the same result with both `SOMz` and `BPZ`. Thus we can use the OOB data error estimate to rescale the PDF for the test data by using the results computed from the OOB data.

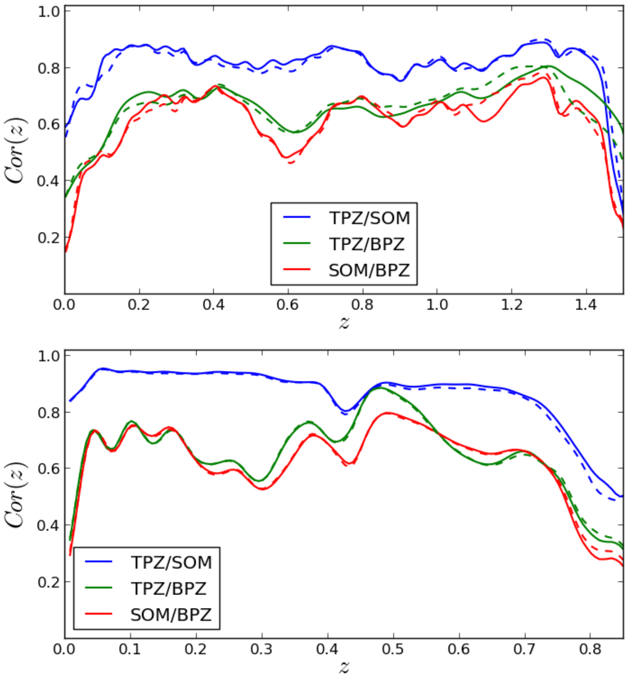

Figure 5. The Pearson correlation coefficient between the individual photo-z PDF estimation methods as a function of redshift for the DEEP2 (top) and SDSS (bottom) data. The coefficients measured from the cross-validation (OOB) data (dashed line) and from the test data (solid line) are nearly identical, indicating the utility of the OOB data in predicting the performance of an algorithm on blind test data. Note that a positive correlation is beneficial since this measures the relative performance of different techniques in predicting redshifts.

5.2 Photo-z PDF Combination for DEEP2

To combine the three photo-z PDF techniques discussed in §2, we employ a binning strategy to allow different method combinations to be used in different parts of parameter space. We first create a two dimensional, 10 × 10 SOM representation of the full 14-dimensional space (eight magnitudes and six colors, note that we do not compute a color between the two different photometric input surveys) by using a rectangular topology to facilitate visualization. With this map we can perform an analysis of all galaxies that lie within the same cell, in a similar process to that described in CB14, but now instead of predicting a photo-z, we are computing the optimal model combination. We apply all seven combination methods presented in Table 1 to all galaxies within each cell by using the OOB data that are also contained within the same cell. We

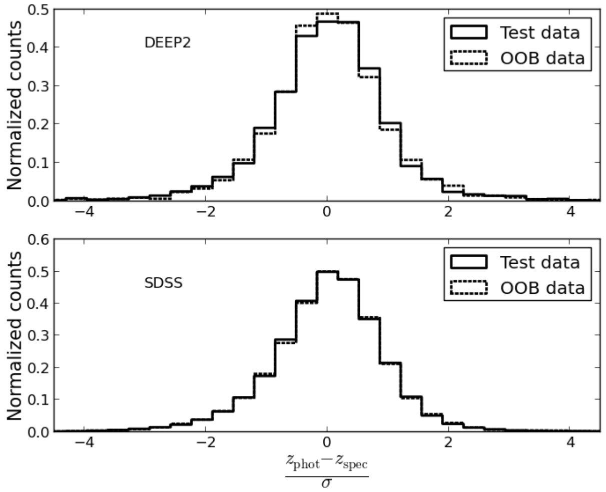

Figure 6. The normalized error distributions for galaxies in DEEP2 (top) and SDSS (bottom). The error distribution computed from the test data is shown in red, while the error distribution for the cross-validation (OOB data) is shown in black. The excellent agreement highlights the importance of the OOB data in predicting the results of blind test data predictions.

note that the $\mathrm{WA_{flat}}$ and $\mathrm{WA_{shape}}$ methods do not depend on this binning, and can, therefore, be used without OOB data. We also could employ the HB approach without using this map, but in this case we would need to define $P_{def}(z \mid M_k, \theta_k)$ and perform the marginalization over the entire range of θ_k without any prior on this value.

We present a summary of the results obtained by applying the seven different combination techniques to all the galaxies within the DEEP2 data in Table 3. The bold entries in this Table highlight the best technique for any particular metric. The first three rows in this Table show the individual photo-z PDF estimation techniques, of which TPZ generally performs the best and is thus shown in the first row as the benchmark. This Table also clearly indicates that the seven different combination techniques generally have a similar performance, and, as shown in the last four rows, often perform better than TPZ.

We observe that the last four methods: $\mathrm{WA_{fit}}$, BMA, BMC, and HB all use the binned model combination approach, and thus can take advantage of the different performance characteristics of individual codes. In this case, BMC provides the best performance as measured by the I-score $I_{\Delta z'}$, the bias $< \Delta z' >$, the scatter $\sigma_{\Delta z'}$, and the outlier fraction $\mathrm{out}_{0.1}$. Overall, the differences are close to 5% for many of the metrics, which, while small, are still significant since these are averaged metrics over the full test galaxy sample.

In Figure 7, we present a visual comparison between the ten different photo-z estimation techniques for five different metrics: bias, scatter, outlier fraction, KS test, and the I-score. In each panel, the horizontal dashed line shows the best value from the individual photo-z PDF estimation methods and the shaded area separates the individual from the combined methods. This Figure demonstrates that the Bayesian modeling techniques provide better performance than the best individual method over all five metrics, and also that by employing the binning scheme to optimize the combination approach we achieve better performance than for the best individual technique.

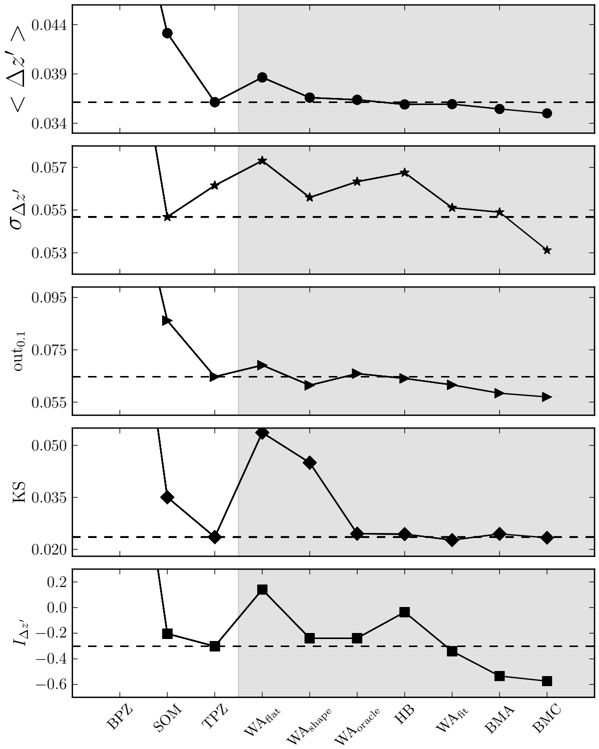

Figure 7. A comparison of the average performance for the three individual photo-z PDF estimation methods and the seven different photo-z PDF combination approaches for five different metrics as defined in Table 2 for the DEEP2 data. The horizontal dashed line indicates the best result for a given statistic among the three individual methods (note, BPZ is not always shown at the provided scale), and the shaded area separates the individual methods from the combined approaches. All values are presented in Table 3.

We compare the actual photo-z PDF for a single galaxy selected from the DEEP2 survey as estimated by the three individual techniques with the photo-z PDF estimated by the BMC method in Figure 8. This Figure clearly shows how the re-normalized combined PDF from the three individual photo-z PDF estimation techniques has been improved as the BMC result is closer to the true galaxy redshift, shown by the vertical line. These combination techniques identify which individual method works best in different cells, and can use that information to either weight the individual photo-z PDFs accordingly, or in the case of BMC to marginalize over the uncertainty in the correct weights to produce the best combination.

