Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **468,** 4323–4339 (2017)
Advance Access publication 2017 March 22

doi:10.1093/mnras/stx687

On the realistic validation of photometric redshifts

R. Beck,[1]⋆ C.-A. Lin,[2,3] E. E. O. Ishida,[4] F. Gieseke,[5] R. S. de Souza,[6,7] M. V. Costa-Duarte,[7,8] M. W. Hattab,[9] A. Krone-Martins[10] and for the COIN Collaboration

[1] *Department of Physics of Complex Systems, Eötvös Loránd University, Budapest 1117, Hungary*
[2] *Service d' Astrophysique, CEA Saclay, Orme des Merisiers, Bât 709, F-91191 Gif-sur-Yvette, France*
[3] *Fenglin Veteran Hospital, 2 Zhongzheng Rd. Section 1, Fenglin Township, Hualien 97544, Taiwan*
[4] *Laboratoire de Physique Corpusculaire, Université Clermont-Auvergne, 4 Avenue Blaise Pascal, F-63178 Aubière Cedex, France*
[5] *University of Copenhagen, Sigurdsgade 41, DK-2200 Copenhagen, Denmark*
[6] *MTA Eötvös University, EIRSA 'Lendulet' Astrophysics Research Group, Budapest 1117, Hungary*
[7] *Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, R. do Matão 1226, 05508-090 SP, Brazil*
[8] *Leiden Observatory, Leiden University, Niels Bohrweg 2, NL-2333 CA Leiden, the Netherlands.*
[9] *Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA*
[10] *CENTRA/SIM, Faculdade de Ciências, Universidade de Lisboa, Ed. C8, Campo Grande, P-1749-016 Lisboa, Portugal*

Accepted 2017 March 16. Received 2017 March 16; in original form 2017 February 8

ABSTRACT

Two of the main problems encountered in the development and accurate validation of photometric redshift (photo-z) techniques are the lack of spectroscopic coverage in the feature space (e.g. colours and magnitudes) and the mismatch between the photometric error distributions associated with the spectroscopic and photometric samples. Although these issues are well known, there is currently no standard benchmark allowing a quantitative analysis of their impact on the final photo-z estimation. In this work, we present two galaxy catalogues, Teddy and Happy, built to enable a more demanding and realistic test of photo-z methods. Using photometry from the Sloan Digital Sky Survey and spectroscopy from a collection of sources, we constructed data sets that mimic the biases between the underlying probability distribution of the real spectroscopic and photometric sample. We demonstrate the potential of these catalogues by submitting them to the scrutiny of different photo-z methods, including machine learning (ML) and template fitting approaches. Beyond the expected bad results from most ML algorithms for cases with missing coverage in the feature space, we were able to recognize the superiority of global models in the same situation and the general failure across all types of methods when incomplete coverage is convoluted with the presence of photometric errors – a data situation which photo-z methods were not trained to deal with up to now and which must be addressed by future large-scale surveys. Our catalogues represent the first controlled environment allowing a straightforward implementation of such tests. The data are publicly available within the COINtoolbox (https://github.com/COINtoolbox/photoz_catalogues).

Key words: methods: data analysis – methods: statistical – techniques: photometric – catalogues – galaxies: distances and redshifts.

1 INTRODUCTION

Photometric redshift (photo-z) estimation has become a widespread and vital tool in the astronomical field. Although compared to their higher resolution counterpart, the spectroscopic technique (spec-z), photo-z measurements are subject to higher uncertainty, they are also more efficient, cheaper and able to probe more distant objects (e.g. Hildebrandt H. et al. 2008). These characteristics make them more suitable for some astrophysical problems. An example of the former is the weak gravitational lensing (Abdalla et al. 2008), which measures the coherent galaxy shape distortion by gravitational potentials, and is relatively less sensitive to the redshift measurement. The lensing signal is a convolution of the density contrast distribution with a broad kernel that has the effect to smooth the sensitivity of the signal to the redshift accuracy. Therefore, in order to obtain a faster measurement and to maximize the data volume, photo-z is commonly used in weak lensing studies.

⋆E-mail: beckrob23@caesar.elte.hu

However, with the arrival of the Stage-IV lensing surveys (Natarajan et al. 2014), the goal on cosmological constraints becomes more ambitious, and consequently, the requirements on photo-z precision and accuracy equally increase. For instance, for the Euclid[1] mission of the European Space Agency, the initial requirements on the bias and the scatter in each redshift bin were 0.002 and 0.05 for a total number of 2 billion galaxies, respectively (Laureijs et al. 2011). Spectroscopic follow-up for such a large number of objects is infeasible and such stringent requirements on redshift measurements are extremely challenging for current photo-z methods. Apart from weak lensing, other applications exist such as large-scale structure (Malavasi et al. 2016) and gravitational waves (Antolini & Heyl 2016), which also require improvements in the current techniques.

In the quest for a viable photo-z alternative capable to handle the size and complexity of modern astronomical surveys, a plethora of different methods have been proposed and tested. These are commonly divided into two main classes: (i) template fitting (e.g. Benítez 2000; Bolzonella, Miralles & Pelló 2000; Csabai et al. 2000; Coe et al. 2006; Ilbert et al. 2006; Brammer, van Dokkum & Coppi 2008; Leistedt, Mortlock & Peiris 2016; Beck et al. 2017), (ii) empirical (e.g. Wadadekar 2005; Boris et al. 2007; Miles, Freitas & Serjeant 2007; Budavári 2009; Carliles et al. 2010; O'Mill et al. 2011; Krone-Martins, Ishida & de Souza 2014; Cavuoti et al. 2015; Elliott et al. 2015; Hogan, Fairbairn & Seeburn 2015) and (iii) hybrid techniques (e.g. Beck et al. 2016).

In template fitting techniques, a set of synthetic spectra is determined from synthesized stellar population models for a given set of metallicities, star formation histories and initial mass functions, among other properties. The photo-z is calculated by determining the synthetic photometry (and thus spectral template and redshift) which best fits the photometric observations. Empirical techniques, on the other hand, usually require a data set with spectroscopically measured redshifts in order to train an algorithm which will subsequently be applied to a pure photometric sample. Hybrid methods represent a combination of the previous ones, using an empirical step to first determine the photometric redshift and a template fitting step where physical information provided by the templates can be used to evaluate the accuracy of the photo-z determination.

There have been several notable publications that contrasted the performance of empirical and/or spectral model fitting codes (e.g. Csabai et al. 2003; Hildebrandt et al. 2010; Abdalla et al. 2011; Dahlen et al. 2013). However, when photo-z algorithms are evaluated in the literature, the available spectroscopic set is randomly split into training and validation sets. Roughly, the algorithm is optimized using the training set and its accuracy estimated based on its performance on the validation set. This approach neglects to take into account the fact that the distribution of galaxies in the space of observables (i.e. photometric magnitudes/colours) is generally very different for the spectroscopic and the photometric samples, and even their range in the magnitude/colour space can differ. Moreover, given the quality requirements demanded for spectroscopic observation, the photometric sample encloses a larger photometric uncertainty that may break the direct relation between magnitude/colours and redshift, even when this relation is clearly defined in the spectroscopic sample. In summary, the performance metrics obtained on the former is not representative of results for the latter. In the worst-case scenario, these issues are ignored altogether, as in the benchmark papers aforementioned. In other cases, error flags or feature selection results are provided alongside the photo-z estimation, notifying users when extrapolation is performed and results should not be trusted (Brescia et al. 2014; Beck et al. 2016; Stensbo-Smidt et al. 2017) – essentially cutting the coverage of the photometric sample.