We apply a SOM to the DEEP2 field 1 data in order to construct a two-dimensional, binned combination of the three individual photo-z PDF estimation methods. We use this SOM to determine the weights for the three individual methods for each cell, and present the results in Figure 9 when using the BMA approach as it is easy to interpret. We also show the mean DEEP2 R-band magnitude for all galaxies in a given cell in the lower right panel, which clearly indicates the ability of the SOM to preserve relationships between galaxies when projecting from the higher dimensional space to the two-dimensional map. Of course, the SOM mapping is a non-linear representation of all magnitudes and colors, thus the DEEP2 R-band map should only be used to provide guidance.

In the three weight maps, a redder color indicates a higher weight, or equivalently that the corresponding method performs better in that region. These weight maps demonstrate the variation in the performance of the individual techniques

Table 3. A summary of the performance results for the three individual methods and the seven different photo-z PDF combination methods as applied to the DEEP2 data, no magnitude cut was applied during the training phase. The bold entries highlight the best value within each column to aid in the interpretation of the table (c.f. Figure 7).

Combination method	$<\Delta z'>$	$\lvert\Delta z'\rvert_{50}$	$\sigma_{\Delta z'}$	σ_{68}	$\sigma_{\rm MAD}$	KS	out$_{0.1}$	out$_{2\sigma}$	out$_{3\sigma}$	$I_{\Delta z'}$
TPZ	0.0361	0.0205	0.0561	0.0257	0.0139	0.0235	0.0647	0.0307	0.0184	-0.3021
SOM	0.0431	0.0291	0.0547	0.0325	0.0188	0.0350	0.0862	**0.0284**	**0.0150**	-0.2035
BPZ	0.0635	0.0476	0.0679	0.0428	0.0273	0.1342	0.1636	0.0338	0.0170	2.3255
WA$_{\rm flat}$	0.0386	0.0231	0.0573	0.0285	0.0155	0.0537	0.0691	0.0313	0.0192	0.1409
WA$_{\rm oracle}$	0.0364	0.0206	0.0563	0.0260	0.0139	0.0245	0.0659	0.0313	0.0184	-0.2385
WA$_{\rm shape}$	0.0366	0.0217	0.0556	0.0268	0.0146	0.0450	0.0614	0.0297	0.0186	-0.2392
WA$_{\rm fit}$	0.0359	0.0208	0.0551	**0.0253**	**0.0137**	**0.0227**	0.0616	0.0318	0.0178	-0.3404
BMA	0.0355	0.0211	0.0549	0.0257	0.0140	0.0245	0.0584	0.0289	0.0178	-0.5339
BMC	**0.0350**	0.0208	**0.0531**	0.0255	0.0140	0.0233	**0.0570**	0.0297	0.0176	**-0.5734**
HB	0.0359	**0.0199**	0.0568	0.0259	**0.0137**	0.0244	0.0641	0.0329	0.0196	-0.0354

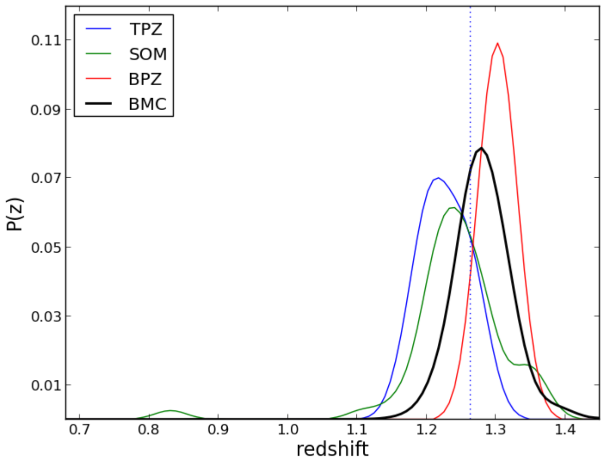

Figure 8. An comparison between the three individual photo-z PDF estimation techniques and a combined PDF computed by using BMC and Equation 12 for a single example galaxy taken from the DEEP2. The vertical line indicates the true source redshift.

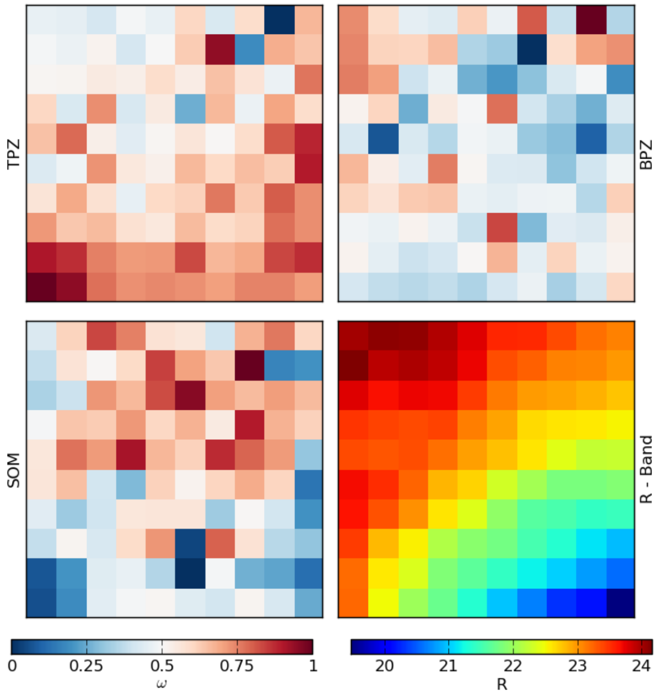

Figure 9. A two-dimensional SOM showing the relative weights for the BMA combination scheme applied to the three individual methods for the DEEP2 field 1 data (`TPZ` is top left, `BPZ` is top right, and `SOM`z is bottom left). In each panel, the color map indicates the value of the weight relative to the other cells in the map. The bottom right panel shows the same cells colored by the mean R-band magnitude for the cross validation galaxies.

across the two-dimensional parameter space defined by the SOM. For example, `BPZ` performs the best, as expected, in the upper left corner of the map, which is approximately where the faintest galaxies, at least in the DEEP2 R-band, are stored. On the other hand, `TPZ` performs better in the lower sections of the map, which approximates to brighter DEEP2 R-band magnitudes. Interestingly, `SOM`z performs relatively better in the upper middle of the map, which corresponds to the middle range $21 \lesssim R \lesssim 23$. The overall variation in weights across the map reflects the performance differences between the individual methods, which are exploited by the combination algorithms in order to identify the optimal combined performance.

We can also compare the global performance of the BMC method with the three individual photo-z PDF methods as a function of the spectroscopic redshift as shown in Figure 10. In this Figure, the photometric redshifts are the computed as the mean of each PDF, and the median is shown as black points along with the tenth and ninetieth percentiles as vertical error bars, enclosing 80% of the distribution on each redshift bin. The performance of the BMC method is generally more accurate, resulting in a tighter distribution that suffers fewer outliers when compared to the benchmark `TPZ` method. Interestingly, the `SOM`z performance is similar to `TPZ`, while `BPZ` is worse, with wider spread and several discontinuities. Nevertheless, the combined method still uses `BPZ`, as shown in the weight maps, as appropriate to generate an overall improved performance, especially for the faintest galaxies as discussed previously. We note, however, that the number counts in the last few bins are very low for the DEEP2 training and testing sets as shown in Figure 11. Therefore, although on average `BPZ` has better performance statistics over those bins (with large error bars), the photo-z results remain subject to Poissonian fluctuations (which is important when constructing a SOM to subdivide the galaxies when applying the combination models), thus the BMC results do not emphasize the `BPZ` results in the highest redshift bins.