Intermediate approaches, aiming at dealing with at least one of these problems have also been reported. In situations where spectroscopic and photometric samples share the same coverage in magnitude/colour space, it is possible to adapt the spectroscopic sample[2] distribution in order to get it closer to the photometric one (e.g. Lima et al. 2008; Sánchez et al. 2014; Kremer et al. 2015). However, in real-data scenarios, especially when upcoming surveys are considered, even that assumption will not hold. To obtain a measure of photo-z performance that is realistic for the actual use case, i.e. when the photometric sample has a much wider coverage in colour space than the spectroscopic sample and, at the same time, there is a correlation between colours and photometric errors, it is crucial that we evaluate photo-z methods in more realistic data situations.

In this paper, we provide for the first time a complete benchmark template to allow a realistic evaluation of the performance of photometric redshift estimators. Using photometry from the Sloan Digital Sky Survey Data Release 12 (SDSS-DR12; Alam et al. 2015) and spectroscopic redshift measurements from a variety of different sources, we were able to construct validation samples that follow the colour coverage and shape distribution from the original SDSS-DR12 photometric sample. These enable an unprecedented realistic view into the accuracy of current photo-z methods and provide a starting point to the development of new techniques which take these issues into account.

The outline of this paper is as follow. In Section 2, we show how we built our benchmark samples from a combination of spectroscopic and photometric data. In Section 3, we describe the methods we have used to access the impact of non-representativeness. In Section 4, we compare the summary statistics and performance of different photo-z estimators. We present a discussion of our results in Section 5.

Throughout the paper, we use SDSS modelMag broad-band magnitudes in the SDSS asinh magnitude system (Lupton, Gunn & Szalay 1999), which have been corrected for Galactic extinction according to Schlegel, Finkbeiner & Davis (1998).

2 THE CATALOGUES

To enable realistic performance estimation for photo-z methods, we present two data sets built to mimic the main causes of non-representativeness between spectroscopic (training) and photometric (test) samples: the disparity in colour-space coverage (Teddy) and the differences between photometric error distributions (Happy). Each catalogue is composed of four samples with known spec-z: one following the characteristics of the real spectroscopic sample, which should be used for training/calibration purposes (A) and three holding increasing degrees of non-representativeness of A, which should be used as test (or validation) sets (B, C and D). In what follows, we describe how the catalogues were constructed and the main effects they allow us to probe.

[1] http://sci.esa.int/euclid/

[2] In machine-learning jargon such methods are a subclass of *domain adaptation* techniques.

Table 1. Numbers of galaxies contained in different samples of the Teddy and Happy catalogues. The total number of galaxies in the SDSS-DR12 spectroscopic sample is 2040 465, which we extended to 2209 299 (see Section 2.2).

	Sample A	Sample B	Sample C	Sample D
Teddy	74 309	74 557	97 980	75 924
Happy	74 950	74 900	60 315	74 642

2.1 Teddy: the effect of colour coverage

The disparity in the feature-space distribution between training and test sets has been known to impact classification and regression tasks in machine learning (Quionero-Candela et al. 2009). Specifically in the photo-z case, this translates into a spectroscopic sample which fails to cover the entire domain occupied by the photometric sample in colour–magnitude parameter space. It has been already reported that such a gap introduces significant biases in redshift estimation for empirical (e.g. MacDonald & Bernstein 2010) as well as for template fitting techniques (e.g. Dahlen et al. 2013).

In the context of empirical methods, in which there is no input information beyond magnitudes and colours from the spectroscopic sample, it is straightforward to expect that results will not be reliable beyond the training domain. Machine-learning methods only learn by example, thus their results should not be extrapolated. For template fitting methods, the region of the parameter space not covered by the spectroscopic sample can be directly related to fainter objects having active galactic nucleus activity and metallicity levels not represented in the template library (MacDonald & Bernstein 2010). Although these issues are widely known, at the moment there is no standard ground to quantify the sensitivity of each photo-z method in the presence of such gap. The Teddy catalogue presented here was designed to fulfil this role. It was completely built using SDSS-DR12 spectroscopic sample, which ensures the quality of photometric measurements and allow us to isolate the effect of colour–magnitude coverage from its correlation with photometric errors.

For each of the four SDSS colours, we defined intervals forming a four-dimensional rectangular parallelepiped ($1.0 < u - g < 2.2$, $1.2 < g - r < 2.0$, $0.1 < r - i < 0.8$ and $0.2 < i - z < 0.5$). We then selected ≈150 000 objects from the SDSS-DR12 spectroscopic sample whose colours were within this space and with an extra constraint on r-band magnitude, $r < 21$ – this is called the *narrow set*. All other objects were grouped into what we called the *wide set*.

From the narrow and wide sets, we constructed one training sample (named A) and three test samples (named B, C, and D) for comparing different scenarios. Sample A and B were created by splitting the narrow set into two comparable parts, sample C was constructed by truncating the wide set on the colour coverage of the narrow set and finally, sample D was made by sampling randomly from the wide set. The number of objects contained in each sample is shown in Table 1. In summary, samples A and B follow exactly the same feature-space distribution, sample C has the same coverage of A, but a different distribution and sample D has a wider coverage in colour–magnitude space.

The r-band magnitude and colour distributions for all samples in the Teddy catalogue are shown in Fig. 1. We observe at ease the colour cut imposed on samples A, B and C. By construction, samples C and D follow the same distribution in the space region, but their probability density functions (PDFs) differ due to marginalization. A significant number of galaxies with colours outside the four-dimensional parallelepiped contribute to the disparity between the two curves.

We consider A as the spectroscopic sample (used for training) and B, C and D as distinct photometric samples (used for validation/test). These correspond to increasingly complex data situations. Training in A and testing in B represent the best case scenario where distributions are completely representative of each other – and thus must yield the best possible results. Training in A and testing in C correspond to the situation of simply ignoring data outside spectroscopic coverage – in this case, methods like applying weights should provide slightly better results than the standard approach. Training in A and testing in D corresponds to a situation with incomplete coverage, where pure machine-learning methods are not expected to provide good results.

It is important to emphasize that the colour and magnitude distributions in Teddy are not realistic, but they do provide a test bench to probe the robustness of photo-z methods against feature distribution shape and coverage. For an ideal photo-z estimator, results from all three configurations should be equivalent.

2.2 Happy: the effect of photometric errors

Once we have a data set that enables probing the robustness of photo-z methods in the case of incomplete colour and r-magnitude coverage, we approach other important differences between spectroscopic and photometric samples: the presence of photometric errors and their correlation with colour coverage. In constructing Teddy, we rearranged data from the original SDSS-DR12 spectroscopic data set. Thus, all its samples share the same spectroscopic data quality level. This is not what happens in the real situation, where photometric observations are statistically fainter and of poorer quality than the spectroscopic ones.

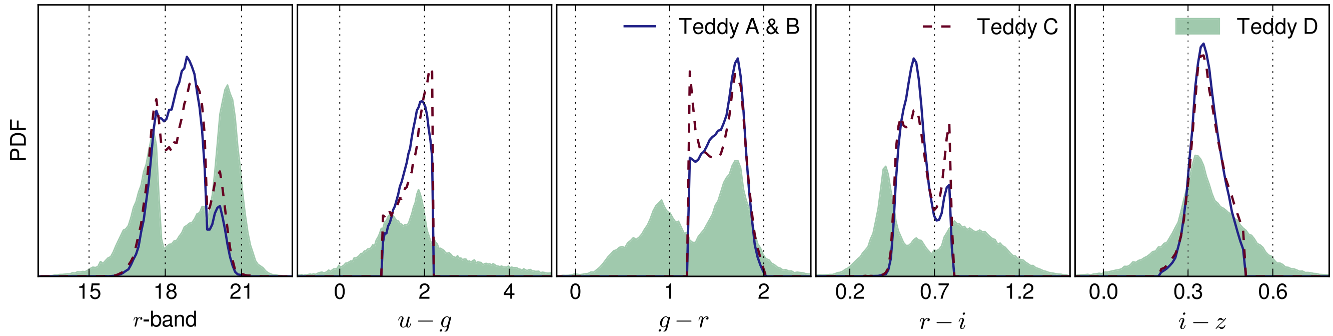

Figure 1. Distributions for r-band magnitude and four SDSS colours for the Teddy catalogue.

The Happy catalogue was created to allow a clear assessment of the impact of photometric errors on photo-z estimation, while at the same time more closely resembling the colour space differences between the original SDSS-DR12 spectroscopic and photometric data sets.

Our first goal was to reproduce the colour–magnitude space distribution of the SDSS-DR12 photometric sample, only with objects with measured spectroscopic redshifts, so that the photo-z methods could be properly evaluated. However, as the DR12 spectroscopic set does not contain objects with more extreme colours and fainter magnitudes, it is not an adequate source of example galaxies. Thus, we chose to extend the DR12 spectroscopic set (S1) by cross-matching SDSS photometric measurements with galaxies from other spectroscopic surveys. This approach provides a deeper sample of spectroscopic galaxies, while also keeping the use of actual SDSS photometry and its inherent systematics.

We followed the Bayesian cross-matching methodology described in Budavári & Szalay (2008), only accepting relatively secure matches, with a Bayes factor above 10 000 (equation 16 in Budavári & Szalay 2008). The following surveys were used in the match:

(i) 2dF (Colless et al. 2001, 2003, 770 matches),

(ii) 6dF (Jones et al. 2004, 2009, 765 matches),

(iii) DEEP2 (Davis et al. 2003; Newman et al. 2013, 7456 matches),

(iv) GAMA (Driver et al. 2011; Baldry et al. 2014, 53 373 matches),

(v) PRIMUS (Coil et al. 2011; Cool et al. 2013, 32 459 matches),

(vi) VIPERS (Garilli et al. 2014; Guzzo et al. 2014, 18 967 matches),

(vii) VVDS (Le Fèvre et al. 2004; Garilli et al. 2008, 8381 matches),

(viii) WiggleZ (Drinkwater et al. 2010; Parkinson et al. 2012, 43 874 matches) and

(ix) zCOSMOS (Lilly et al. 2007, 2009, 2789 matches).

Refer to Beck et al. (2016) for additional details regarding the cross-match – the same data were used here, with the important distinction that photometric colour and error cuts were not applied.

This procedure enabled us to find 168 834 matches, which extended the total number of galaxies with spectroscopic redshifts to 2209 299, and also extended the colour–magnitude space coverage of the sample such that the parameter range of the SDSS photometric set was covered.

In order to build, from the new extended spectroscopic sample (E1), a subset that follows the same colour–magnitude distribution as the original SDSS-DR12 photometric set, we randomly selected 75 000 objects from the SDSS-DR12 photometric sample (S2)[3] and performed a nearest neighbour (NN) search in E1 (in colour/r-magnitude space).

For each object in S2, we search for its first NN to include into our new set, Happy D. To avoid duplicate entries, if the given NN was already included, we select the next closest NN (second, third, etc.) which was not already in Happy D.

Then, we similarly constructed two new subsets to represent the DR12 spectroscopic sample, Happy A and Happy B. The former will act as a training set/spectroscopic sample, while the latter will be a test set that has the same distribution of photometric properties as the training set. Thus, we randomly selected 2 × 75 000 objects from S1, and searched for their NN in E1 using the method outline above, again avoiding any duplicates within Happy sets.

Finally, to create an intermediate sample that is between the photometric error properties of S1 and S2, we decided to perform a photometric error cut. Our goal was to reproduce the same range of photometric errors as in S1, but with a distribution that resembles S2, being more weighted towards higher errors. Thus, the cut was chosen to be at the 98th percentile (to discard outliers) of the photometric error distribution of S1 for each observed feature. We randomly selected 150 000 objects from S2, searched for their NNs in E1 following the same procedure, and applied the error cut. This yielded the set Happy C, composed of ≈60 000 galaxies. We note that contrary to all other Happy set pairings, Happy C and Happy D were allowed to overlap to avoid excessively selecting from the less populated faint end of E1.

Cutting the photometric error range of Happy C to match that of Happy A and B had an important side effect: the magnitude and colour ranges (note: the range, not the shape of the distribution) were also essentially cut to the limits of Happy A and B. This shows that the effects of photometric error and magnitude/colour coverage are in fact very much intertwined, one cannot modify one without affecting the other. There are two feasible explanations for why the colour ranges shrink because of the error cut. First, this observation could indicate that the wider colour distributions in the photometric sample are mainly caused by the higher photometric errors smearing the distribution, not by containing physically different extreme galaxies that are missing from the spectroscopic sample. Secondly, it could be a consequence of galaxy types with extreme colours being significantly more likely to have high measurement errors, therefore these would be preferentially eliminated by the cut and also would not be presented in the spectroscopic sample. Fig. 2 shows the r-magnitude and colour PDFs for all samples in Happy, and the number of objects contained in each set is shown in Table 1.

Following the same reasoning as presented in the previous section, we built Happy A to act as a spectroscopic sample and, as such, should be used for training. Happy B is completely representative of Happy A and so mirrors the equivalent scenario of traditional photo-z validation exercises. Happy C illustrates a photometric sample that has been cut to conform to the training sample, with a larger proportion of objects having high photometric errors. Happy D serves as a complete photometric sample, with both a wider parameter coverage and higher measurement errors. Thus, Happy C and D must only be used for testing, representing increasing degrees of complexity and similarity with the real photometric situation.

3 PHOTOMETRIC REDSHIFT ESTIMATION

Aiming at quantifying the impact of the above discussed effects on the traditional photo-z estimation methods, we selected a few examples to illustrate how the colour–magnitude coverage and its correlations with photometric redshift accuracy can be quantified in future discussions on photo-z methods.

3.1 Empirical methods

In what follows, we introduce a selection of empirical photo-z estimation techniques that were chosen to represent this class of methods.

[3] To be precise, the objects were selected from a 2 million element random sub-sample of S2.