Of all of the ten different metrics presented in Table 3, only the KS test does not show a marked improvement over the benchmark `TPZ` method. This metric does not depend on

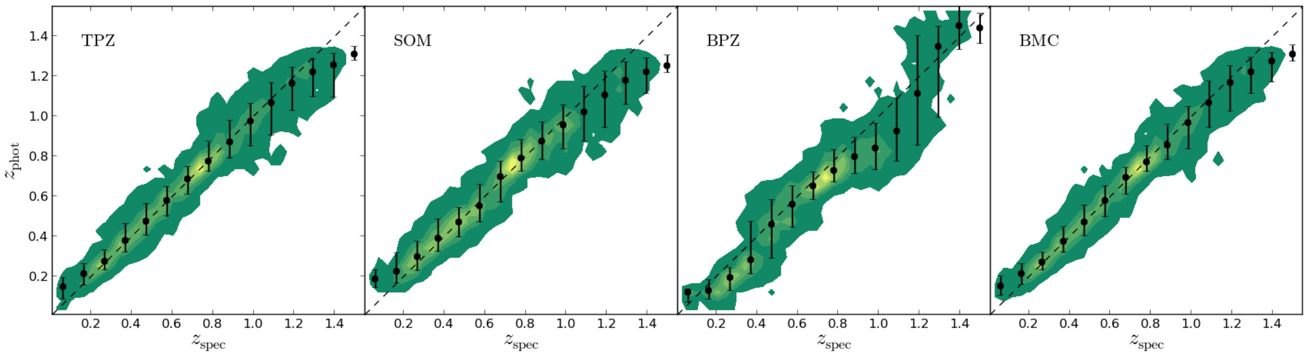

Figure 10. A comparison of the photometric and the spectroscopic redshifts for all DEEP2 field1 galaxies. From left to right, the comparison is for the `TPZ`, `SOM`z, `BPZ`, and the *BMC* techniques.The black dots are the median values of $z_{\rm phot}$ and the errors bars correspond to the tenth and ninetieth percentiles within a given spectroscopic redshift bin of width $\Delta z = 0.1$

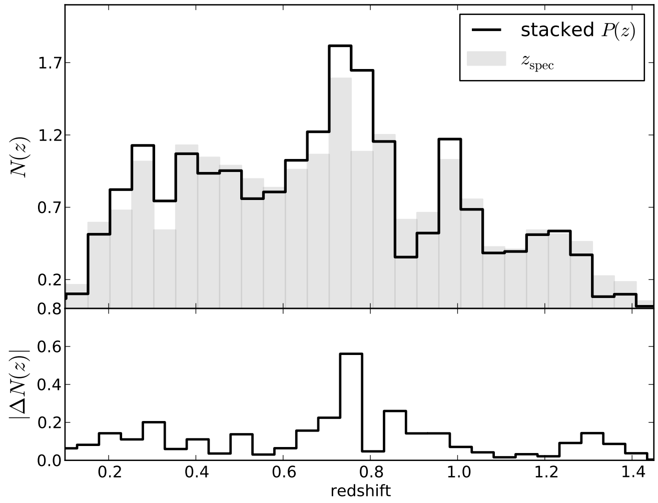

Figure 11. Top panel: The $N(z)$ for the DEEP2 sample computed directly from the spectroscopic redshifts (gray) and by stacking the photo-z PDF estimates from the *BMC* method (black). Bottom Panel: The absolute difference between these two $N(z)$ distributions.

the redshift binning and it is computed by using the stacked PDF for each method. As a result, this metric is expected to be less sensitive to a combination approach, since stacking the PDF smooths out little discrepancies between the models. After integrating over a large number of galaxies PDFs, the individual methods will not differ significantly from one another and the final $N(z)$ distribution will resemble the one from the benchmark method.

Figure 11 shows the final $N(z)$ produced by stacking the PDFs from the BMC technique for galaxies from the DEEP2 (in solid black) and the corresponding DEEP2 spectroscopic $N(z)$ for the same galaxies (in gray). As also seen in CB13 and CB14 for `TPZ` and `SOM`z respectively, both distributions match exceedingly well.

5.3 Photo-z PDF Combination for the SDSS

We now change our focus to the analysis of the SDSS galaxy sample, which consists of 1,097,397 galaxies taken from the SDSS-DR10 data; we now retain 50,000 galaxies for training purposes. We apply the same three photo-z PDF estimation methods and seven different combination methods. We construct a SOM-defined, 10×10 two-dimensional map to subdivide the multi-dimensional magnitude and color space by using a rectangular topology to facilitate visualization. As before, we use cross-validation data to identify the best set of model parameters within each individual cell in our two-dimensional map. As shown in Figures 5 and 6, the photo-z PDFs computed by using the cross-validation and testing data sets are comparable and unbiased.

We present in Table 4 the same ten metrics for each method, and in bold we highlight the best method for each metric. Overall, the results obtained for this data set are remarkable, especially for the outlier fraction and the dispersion. We once again treat `TPZ` as the benchmark method; but note that, interestingly enough, in two cases, including the *KS* metric, `TPZ` does provide the best result. In addition, both BMA and BMC have very similar results, with the latter being slightly better.

After these two models, WA$_{\rm shape}$, which is OOB data independent, shows good performance, especially when looking at the $I_{\Delta z'}$ score. For any given individual metric, however, it does not perform better than other combination methods. For this data, `BPZ` provides good results; thus we expect that the set of template described in §2.3 are a good representation of the galaxies in the SDSS photometric data. In particular, this seems true of the LRGs that dominate this sample for $z \gtrsim 0.3$.

We present the performance of the three individual and seven combination methods when applied to the SDSS data for five of the most common metrics in Figure 12. As was the case with the DEEP2 data, the Bayesian combination methods provide good performance. We also see the same variation in the *KS* metric, especially when comparing the combination methods to `TPZ`. However, `TPZ` is not always the best performer among the individual techniques, for example `SOM`z displays the best performance as measured by $\sigma_{\Delta z'}$ and out$_{0.1}$.

As we discussed in CB14, `SOM`z performs quite well when using a spherical topology; in the current application to the SDSS data, we have used a random atlas containing 300 maps that use spherical topology each with 3072 total cells. Interestingly, the WA$_{\rm oracle}$ method, which selects the best method within each binned cell, often selects the `SOM`z result as we can infer from Figure 12. Although in general the *oracle* combination method is not the best possible combination, as shown by the overall performance of the BMA and BMC combination methods on this data.

We also display the SOM-defined, 10×10 two-dimensional

Table 4. A summary of the performance results for the three individual methods and the seven different photo-z PDF combination methods as applied to the SDSS-DR10 data, with no magnitude cut applied to the training data set. The bold entries highlight the best value within each column to aid in the interpretation of the table (c.f. Figure 12).

Combination method	$< \Delta z' >$	$\lvert\Delta z'\rvert_{50}$	$\sigma_{\Delta z'}$	σ_{68}	$\sigma_{\rm MAD}$	KS	$\rm out_{0.1}$	$\rm out_{2\sigma}$	$\rm out_{3\sigma}$	$I_{\Delta z'}$
TPZ	0.0188	0.0137	0.0219	0.0139	**0.0082**	**0.0260**	0.0078	0.0297	0.0121	-0.2875
SOM	0.0201	0.0149	0.0209	0.0152	0.0094	0.0381	0.0070	0.0334	0.0125	0.7836
BPZ	0.0230	0.0164	0.0289	0.0167	0.0103	0.0367	0.0134	**0.0228**	0.0111	1.7143
$\rm WA_{flat}$	0.0195	0.0139	0.0235	0.0145	0.0088	0.0292	0.0082	0.0251	0.0104	-0.2507
$\rm WA_{oracle}$	0.0193	0.0141	0.0220	0.0145	0.0089	0.0373	0.0067	0.0266	**0.0100**	-0.1495
$\rm WA_{shape}$	0.0192	0.0136	0.0236	0.0143	0.0086	0.0297	0.0081	0.0243	0.0102	-0.4114
$\rm WA_{fit}$	0.0200	0.0141	0.0242	0.0149	0.0090	0.0274	0.0090	0.0255	0.0107	0.0244
BMA	0.0183	0.0133	0.0209	0.0139	0.0084	0.0261	0.0060	0.0296	0.0110	-0.6384
BMC	**0.0183**	**0.0133**	**0.0203**	**0.0138**	0.0084	0.0267	**0.0059**	0.0296	0.0109	**-0.6873**
HB	0.0198	0.0143	0.0237	0.0147	0.0090	0.0271	0.0084	0.0251	0.0106	-0.0975