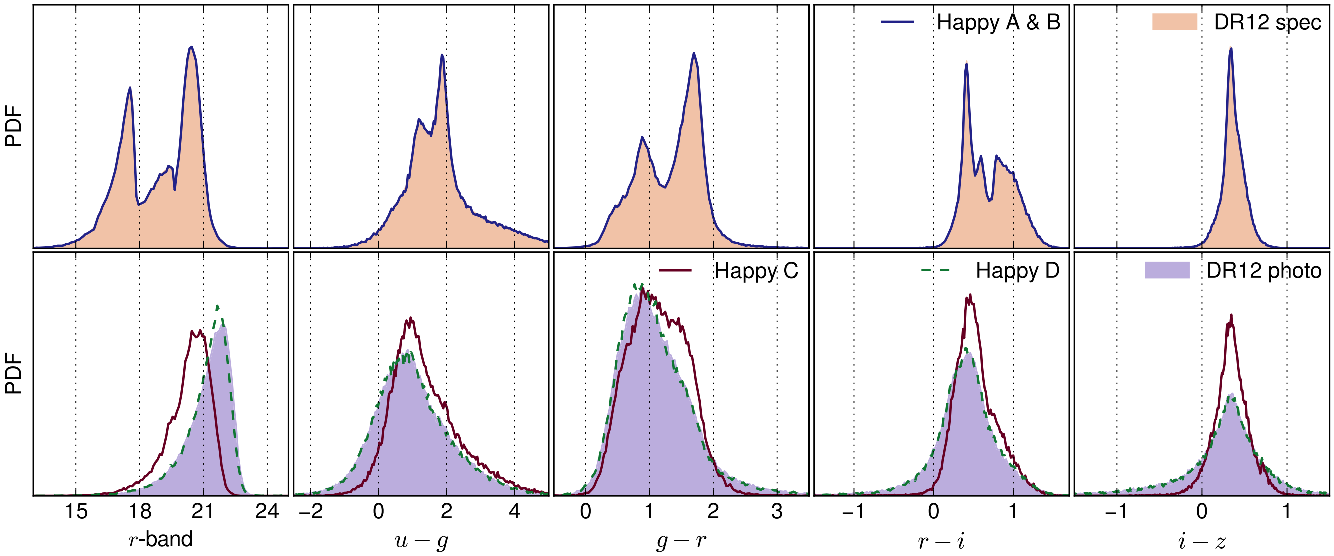

Figure 2. Distributions for magnitude in r band and four colours for the Happy catalogue compared with the full spectroscopic (red) and photometric (purple) distributions from SDSS-DR12.

3.1.1 *Artificial neural network (ANNZ)*

ANNZ (Collister & Lahav 2004) implements a particular type of artificial neural networks known as multilayer perceptron, which is formed by a set of layers, each one of them populated by a number of nodes. The first layer receives the observed magnitudes, or colours, the final layer returns the estimated photo-z values and the intermediate layers are considered hidden – since they can contain any number of nodes. All nodes in a given layer hold an activation function and are connected to all the nodes in adjacent layers, with each connection corresponding to a weight parameter $w_{i,j}$.[4]

Given a training set with measured magnitudes and spectroscopic redshifts, the network is trained by determining the set weight parameter values, **w**, which minimizes the cost function

$$E=\sum_{i=1}^{N}\left[z_p(\mathbf{w},\boldsymbol{m})-z_{\rm spec}\right]^2, \qquad (1)$$

where $\boldsymbol{m}$ is the set of observed magnitudes, z_p is the output given by the last layer and $z_{\rm spec}$ is the spectroscopic redshift (Abdalla et al. 2011).

The neural network used in this study is configured to have two intermediate layers, resulting in four layers in total. The first layer receives five input features: the r-band magnitude and four colours, normalized by their respective mean and standard deviation. The two intermediate layers contain each 10 nodes. The final layer outputs one node for the photo-z prediction, so that the whole network has a 5-10-10-1 structure.

In what follows, the network was trained in Teddy A and Happy A and subsequently applied to the other samples in each catalogue. We did not consider measurement errors in this work.

3.1.2 *Local linear regression (LLR)*

The LLR method finds the k-NNs of the test galaxy in colour space from a training set, and performs a hyperplane fit on these neighbours. This way, a functional form is fitted that also follows the local trends, and therefore can be quite flexible. The implementation used here is the same as in Beck et al. (2016), refer to that paper for more details. We note that while Beck et al. (2016) performed a template fitting step after the photo-z estimation itself, in this paper we do not utilize the extra physical information, therefore that step was omitted entirely. Thus, the method used here can be categorized as purely empirical.

The number of NNs used in Beck et al. (2016) is $k=100$, but since we are also testing extrapolation capabilities, that number was increased to $k=1000$ here. The increase does not noticeably affect estimation results when within the coverage of the training set, but taking a larger colour space region into account better determines the functional form, and significantly reduces scatter when extrapolating. Of course, there is a trade-off in computational performance when using more neighbours.

3.1.3 *Generalized additive model (GAM)*

Generalized linear models (GLMs), as its name suggests, assume – through a link function – a linear relationship between the response variable y and set of predictors $\boldsymbol{x}$. The distribution of y is a member of the exponential family and the link function is monotonic and differentiable. However, GLMs also allow different relationships between the mean and the variance. For example, in linear models (LMs) the response mean is independent of the variance and given simply by $E(y)=\boldsymbol{x}^T\boldsymbol{\beta}$. For Poisson models with log link $\log E(y)=\boldsymbol{x}^T\boldsymbol{\beta}$, the mean is equal to the variance. In this context, $\boldsymbol{\beta}$ is a vector of unobservable regression coefficients to be estimated from the data using maximum likelihood (ML) methods.

Let $\hat{\boldsymbol{\beta}}$ be the ML estimate of $\boldsymbol{\beta}$. A prediction at a new point $\boldsymbol{x}_0$ is obtained by $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ for LMs and by $\exp(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})$ for Poisson models. GLMs also enclose logistic and gamma regression, among many others. For the following discussion, we will only consider LMs but the same reasoning can be followed for any member of GLMs. For a comprehensive treatment on LMs, see Isobe et al. (1990), Dobson (2001), Kutner (2005), Christensen (2011) and Myers et al. (2012) and for GLMs see Nelder & Wedderburn (1972), Dobson (2001) and Myers et al. (2012). For GLM applications in astronomy,

[4] The indexes indicate the two nodes connected by this parameter.

the reader is refereed to Andreon & Hurn (2010), De Souza et al. (2015a), De Souza et al. (2015b), Elliott et al. (2015) and De Souza et al. (2016).

It is understood that the exact functional relationship between $E(y)$ and $\boldsymbol{x}$ is unknown. In other words, $E(y) = f(\boldsymbol{x})$ for unknown *f*. LMs assume that *f* can be approximated by $\boldsymbol{x}^T\boldsymbol{\beta}$. Despite its simplicity, LMs have been successfully applied in many areas. However, it has been found that the linearity assumption can be restrictive and too simple to account for non-trivial relationships in the data. Non-parametric regression relaxes this assumption and allows us to estimate *f* directly without imposing any specific functional formula. In fact, $f(\boldsymbol{x}) = \sum_{k=1}^{\infty} \alpha_k \phi_k(\boldsymbol{x})$ where ϕ_k's are known basis functions. Restricting the upper bound of *k* to a reasonable number *K*, $f(\boldsymbol{x}) \approx \sum_{k=1}^{K} \alpha_k \phi_k(\boldsymbol{x})$. Then, one can estimate α's as usual using parametric regression methods. Examples of basis functions include cubic spline basis, B-splines, Haar wavelet basis functions and radial basis functions.

This set-up does not impose any assumption on *f*. Even with a modest number of predictors, this attractive property requires estimating a huge number of parameters. Thus, *f* cannot be estimated properly due to the curse of dimensionality. To tackle this problem, Hastie & Tibshirani (1990) suggests to fit GAMs.

GAMs assume that $f(\boldsymbol{x}) = f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$. If each f_j demands *m* components then the total number of parameters is $p \times m$ which is reasonable even for moderate-sized studies. Penalized least squares (e.g. Ruppert, Wand & Carroll 2003) is a powerful approach to fit GAMs. Traditionally, GAMs are fitted using backfitting algorithm (Hastie & Tibshirani 1990). If each f_j estimated by $\hat{f}_j$ then $f(\boldsymbol{x})$ can be estimated by $\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{p} \hat{f}_j(x_j)$. A prediction at $\boldsymbol{x}_0$ is obtained by $\hat{f}(\boldsymbol{x}_0) = \sum_{j=1}^{p} \hat{f}_j(x_{0j})$. Additional details on fitting models and conducting inferential statistics using GAMS can be found in Hastie & Tibshirani (1990), Ruppert et al. (2003) and Wood (2006). We should note that the GAM methodology herein developed is a more general extension of the COSMOPHOTOZ (Elliott et al. 2015) package, who first introduced the use of GLMs for redshift estimation.

3.1.4 *Random forest*

Random forests depict ensembles of individual classification or regression trees, which are fit given the available training data. Each tree is built from top to bottom, where the root corresponds to all considered training instances and the leaves to subsets of instances. The internal nodes of each tree are usually split recursively (into two) based on certain splitting criteria. The overall recursive process stops as soon as some stopping criterion is fulfilled per leaf node. In case of fully grown trees, each leaf corresponds to single pattern or to a group of 'pure' patterns (i.e. a group of patterns having all the same label).

The original way of constructing random forests is to consider, for each individual tree, a subset of the training patterns, called *bootstrap sample* (Breiman 2001). These bootstrap samples are drawn uniformly at random (with replacement) to generate slightly different training sets and, hence, slightly different individual trees. Another way is to consider slightly different splits at each internal node by, e.g. considering different feature dimensions or random splitting thresholds. The overall performance of such a forest of trees is usually much better than the one of the individual trees (due to the variance reduction that stems from combining the predictions made by the trees).

The splitting processes taking place at the internal nodes is based on measuring the gain in 'purity', which can be, in turn, measured via different scores. Typical ones, measuring how impure a set of patterns associated with a node is, are the *mean squared error* (MSE) for regression problems or the *Gini index* for classification problems. We refer to Breiman (2001) for a detailed description.

In this work, we consider random forest regressors as we are interested in real-valued redshift estimates. While various parameters can be set for random forests, the performances of the induced models are often very similar among each other as long as reasonable parameter assignments are chosen. In the remainder of this work, we consider the following set-up: for all random forest models, 500 individual fully grown trees are fitted given bootstrap samples, where $\sqrt{d}$ features are tested per internal node split using the MSE as impurity measure.

3.2 Template fitting methods

In this section, we continue on to outlining the details of the spectral template fitting methods that were analysed in this paper.

3.2.1 *Bayesian photometric redshifts* (BPZ)

BPZ[5] applies Bayesian inference to the problem of photometric redshift estimation (Benítez 2000). In this context, the probability of a galaxy with measured colour and magnitudes, $\{C, \boldsymbol{m}\}$, to have a redshift *z* can be described as

$$p(z|C, \boldsymbol{m}) \propto p(z|\boldsymbol{m})p(C|z), \tag{2}$$

where $p(C|z)$ is the probability of the observed colours given a galaxy at redshift *z* (likelihood) and $p(z|\boldsymbol{m})$ is the expected redshift distribution for galaxies with measured magnitudes $\boldsymbol{m}$ (prior). This description assumes that the likelihood depends only on the measured magnitudes and morphological galaxy type (Benítez 2000). The feature that differentiates this approach from the others which came before it is the introduction of the prior. It improves over the simple template fitting χ^2 minimization by allowing the introduction of information about the galaxy morphological type and helps avoiding unrealistic redshift estimations by using simple assumptions as the range of redshift expected for a particular survey.

In this work, we present results using the default set of eight spectral energy distributions (SEDs) based on Coleman, Wu & Weedman (1980) and Kinney et al. (1996) and two extra interpolated ones between each pair (default option). The filters zero-points were calibrated using samples A.

3.2.2 EAZY

Easy and Accurate Redshifts from Yale[6] (EAZY) minimizes the dependence on spectroscopically measured spectra by relying on synthetic SEDs from semi-analytical models. This set, despite not enclosing all possible galaxy types and dust models, provides further completeness to UV and NIR wavelengths than SEDs built from spectroscopic observations (Brammer et al. 2008).

The default implementation, used in this work, constructs a minimum representative template set of five spectra derived from the application of a *non-negative matrix factorization* (NMF – Blanton & Roweis 2007) algorithm to the set of 485 synthetic spectra provided by Bruzual & Charlot (2003). NMF can be considered a kind of 'principal component' derivation, with the additional constraint

[5] http://www.stsci.edu/~dcoe/BPZ/

[6] http://www.astro.yale.edu/eazy/

that the linear combination coefficients need to be non-negative. Results are thus more interpretable than the ones derived from a standard principal component analysis (Ishida & de Souza 2011, 2013; Jolliffe 2013; De Souza et al. 2014).

Although the templates used in this method do provide a larger wavelength coverage, it is important to emphasize that the semi-analytical models themselves are constructed from detailed visual and NIR observations of nearby objects. Thus, the accuracy of these models in predicting the spectral behaviour of high-redshift galaxies is also limited. As the authors pointed out themselves, the discrepancy in UV and NIR fluxes between observed spectra and the one chosen by EAZY is larger in the rest-frame UV and NIR wavelengths. Consequently, its results will also be subjected to the lack of representativeness discussed above.

3.2.3 PHOTO-Z-SQL

PHOTO-Z-SQL[7] is a recent Bayesian template fitting photo-z implementation in C# that can be integrated into a data base running Microsoft SQL Server (Beck et al. 2017). The code is fairly flexible in the choice of templates and priors and thus can easily adopt successful approaches found in the literature. Moreover, it features an iterative photometric zero-point calibration to optionally take into account a spectroscopic training set.

We used the stand-alone version of the code, searching for the maximum probability photo-z value using Bayesian estimation. We present results for two configurations, both computed on a redshift grid with a linear step size of 0.01 between $z = 0$ and $z = 1$.

The first configuration, which we refer to as SQL BPZ, uses the BPZ spectral template set of Coe et al. (2006), and the prior of Benítez (2000) that had been derived from Hubble Deep Field North (HDF-N) data. It also utilizes an empirical filter zero-point calibration based on the training sets, and adds a separate photometric error term of 0.02 mag to account for template mismatch. The two major differences between this and the earlier BPZ approach (Section 3.2.1) are the differing calibration, and the larger number of interpolated templates (10 between each pair).

The second configuration – denoted by SQL LP – uses the LE PHARE (LP) spectral templates of Ilbert et al. (2009) in conjunction with a flat prior. In this case, zero-point calibration was not used, and the extra error term was chosen to be only 0.01 mag due to the larger and more detailed set of templates (641 for SQL LP, as opposed to 71 for SQL BPZ).

These configurations were selected to optimize results on the Teddy B and Happy B samples, which correspond to the usual case of only having validation information based on a spectroscopic set. In those samples, the SQL LP configuration did not benefit from either the calibration or applying the HDF-N prior, while the SQL BPZ case was improved by both.

For reference, the SQL BPZ and SQL LP configurations correspond to the notation *BPZ HDF Err ZP* and *LP Flat Err*, respectively, in Beck et al. (2017).

4 RESULTS

4.1 Diagnostics of estimators

Following earlier works that compare photo-z methods (Hildebrandt et al. 2010; Dahlen et al. 2013), we selected four summary statistics to quantify the photo-z estimation quality of the various methods tested here. We consider the normalized redshift error, $\Delta z_{\rm norm} = (z_{\rm spec} - z_{\rm photo})/(1 + z_{\rm spec})$, and from the distribution of $\Delta z_{\rm norm}$, we compute its mean (which is also the average bias), standard deviation (std), median absolute deviation (MAD) and outlier rate. Outliers are defined by $|\Delta z_{\rm norm}| > 0.15$. The median absolute deviation MAD = median($|\Delta z_{\rm norm}|$) is computed with outliers included; and the mean $\overline{\Delta z_{\rm norm}}$ and standard deviation $\sigma(\Delta z_{\rm norm})$ are computed with the outliers removed from the samples.