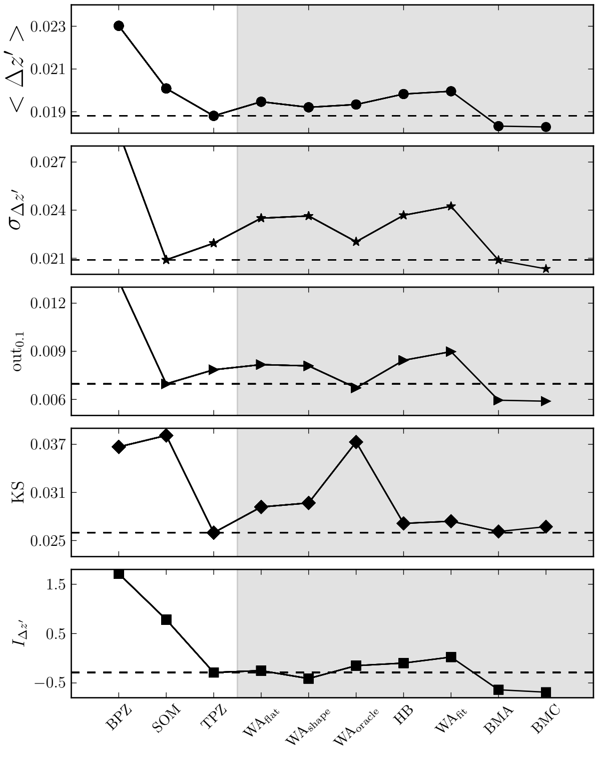

Figure 12. A comparison of the average performance for the three individual photo-z PDF estimation methods and the seven different photo-z PDF combination approaches for five different metrics as defined in Table 2 for the SDSS data. The horizontal dashed line indicates the best result for a given statistic among the three individual methods, and the shaded area separates the individual methods from the combined approaches. All values are presented in Table 4.

map used to determine the weights for the three individual methods for each cell in Figure 13. In this map, we identify galaxies within the OOB and test data to determine the parameters for the combination models. One of the benefits of using an unsupervised learning method for this mapping is that we can use any property from the galaxies within this map to construct a representation, such as the mean SDSS r-band magnitude map shown in the bottom right panel of Figure 13. In this panel the brighter galaxies are generally on the right while the fainter galaxies are on the left, even though all five magnitudes and four colors were used to construct the SOM-defined, two-dimensional map.

The weighting for the three individual methods show interesting patterns, and `TPZ` and `SOMz` seem complimentary in that `TPZ` is weighted most strongly at fainter r-band magnitudes (the left side of the map) while `SOMz` is weighted most strongly at brighter r-band magnitudes (the right side of the map). This result is most likely an artifact from the bimodality of the training data, which is dominated at low redshift by the SDSS main galaxy sample and at high redshifts by the SDSS-III LRG sample. At brighter magnitudes and lower redshifts, the `SOMz` approach where a high-dimensional space is projected to two-dimensions does a better job of maintaining complex relationships within the data. At fainter magnitudes and higher redshifts, however, the data are dominated by the homogeneous LRG sample. The `TPZ` approach performs better for this sample, since the high-dimensional space is recursively sub-divided by `TPZ` to maximize the information gain, which may only require one or two dimensions.

Another interesting observation from these weight maps is that `BPZ` performs well over much of the parameter space, with a particular strong weighting in a narrow vertical band on the extreme left of the map and again in the center of the map. Given the nature of the input galaxy sample, it seems reasonable to expect that these areas of the map are dominated by Elliptical galaxies. Another interesting observation is that there are six cells in the second column from the left that all have the same value in each weight map (pink for `TPZ`, white for `BPZ`, and light blue for `SOMz`). These cells are primarily empty, i.e., they contain weights and training data but they lack test galaxies and thus have a constant value, which illustrates how strongly the galaxies (i.e., MGS or LRG) are concentrated in this SOM-defined, two-dimensional topology.

The number of galaxies, either for training or testing, within each cell can vary significantly, which is simply due to the fact that we used a fixed number of cells (in this case 100) to represent the higher dimensional space when fewer cells would have been sufficient. However, the empty cells do not affect the performance of the photo-z combination methods, they are simply not used during the analysis. It is the fact that these individual methods perform differently across these cells that makes the combination approach a powerful technique to maximally extract information from the available data.

We next provide a comparison between the photo-z PDFs computed by the three individual techniques and the BMC technique and the SDSS spectroscopic redshift for all 1,097,397

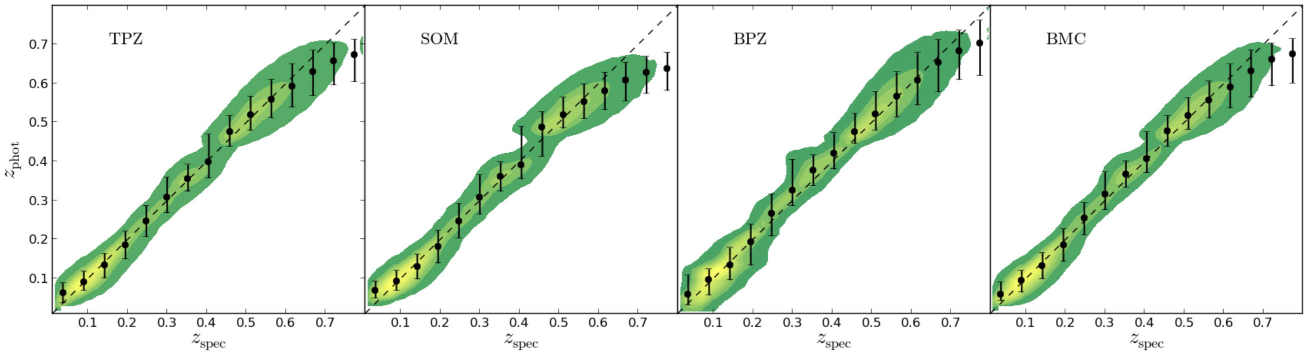

Figure 14. A comparison of the photometric and the spectroscopic redshifts for all SDSS galaxies. From left to right, the comparison is for the `TPZ`, `SOM`z, `BPZ`, and the *BMC* techniques.The black dots are the median values of $z_{\rm phot}$ and the errors bars correspond to the tenth and ninetieth percentiles within a given spectroscopic redshift bin of width $\Delta z = 0.05$

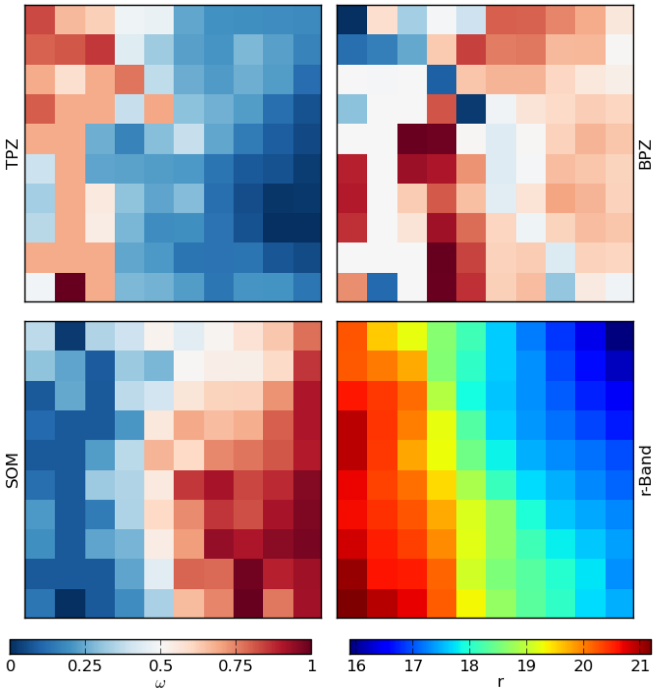

Figure 13. A two-dimensional SOM showing the relative weights for the BMA combination scheme applied to the three individual methods for the SDSS data (`TPZ` is top left, `BPZ` is top right, and `SOM`z is bottom left). In each panel, the color map indicates the value of the weight relative to the other cells in the map. The bottom right panel shows the same cells colored by the mean SDSS r-band magnitude for the cross validation galaxies.