4.2 Results from Teddy

In order to quantify the impact of the lack of r-magnitude/colour coverage between spectroscopic and photometric samples, we applied the photo-z methods described above to the Teddy catalogue. Methods were trained/calibrated on Teddy A and tested on Teddy B, C and D to represent increasing levels of mismatch.

4.2.1 Empirical methods

Fig. 3 shows the photo-z estimations versus their spec-z values. Each row represents one of the four machine-learning methods described in Section 3.1. Teddy B, represented on the left-hand panels of Fig. 3, yields very satisfactory results in general. This observation is not surprising since this testing set shares, by construction, the same feature-space properties of the training set, Teddy A. The diagnostics of results, given by the left-hand panel of Table 2, show that the absolute value of the normalized bias does not exceed 7×10^{-4} and the outlier rate 0.2 per cent for all four methods. We also see that, due to the colour cut, the spec-z of most galaxies of set B is between 0.15 and 0.6.

Results for Teddy C are shown on the middle panels of Fig. 3. While set C shares the same colour coverage as set A, the distribution differs. This disparity of colour distribution has a minor impact on the photo-z scatter. As shown by Table 2, the std and the outlier rate of Teddy C is slightly higher than that of Teddy B, whereas the mean and the MAD are not affected by a significant change. Readers may notice a horizontal feature privileging a prediction around $z_{\rm photo} \approx 0.45$ for both sets B and C. This feature can be explained by the lack of objects in the $r - i$ distribution around 0.7. Fig. 1 shows that $r - i$ peaks at 0.6 and 0.8. By examining the colour of these objects, we discovered that photo-z predictions for most galaxies with $r - i > 0.7$ are located around $z_{\rm photo} \approx 0.46$ and predictions for those with $0.6 < r - i < 0.7$ are found around $z_{\rm photo} \approx 0.36$. Therefore, a local minimum in the galaxy population at $r - i \approx 0.7$ would yield a deficit of predictions around $z_{\rm photo} \approx 0.41$, which corresponds to the 'neck' we observe below the apparent horizontal feature in Fig. 3. We conclude that the $r - i$ distributions of sets B and C are responsible for this result.

Results for Teddy D are shown on the right-hand panels of Fig. 3 and the left-hand panel of Table 2. Compared to the two previous cases, the std, MAD and outlier rate are significantly larger for all methods. As expected, this confirms that if we do not account for the difference of colour coverage between spec-z and photo-z, the assumed error on photo-z will be underestimated.

Apart from this general outcome across empirical methods, two distinct behaviours have been observed on set D. The $z_{\rm spec}$–$z_{\rm photo}$ scatter from GAM and LLR stays close to the diagonal, while ANNZ and random forest show two horizontal features, resulting in wrong predictions for galaxies with a true redshift at $z < 0.2$ or $z > 0.45$. Essentially, in the latter case $z_{\rm photo}$ values are truncated at the end

[7] https://github.com/beckrob/Photo-z-SQL

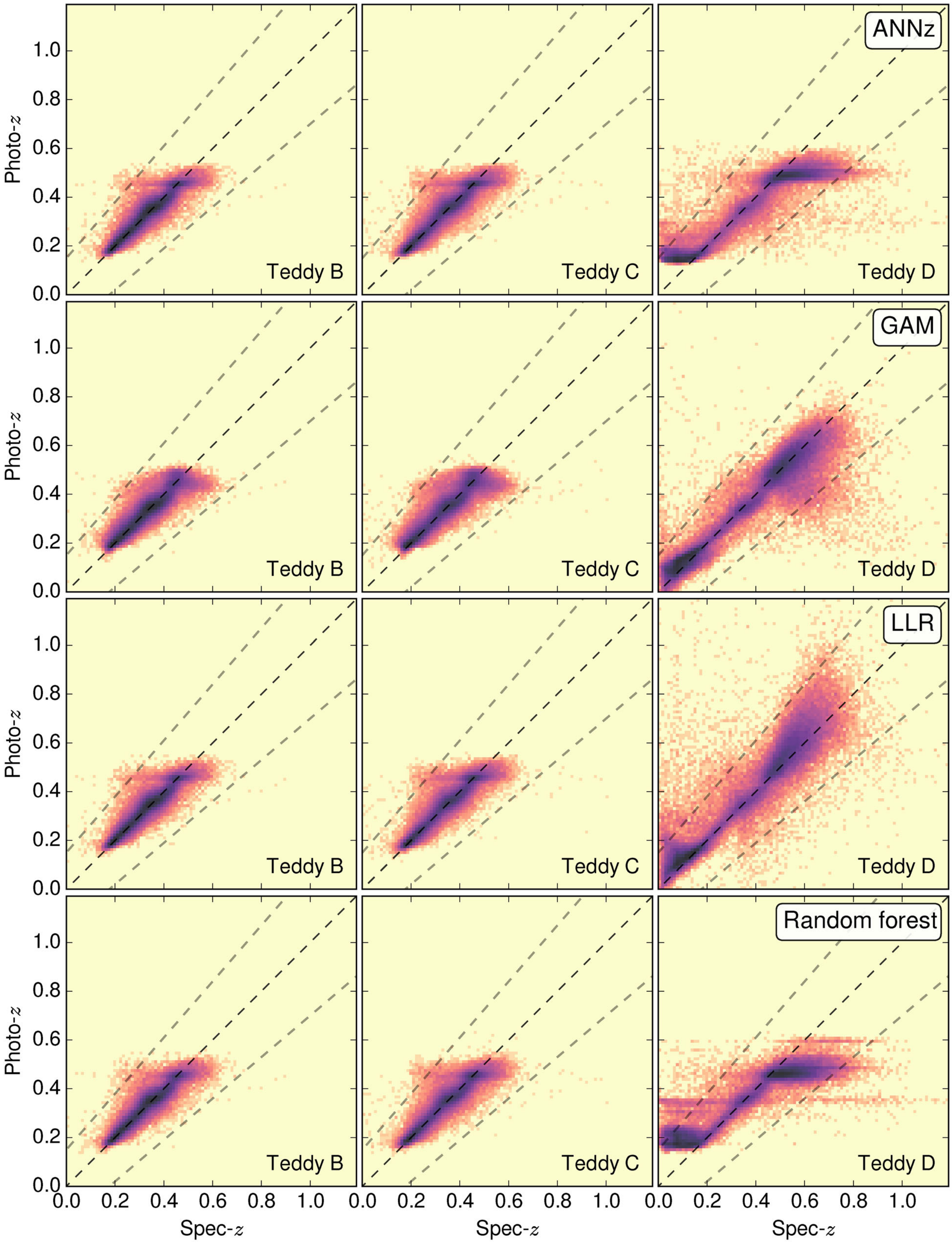

Figure 3. Results on three testing sets of the Teddy catalogue (columns) obtained from four empirical photo-z methods (lines). The colour gradient shows the logarithmic density. The dashed lines define the perfect prediction (centre) and the limits for being considered outliers. Numerical results are shown in Table 2 – left-hand panel.

Table 2. Results for the Teddy catalogue.