galaxies in Figure 14. The first observation from the figure is the bi-modality of the sample, which is the result of the two primary sub-populations (i.e., MGS and LRGs). Overall, the results are quite good with a very tight correlation, especially in areas of high source density areas. The main exception is at the highest redshifts where there is a slight underestimation; and, as seen before, we can observe how these different approaches provide similar results, which are therefore correlated, while still differing in other areas where one method may outperform the others. The most right panel is the BMC which shows a slightly tighter distribution in comparison to the others.

Finally, in Figure 15 we present the galaxy redshift distribution for both the spectroscopic sample (in gray) and the photometric redshift distribution, computed by stacking the individual galaxy PDFs (in black). This Figure highlights that the underestimation of the photo-z at high redshifts in Figure 14 coincides with the strong decline in the number of galaxies after $z = 0.75$. More importantly, however, this $N(z)$ figure shows the excellent agreement between the photometric and spectroscopic galaxy redshift distributions. Given the fact that the SDSS galaxy sample contains two distinct populations, this agreement is remarkable.

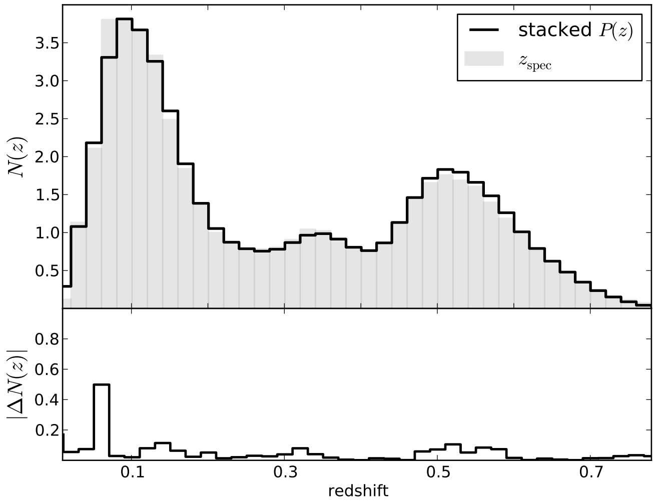

Figure 15. Top panel: The $N(z)$ computed directly from the spectroscopic redshifts (gray) and by stacking the photo-z PDF estimates from the *BMC* method (black). Bottom Panel: The absolute difference between these two $N(z)$ distributions.

6 OUTLIERS IDENTIFICATION

As we have discussed previously, aggregating information from multiple photo-z PDFs estimation techniques can improve the overall photo-z solution. In this section, however, we explore how this information can be combined to improve the identification of outliers within the test data. In particular, we attempt to use all possible information in order to identify these objects, from the shape of each photo-z PDF as computed by all individual methods to the differences in their predicted photo-z. We adopt a Naïve Bayes Classifier (NBC) (Zhang

2004) to identify these two groups, a technique that has found widespread adoption to identify spam email messages. The advantage of this approach is that it is easy to implement, is fast and efficient for large dimensional data, and can be very competitive with other classifiers (Domingos & Pazzani 1997; Frank et al. 2000).

Let $\boldsymbol{\theta}$ be the set of N_θ parameters, θ_i, we will use to identify the outliers. By using the Bayes Theorem, we can compute the probability for an object to be an outlier, given $\boldsymbol{\theta}$ as:

$$P(\mathrm{out} \mid \boldsymbol{\theta}) = \frac{P(\mathrm{out})P(\boldsymbol{\theta} \mid \mathrm{out})}{P(\boldsymbol{\theta})} \tag{26}$$

where the *evidence*, $P(\boldsymbol{\theta})$ is given by

$$P(\boldsymbol{\theta}) = P(\boldsymbol{\theta} \mid \mathrm{out}) + P(\boldsymbol{\theta} \mid \mathrm{in}) \tag{27}$$

and *out* refers to outliers and *in* refers to inliers, the only two classes we identify in this analysis. The Naïve Bayes Classifier assumes that all θ_i variables are independent, even if their independence is weak or even if there is a strong dependence between any of them. Each variable provides information about these two classes, and this information can be combined to make a stronger classifier (Zhang 2004). For instance, in CB13 we showed that outliers tend to have a broader (larger values of $zConf$) and multi-peaked PDFs, and herein we treat these values as independent data even though multi-peaked PDFs are indeed generally broader.

By using this assumption, we can write:

$$P(\boldsymbol{\theta} \mid \mathrm{out}) = P(\theta_1, \theta_2, \ldots, \theta_{N_\theta} \mid \mathrm{out}) = \prod_{i=1}^{N_\theta} P(\theta_i \mid \mathrm{out}) \tag{28}$$

and similarly,

$$P(\boldsymbol{\theta} \mid \mathrm{in}) = \prod_{i=1}^{N_\theta} P(\theta_i \mid \mathrm{in}) \tag{29}$$

We can now rewrite Equation 26:

$$P(\mathrm{out} \mid \boldsymbol{\theta}) = \frac{P(\mathrm{out}) \prod P(\theta_i \mid \mathrm{out})}{\prod P(\theta_i \mid \mathrm{out}) + \prod P(\theta_i \mid \mathrm{in})}, \tag{30}$$

which is similar to the method used by Gorecki et al. (2014), who demonstrated the potential of this approach to identify photo-z outliers. Here, however, we use a different set of variables that are generated for all three individual photo-z PDF methods.

In our case we use N_{peak}, the number of peaks in each photo-z PDF; r_{peak}, the logarithm of the ratio between the height of the first peak and the height of the second peak; z_{mean}, the mean of each photo-z PDF; z_{mode}, the mode of each PDF;$zConf$, measured with respect to the mean and the mode of the photo-z PDF; and the difference in the photo-z , as enumerated by the mean and the mode between each of the three methods. Thus, we have six metrics computed individually for each of our three photo-z PDF estimation techniques, and an additional six metrics for the difference in photo-z mean and mode between the three techniques. As a result, we have a total of twenty-four metrics, to which we can add the input data for each survey.

We, therefore, have a total of thirty-eight variables for the DEEP2 survey, while for the SDSS we have a total of thirty-three variables to use for outlier detection. For convenience, we rescale each of these variables to lie between zero and one. $P(\theta_i \mid \mathrm{in})$ and $P(\theta_i \mid \mathrm{out})$ are evaluated by using the OOB or cross-validation data, which we have shown can reliably predict the results on the test data. Once computed, these distributions are evaluated for the test data, where $P(\mathrm{out} \mid \boldsymbol{\theta})$ is evaluated separately for each galaxy in the test data.

Figure 16 presents the normalized distributions of four rescaled variables (i.e., θ_i) taken from the DEEP2 test data. Note that the inlier and outlier distributions are normalized to have unit area, thus these distributions illustrate how these

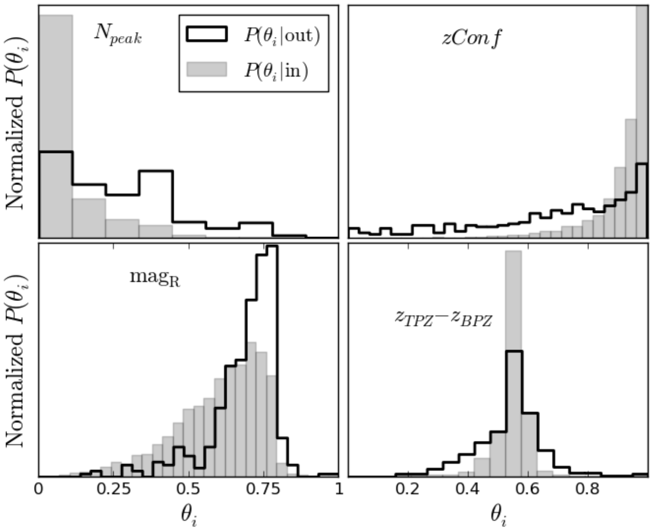

Figure 16. The normalized distributions of four of the set of thirty-eight (rescaled) $\boldsymbol{\theta}$ variables from the DEEP2 data that are used for outlier detection. The variables are binned as outliers (black line histograms) or inliers (gray histogram). From the top left and following in a clockwise direction: N_{peak}, the number of peaks in the TPZ PDF; $zConf$, as computed from TPZ, the R-band magnitude, and the difference between the photo-z computed by using the mean of the TPZ and BPZ PDFs.

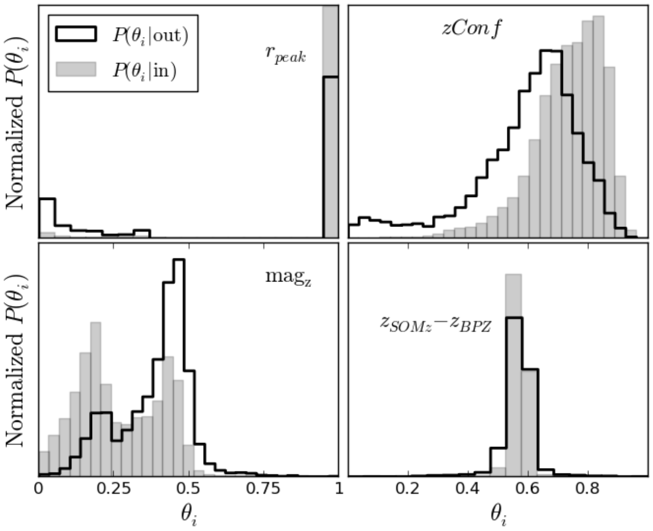

Figure 17. The normalized distributions of four of the set of thirty-three (rescaled) $\boldsymbol{\theta}$ variables from the SDSS data that are used for outlier detection. The variables are binned as outliers (black line histograms) or inliers (gray histogram). From the top left and following in a clockwise direction: r_{peak}, the logarithmic ratio of the first two peaks in the TPZ PDF; $zConf$, as computed from SOMz, the SDSS z-band magnitude, and the difference between the photo-z computed by using the mode of the SOMz and BPZ PDFs.

two populations differ and not how the relative numbers between the inlier and outlier populations vary. The four variables shown in this Figure include the number of peaks in the TPZ PDFs, $zConf$ computed by TPZ, the R-band magnitude, and the difference between the mean of the TPZ and BPZ photo-z PDFs. In just these four distributions, there is clear separation between the galaxies labeled as outliers (black line) and inliers (gray shaded area), where the outlier identification metrics are defined by using Table 2. In particular, for this Figure we use $\text{out}_{0.1}$, i.e., galaxies for which $\Delta z' > 0.1$. While not shown, a similar result is seen for the other distributions. The result that outliers and inliers follow distinct distributions is what makes this a powerful approach. In effect, all information is assumed to be independent, and when combined allows an efficient identification of catastrophic outliers.

We see a similar trend in Figure 17, but now for galaxies in the SDSS test data. In this Figure, we have selected four different rescaled variables; namely, the logarithmic ratio between the first and the second peaks of the TPZ PDF (note that if the PDF has one peak, we fix this value to be four), the $zConf$ computed from SOMz, the SDSS z-band magnitude, and the difference between the mode of the SOMz and BPZ photo-z PDFs. Once again, this Figure highlights that in each of these distributions there is a separation between the outliers and inliers, and that in combination we obtain an even better discriminant between these two classes.

By using Equation 30, we can combine the values of all of the rescaled variables (i.e., θ_i) to compute $P(\text{out} \mid \boldsymbol{\theta})$ for each galaxy in the DEEP2 and SDSS, both for the OOB and the test data. We present these $P(\text{out} \mid \boldsymbol{\theta})$ distributions for the DEEP2 in Figure 18 and for the SDSS in Figure 19. Both Figures are similar, showing a clear separation between the outliers and inliers in both data sets. The probability ranges between zero and one, and the outliers are generally concentrated near one, while the inliers are concentrated near zero. While some mis-classifications remain, the contamination has been greatly reduced, meaning we can successfully identify a majority of the outlier population. Lastly, while there are a few galaxies with probabilities lying somewhere between zero and one, these distributions are highly bimodal, which reinforces the belief that this method provides a remarkably good discriminant between these two populations.

Once again, in both Figures 18 and 19, the OOB and test data distributions show strong similarities. As a result, we can expect that any cut we make on the OOB data will produce similar results in the test data, allowing us to make a robust classification of outliers in potentially blind test data. To quantify this, we show in Table 5 the effects of selecting outliers by using this NBC approach and by using the $zConf$ approach we initially presented in CB13 for the DEEP2 data. To simplify the comparison, we first select inlier galaxies by using the $P(\text{out} \mid \boldsymbol{\theta})$ to cut the test data sample, and subsequently choosing those galaxies in the test data that have the highest $zConf$ so that we have the same number of galaxies selected via both techniques.

The information in this Table demonstrates that the NBC approach produces a sample of galaxies that have a smaller spread in $\Delta z'$ along with a smaller number of outliers than the $zConf$ method, which was previously shown to be beneficial in this regard (CB13). We interpret this result as suggesting that a $zConf$ cut can potentially remove *good* galaxies whose photo-z PDF happens top be broad, while retaining some *bad* galaxies that have a well-localized photo-z PDF. By

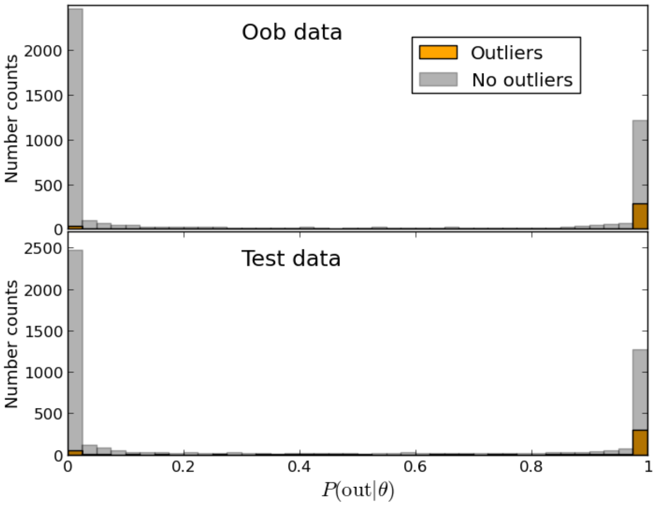

Figure 18. The count distribution of $P(\text{out} \mid \boldsymbol{\theta})$ for the DEEP2 OOB data (top) and test data (bottom) showing both the outliers (orange) and inliers (gray).

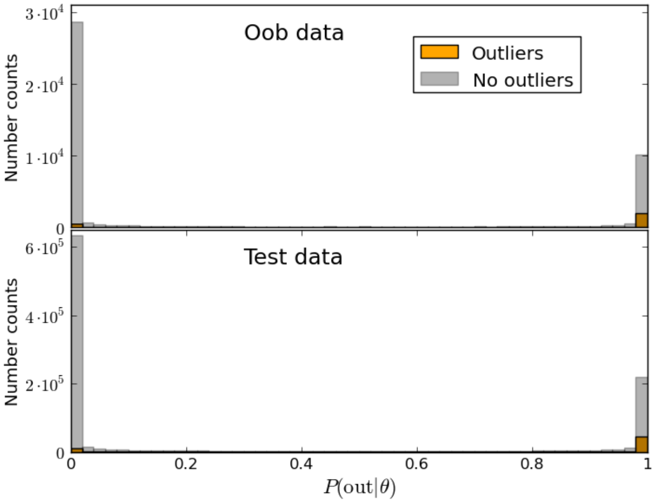

Figure 19. The count distribution of $P(\text{out} \mid \boldsymbol{\theta})$ for the SDSS OOB data (top) and test data (bottom) showing both the outliers (orange) and inliers (gray).