Method	Set	Diagnostics				Method	Set	Diagnostics			
		Mean ($\times 10^{-2}$)	Std ($\times 10^{-2}$)	MAD ($\times 10^{-2}$)	Outlier rate (per cent)			Mean ($\times 10^{-2}$)	Std ($\times 10^{-2}$)	MAD ($\times 10^{-2}$)	Outlier rate (per cent)
ANNZ	B	0.03	2.35	1.16	0.18	BPZ	B	3.51	3.07	3.63	0.49
	C	−0.01	2.45	1.15	0.26		C	3.48	3.30	3.69	0.58
	D	−0.08	5.67	3.61	3.09		D	2.61	4.73	3.66	3.60
GAM	B	0.05	2.62	1.34	0.11	EAZY	B	−2.99	3.82	3.05	2.71
	C	0.06	2.79	1.38	0.18		C	−3.71	4.07	3.57	4.03
	D	−0.06	3.93	2.23	2.28		D	−3.64	4.62	4.34	6.58
LLR	B	0.07	2.35	1.14	0.19	SQL BPZ	B	2.13	3.34	2.28	0.43
	C	0.05	2.44	1.14	0.28		C	1.68	3.40	2.00	0.64
	D	1.76	4.08	2.46	3.80		D	0.94	4.06	2.41	2.01
Random forest	B	0.03	2.38	1.18	0.17	SQL LP	B	−0.45	3.40	1.95	0.53
	C	−0.01	2.49	1.17	0.26		C	−0.7	3.61	2.14	0.70
	D	0.16	6.85	5.24	6.70		D	−0.48	4.19	2.74	3.29

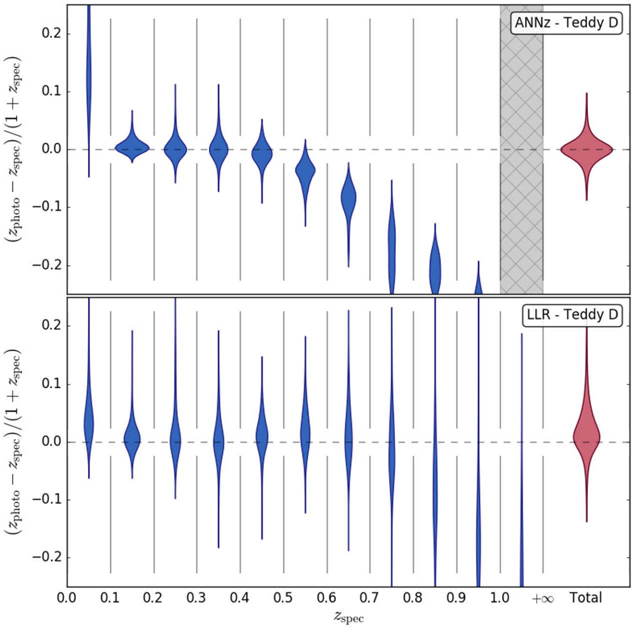

Figure 4. Violin plot of the normalized photo-z estimation error on Teddy set D, for the ANNZ and LLR methods. The redshift bins have a width of 0.1.

of the training set coverage. This effect has also been illustrated in Fig. 4. ANNZ clearly has strong $z_{\rm spec}$-dependent bias, while its overall bias is rather small due to positively and negatively biased regions cancelling out. In contrast, LLR has much smaller redshift-dependent bias, but the overall bias is higher. A similar comparison between random forest and GAM presents almost the same picture, with the exception that GAM also has low overall bias.

This result is a consequence of intrinsic differences between the methods we tested. GAM is a method that fits a rather general set of smooth functions to the training data. There is a 'global' relationship between covariates and the response variable, which is why we will refer to GAM and similar methods as 'global' methods. Of course, the function can be evaluated even when there is no coverage of the training set, which is why 'global' methods are expected to perform well when extrapolating, and indeed that is what we observe on Teddy D.

On the other hand, methods that strictly depend on the examples present in the training set and do not attempt to extrapolate its features, e.g. NNs and random forests, are not expected to be able to perform well beyond the boundaries of the training set – hence, the observed truncation on set D. For such models, the maximum redshift values that can be predicted are determined by the redshifts given in the training data. Neural networks could, in principle, fit an arbitrary 'global' functional formula, but in practice we observe the same behaviour with ANNZ as with the random forest, i.e. dependence dominated by colour–magnitude space neighbours. Therefore, we will refer to random forest, ANNZ and similar methods as 'local' methods.

LLR is an interesting hybrid of the previous two classes: it is based on NNs, thus it should be 'local', but it also fits a linear functional formula that can be used to extrapolate. Indeed, with enough neighbours ($k = 1000$), LLR does extrapolate reasonably well on Teddy D.

4.2.2 *Template fitting methods*

The photo-z estimation results on the catalogue Teddy for the four template fitting methods (introduced in Section 3.2) are shown in Fig. 5, with each row of scatterplots corresponding to a method. Each column represents a subsample of Teddy: in order, sets B, C and D. Teddy A was used as the calibration sample for methods where it was applicable. Numerical diagnostics are presented in the right-hand panel of Table 2.

The results do not change significantly between Teddy B and C for any of the four methods, only becoming slightly worse for the latter. The std and MAD values are all in the neighbourhood of 0.03, and the outlier fraction is ≈0.5 per cent (with the exception of EAZY, where 3–4 per cent). However, most of the bias values are relatively high, comparable to the scatter: ≈0.035 for BPZ and EAZY, and ≈0.02 for SQL BPZ. SQL LP has the lowest bias, ≈0.005. On these samples, the machine-learning methods have a clear edge in performance.

A visual inspection of Fig. 5 reveals that the high bias (and outlier rate) of EAZY is due to a sharp turn away from the diagonal above $z_{\rm spec} = 0.4$. In contrast, for the two BPZ template methods, the bias is caused by an upward shift of the entire population.

On the sample Teddy D, which is the case illustrating extrapolation outside the coverage of the training set, we can again observe a high bias around ≈0.03 for the BPZ and EAZY methods, with outlier rates of 3.6 per cent, 6.6 per cent and std rising to ≈0.046. However,