Table 5. The effect of removing outliers from the DEEP2 test data on several, select performance metrics by using the Naïve Bayes Classifier and the $zConf$ cut approach. The two techniques are applied to ensure equal numbers of galaxies are selected, which is indicated by the *Fraction* column.

Method	Criteria	Fraction	$<\Delta z'>$	$\sigma_{\Delta z'}$	$\text{out}_{0.1}$
NBC	< 0.998	83.0 %	0.02819	0.03948	0.0362
$zConf$	> 0.854	83.0 %	0.02868	0.04186	0.0371
NBC	< 0.894	72.0 %	0.02616	0.03548	0.0304
$zConf$	> 0.893	72.0 %	0.02721	0.03895	0.0330
NBC	< 0.174	56.0 %	0.02565	0.03470	0.0251
$zConf$	> 0.918	56.0 %	0.02595	0.03575	0.0289

Table 6. The effect of removing outliers from the SDSS test data on several, select performance metrics by using the Naïve Bayes Classifier and the *zConf* cut approach. The two techniques are applied to ensure equal numbers of galaxies are selected, which is indicated by the *Fraction* column.

Method	Criteria	Fraction	$< \Delta z' >$	$\sigma_{\Delta z'}$	$\mathrm{out}_{0.1}$
NBC	< 0.999	83.0 %	0.01560	0.01533	0.0022
zConf	> 0.7018	83.0 %	0.01589	0.01704	0.0035
NBC	< 0.802	72.0 %	0.01473	0.01411	0.0012
zConf	> 0.755	72.0 %	0.01475	0.01549	0.0026
NBC	< 0.001	56.0 %	0.01387	0.01309	0.0006
zConf	> 0.807	56.0 %	0.01366	0.01410	0.0020

using a Naïve Bayes approach, we collect all information from photo-z PDFs predicted by using different, semi-independent methods, allowing a more robust discriminant between outliers and inliers. Finally, we notice that as always there is a trade-off between completeness, whereby we try to retain as many *good* galaxies, and contamination, whereby we try to minimize the inclusion of *bad* galaxies. The final choice in this conflict should be determined by the scientific application, but by producing a probabilistic value, subsequent researchers can make these cuts more easily.

We performed a similar analysis on the SDSS galaxy sample and present the results in Table 6. As was the case with the DEEP2 galaxies, we see that the NBC approach once again does better in identifying outliers within the sample, as the NBC cuts have a smaller scatter and the fraction of remaining outliers is remarkably small. We also notice that the mean bias is similar between the two approaches, but the number of outliers, defined as $\Delta z' > 0.1$, is significantly reduced when we adopt the Bayesian approach. This is yet another piece of evidence supporting the benefits of aggregating information to make decisions.

We can also test how the definition of an outlier affects this approach. Previously we identified an outlier as a galaxy that had $\Delta z' > 0.1$; but for the purpose of this test, we apply a much more restrictive cut of $\Delta z' > 0.05$. We apply the NBC cut and produce a matched sample by imposing a *zConf* cut to both the DEEP2 and the SDSS galaxies, presenting the information in Table 7. We find, once again, that even for this more restrictive approach we produce a cleaner catalog (of the same size) as compared to using only the *zConf* parameter. Interestingly, even after removing almost 30% of the galaxies from the DEEP2 galaxy sample, we still have over a 10% outlier contamination. On the other hand, this tight cut applied to the SDSS galaxies produces a very small contamination of ∼ 2%, for both methods, albeit the NBC approach is still slightly better.

While producing galaxy samples that are less affected by outliers than competing techniques, the NBC approach has an additional advantage in that it can easily be extended to other variables and to other photo-z algorithms. In effect, any information that might increase the efficacy of outlier identification can be included in order to improve this discriminant while still maximizing the overall galaxy sample size.

7 CONCLUSIONS

We have presented and analyzed different techniques for combining photo-z PDF estimations on galaxy samples from the DEEP2 and SDSS projects. In particular, we use three independent photo-z PDF estimation methods: `TPZ`, a supervised machine learning technique based on prediction trees and a random forest; `SOM`z, an unsupervised machine learning approach based on self organizing maps and a random atlas; and `BPZ`, a standard template-fitting method that we have slightly modified to parallelize the implementation. Both `TPZ` and `SOM`z are currently available within a new software package entitled `MLZ`[5].

Table 7. The effect of removing outliers, defined as $\Delta z' > 0.05$, from the DEEP2 and SDSS test data on several, select performance metrics by using the Naïve Bayes Classifier and the *zConf* cut approach. For each data set, the two techniques are applied to ensure equal numbers of galaxies are selected, which is indicated by the *Fraction* column.

Method	Criteria	Fraction	$< \Delta z' >$	$\sigma_{\Delta z'}$	$\mathrm{out}_{0.05}$
DEEP2					
NBC	< 0.996	72.0 %	0.02780	0.03934	0.138
zConf	> 0.878	72.0 %	0.02809	0.04244	0.141
SDSS					
NBC	< 0.85	72.0 %	0.01461	0.01407	0.0247
zConf	> 0.75	72.0 %	0.01479	0.01554	0.0278

We developed seven different combination methods that employ ensemble learning with cross-validation data to maximize the information extracted. Of these seven methods, four employ a weighted average where the weights can either be selected to be uniform across the input methods, to be determined from the shape of the photo-z PDF (e.g., by using the *zConf* parameter), to be determined by an *oracle* estimator where one (ideally the best) method is preferentially selected, and where the weights are obtained by a fitting procedure applied to the OOB data. Three of the combination methods were Bayesian techniques: Bayesian Model Averaging (BMA), Bayesian Model Combination (BMC), and Hierarchical Bayes (HB).

We expect the individual photo-z PDF estimation techniques to perform differently across the parameter space spanned by our galaxy samples; for example, template-fitting techniques are expected to work better at higher redshifts than machine learning methods, which perform optimally when provided high-quality, representative training data. Thus we construct a two-dimensional, 10 × 10 self-organizing map (SOM) to subdivide the high-dimensional parameter space occupied by the galaxy samples. We apply different photo-z PDF estimation techniques within each cell in this map, since each cell should contain galaxies with similar properties. A visual inspection of these maps indicates that the two machine learning methods: `TPZ` and `SOM`z are generally complementary, and that in combination with a model based technique such as `BPZ` we are able to maximize the coverage of this multidimensional space efficiently.

We also verified that by using the OOB data, as introduced in CB13, we can an obtain an accurate, unbiased and *honest* estimation of the performance of a photo-z PDF estimation technique on the test data. We also computed the correlation coefficient and the error distribution and showed they also behave similarly for the cross-validation (i.e., the OOB data) and the test data. These computations are extremely

[5] http://lcdm.astro.illinois.edu/code/mlz.html

important when combining photo-z PDF techniques as we can learn from the OOB data the optimal parameters needed for a specific ensemble learning approach, and thereby maximize the performance of that combination technique when applied to *blind* test data.

Overall, we found that the BMA and BMC are the best photo-z PDF combination techniques as they have better performance metrics when compared to the individual photo-z PDF estimation techniques, especially when unbiased cross-validation data is available. This result is true for both the DEEP2 and the SDSS data. When OOB data is not available, we can instead use the $zConf$ parameter as a weight for each method after first renormalizing the individual photo-z PDFs. We can also use the Hierarchical Bayes method to combine these predictions, which we demonstrated can also lead to better results.

Within this Bayesian Framework, we also developed a novel, Naïve Bayesian Classifier (NBC) that efficiently identifies outliers within the galaxy sample. The approach we present gathers all available information from the different photo-z PDF estimation techniques regarding the shape of the PDF, the location of the mean and mode, and the magnitudes and colors, which are all *naively* assumed to be independent, in order to compute a Bayesian posterior probability that a certain galaxy is an outlier. The distribution of these probabilities for an entire galaxy sample indicate that this is a very powerful method to separate outliers from inliers (i.e., *good* galaxies), and we further demonstrated that this approach can produce a more accurate and cleaner sample of galaxies than competing techniques, such as the use of the $zConf$ parameter. An important takeaway point is that all information provided by the catalogs and the photo-z PDF methods, no matter how redundant the information might appear, helps in building this discriminant probability. Given the probabilistic nature of this computation, the final application of this technique can be chosen to maximize the scientific utility of the resulting galaxy data for a specific application.

The computational cost to apply these Bayesian models to galaxy samples will depend directly on the size of the data set, the number of photo-z estimation techniques used, and the resolution of the given photo-z PDFs. In Carrasco Kind & Brunner (2014b) we demonstrate how a sparse basis representation can reduce the storage significantly and that manipulation of these PDFs can be improved within the bases framework thereby reducing computational costs. We plan to adopt this representation framework to compute the combination models, which will allow fast and accurate combination of multiple photo-z PDFs.

Finally, we have demonstrated that even when a photo-z PDF technique is very accurate, we can still make improvements by extracting additional information about the distribution of galaxies in the higher dimensional parameter space and the individual performance of the photo-z PDF algorithms. There are currently a large number of published algorithms to compute photo-z 's, many of which also compute photo-z PDFs. Even if their performance is similar, these techniques will all have their own advantages and disadvantages. Thus we believe the combination of different techniques is the future of photo-z research, and we expect additional research to be forthcoming in this area. Overall, the combination of photo-z PDFs is a powerful, new approach that can be easily extended to incorporate new techniques in order to generate a meta-predictor that accelerate our knowledge and understanding of the Universe.

ACKNOWLEDGEMENTS

The authors thank the referee for a careful reading of the manuscript and for comments that improved this work. RJB and MCK acknowledge support from the National Science Foundation Grant No. AST-1313415. MCK has been supported by the Computational Science and Engineering (CSE) fellowship at the University of Illinois at Urbana-Champaign. RJB has been supported in part by the Institute for Advanced Computing Applications and Technologies faculty fellowship at the University of Illinois.

The authors gratefully acknowledge the use of the parallel computing resource provided by the Computational Science and Engineering Program at the University of Illinois. The CSE computing resource, provided as part of the Taub cluster, is devoted to high performance computing in engineering and science. This work also used resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

Funding for the DEEP2 Galaxy Redshift Survey has been provided by NSF grants AST-95-09298, AST-0071048, AST-0507428, and AST-0507483 as well as NASA LTSA grant NNG04GC89G.

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 Astrofisica 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.

REFERENCES

Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, MNRAS, 417, 1891

Ahn C. P. et al., 2013, ArXiv e-prints

Assef R. J. et al., 2010, ApJ, 713, 970

Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12

Baum W. A., 1962, in IAU Symposium, Vol. 15, Problems of Extra-Galactic Research, McVittie G. C., ed., p. 390

Benítez N., 2000, ApJ, 536, 571

Blake C. et al., 2011, MNRAS, 418, 1707

Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, 476

Breiman L., 2001, Machine Learning, 45, 5

Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees, Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A.

Brunner R. J., Connolly A. J., Szalay A. S., Bershady M. A., 1997, ApJL, 482, L21

Carrasco Kind M., Brunner R. J., 2013a, MNRAS, 432, 1483, (CB13)

Carrasco Kind M., Brunner R. J., 2013b, in Astronomical Society of the Pacific Conference Series, Vol. 475, Astronomical Society of the Pacific Conference Series, Friedel D. N., ed., p. 69

Carrasco Kind M., Brunner R. J., 2014a, MNRAS, 438, 3409, (CB14)

Carrasco Kind M., Brunner R. J., 2014b, ArXiv e-prints : 1404.6442

Caruana R., Karampatziakis N., Yessenalina A., 2008, in Proceedings of the 25th international conference on Machine learning, ICML '08, ACM, New York, NY, USA, pp. 96–103

Coleman G. D., Wu C.-C., Weedman D. W., 1980, ApJS, 43, 393

Collister A. A., Lahav O., 2004, PASP, 116, 345

Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655

Cunha C. E., Huterer D., Busha M. T., Wechsler R. H., 2012a, MNRAS, 423, 909

Cunha C. E., Huterer D., Lin H., Busha M. T., Wechsler R. H., 2012b, ArXiv e-prints

Dahlen T. et al., 2013, ApJ, 775, 93

Davis M. et al., 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4834, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Guhathakurta P., ed., pp. 161–172

Dawson K. S. et al., 2013, AJ, 145, 10

Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159

Domingos P., Pazzani M., 1997, Machine Learning, 29, 103

Drinkwater M. J. et al., 2010, MNRAS, 401, 1429

Faber S. M. et al., 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4841, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Iye M., Moorwood A. F. M., eds., pp. 1657–1669

Fadely R., Hogg D. W., Willman B., 2012, ApJ, 760, 15

Feldmann R. et al., 2006, MNRAS, 372, 565

Frank E., Trigg L., Holmes G., Witten I., 2000, Machine Learning, 41, 5

Freeman P. E., Newman J. A., Lee A. B., Richards J. W., Schafer C. M., 2009, MNRAS, 398, 2012

Geach J. E., 2012, MNRAS, 419, 2633

Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823

Gorecki A., Abate A., Ansari R., Barrau A., Baumont S., Moniez M., Ricol J.-S., 2014, A&A, 561, A128

Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759

Gregory P. C., Loredo T. J., 1992, ApJ, 398, 146

Gwyn S. D. J., 2012, AJ, 143, 38

Hildebrandt H. et al., 2010, A&A, 523, A31

Ilbert O. et al., 2006, A&A, 457, 841

Jee M. J., Tyson J. A., Schneider M. D., Wittman D., Schmidt S., Hilbert S., 2013, ApJ, 765, 74

Kinney A. L., Calzetti D., Bohlin R. C., McQuade K., Storchi-Bergmann T., Schmitt H. R., 1996, ApJ, 467, 38

Kohonen T., 1990, Proceedings of the IEEE, 78, 1464

Kohonen T., 2001, Self-Organizing Maps, Physics and astronomy online library. Springer-Verlag GmbH

Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118

Mandelbaum R. et al., 2008, MNRAS, 386, 781

Matthews D. J., Newman J. A., Coil A. L., Cooper M. C., Gwyn S. D. J., 2013, ApJS, 204, 21

Monteith K., Carroll J. L., Seppi K., Martinez T., 2011, The 2011 International Joint Conference on Neural Networks, 2657

Myers A. D., White M., Ball N. M., 2009, MNRAS, 399, 2279

Newman J. et al., 2013b, ArXiv e-prints : 1309.5384

Newman J. A. et al., 2013a, ApJS, 208, 5

Oke J. B. et al., 1995, PASP, 107, 375

Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008, ApJ, 689, 709

Parkinson D., Liddle A. R., 2013, Statistical Analysis and Data Mining, 6, 3

Percival W. J. et al., 2010, MNRAS, 401, 2148

Rokach L., 2010, Artificial Intelligence Review, 33, 1

Sánchez A. G. et al., 2013, MNRAS, 433, 1202

Sánchez C. et al., 2014, in preparation

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Trotta R., 2007, MNRAS, 378, 72

Wadadekar Y., 2005, PASP, 117, 79

Way M. J., Klose C. D., 2012, PASP, 124, 274

York D. G. et al., 2000, AJ, 120, 1579

Zhang H., 2004, in Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004), Barr V., Markov Z., eds., AAAI Press

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