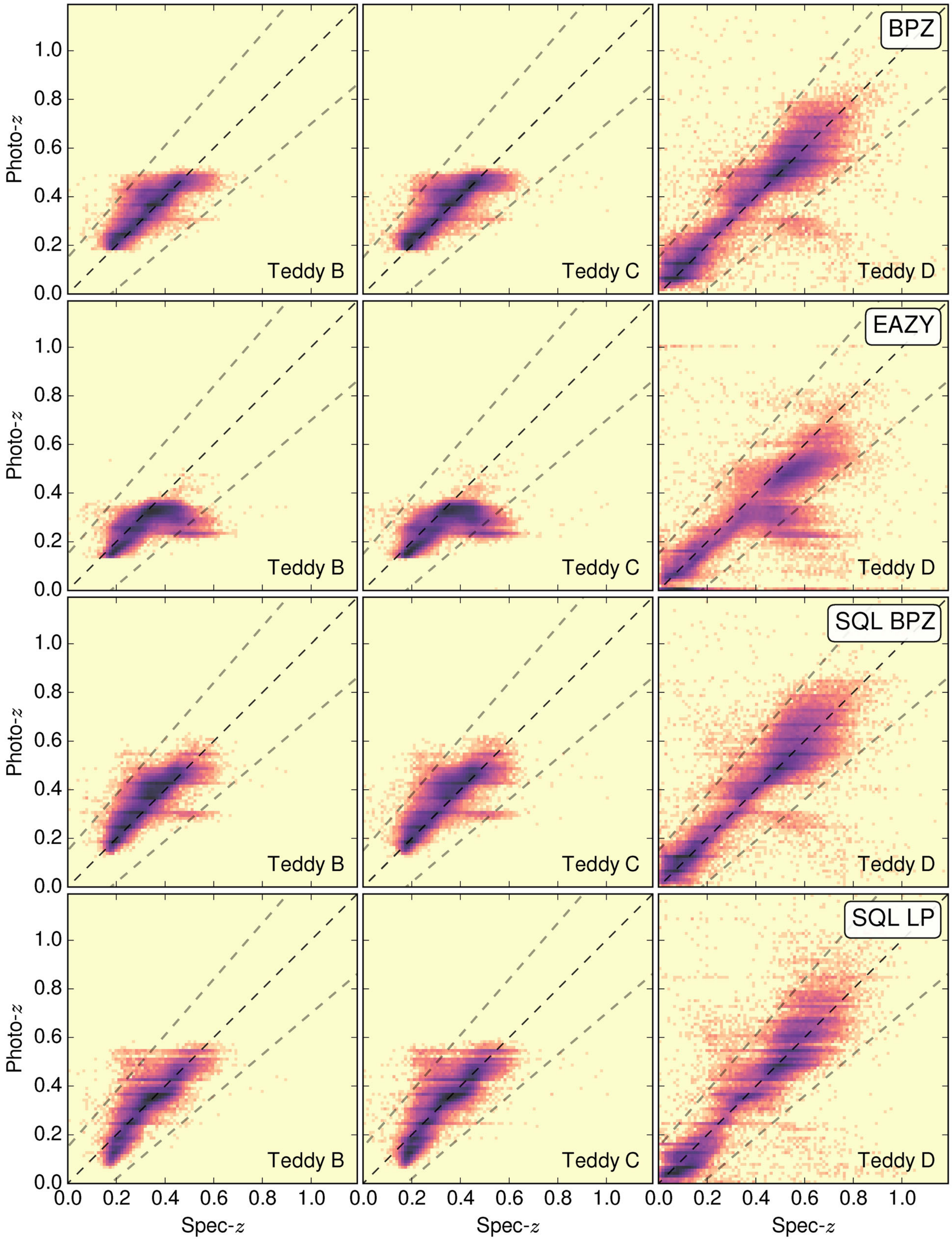

Figure 5. The photo-z estimation results for the four template fitting methods (lines) on the three testing subsamples of the Teddy catalogue (columns). The colour gradient shows the logarithmic density. The dashed lines define the perfect prediction (middle) and the limits for being considered as outliers. Numerical results for all four diagnostics are shown in Table 3 – right-hand panel.

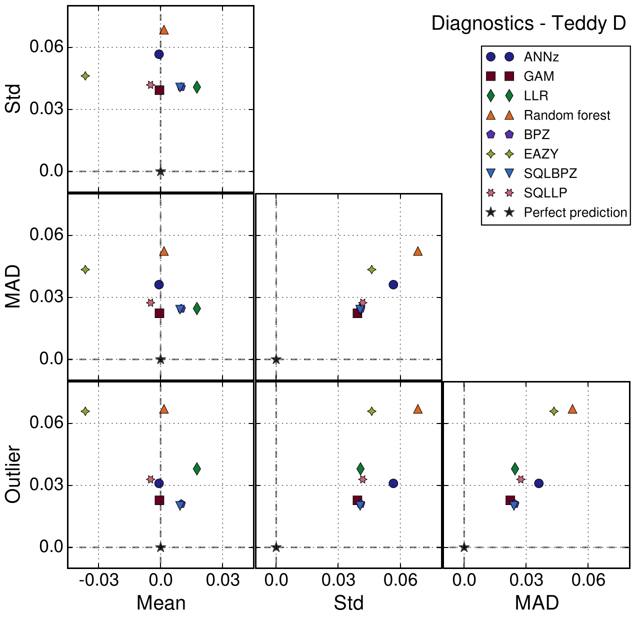

Figure 6. Comparison of numerical diagnostics (panels) for different photo-*z* methods (symbols) from sample Teddy D.

in the case of SQL BPZ and SQL LP, the overall bias remains below 0.01, with a scatter of ≈0.04 and 2–3 per cent outliers.

For the extrapolating case of Teddy D, which is the worst-case scenario in this catalogue, we compare the numerical diagnostics for both empirical and template fitting methods in Fig. 6. In this figure, each photo-*z* method is represented by a different symbol and each panel corresponds to a different diagnostic. The black star in the origin of each panel represents the best-case scenario, where a method would lie. Thus, the closer a symbol is from the black star the better the corresponding method performed according to a given diagnostic. In this visualization, it is possible to note that template fitting methods outperform 'local', non-extrapolating empirical methods on this sample, and are even comparable to 'global' methods as long as their overall bias can be managed e.g. with proper calibration. However, GAM still has a slight edge over the best-performing template fitting methods in this test, SQL LP and SQL BPZ.

4.3 Results from Happy

Although Teddy is a good starting point to probe the bias between photometric and spectroscopic samples, it is still simpler than the real scenario. It was intended to isolate the effect of gap in the feature-space coverage, but also it was built entirely from the SDSS spectroscopic sample, which means that all its objects fulfil the same spectroscopic data quality requirements. In what follows we shall relax this assumption and quantify the performance of photo-*z* estimators in a harder and more realistic scenario. We now present results for the Happy catalogue – specially built to account for differences between the photometric error distributions of the samples and their correlation with the lack of feature-space coverage.

4.3.1 *Machine-learning methods*

Fig. 7 shows the scatterplots of photo-*z* estimation results for the machine-learning methods described in Section 3.1 on different samples of the Happy catalogue. Numerical diagnostics are presented in the left-hand panel of Table 3.

All methods provide reasonably accurate redshifts on Happy B, where the estimated galaxies have the same distribution of magnitude, colour *and* photometric error as the spectroscopic training set (Happy A). Outlier rates were kept around 1 per cent for all tested photo-*z* codes while local methods (ANNZ and random forest) presented smaller biases, $\approx 5 \times 10^{-4}$, then global ones (LLR and GAM), $\approx 1 \times 10^{-3}$. GAM obtained the larger scatter $\approx 3.5 \times 10^{-2}$, a trend that was maintained for the other samples.

Interestingly, even with the same range of photometric errors, and within the coverage of other properties, on Happy C the scatter and proportion of outliers are significantly larger across the board due to the different error distribution (weighted towards higher errors). In this context, ANNZ obtained significantly smaller bias than the other methods, 0.16×10^{-2} – half of the number achieved by random forest, the second largest biased – and GAM presented the largest outlier rates, ≈7 per cent and scatter, ≈6.3. As expected, photo-*z* accuracy drops even more for all methods on Happy D, where many objects are outside the coverage of the training set in all respects. In this scenario, all methods produced outlier rates >10 per cent. For the GAM method, an unwanted feature shows up on Happy C, a linear broadening of the well-populated region between $z_{\text{spec}} \approx 0.2$–$0.6$ that is estimated to be at $z_{\text{photo}} \approx 0.3$. The feature broadens even further on the sample Happy D. This would suggest that the global fitting is more and more disrupted as the photometric errors increase, whereas for the other local methods such an effect is not observed. The numerical results also show that indeed GAM performs worse than the other methods when photometric errors get higher: the MAD value goes from being ≈0.005 worse than the other methods to being ≈0.012 worse, while the outlier rate goes from 0.3 per cent worse to 2 per cent worse.

There are two main takeaways from the results on the Happy catalogue. First, even if an error or colour cut is performed on the photometric sample to make it cover the same parameter range as the spectroscopic training set (this was done in Happy C), the results on a spectroscopic validation set (Happy B) will not be representative of results on such a cut photometric sample, due to the differing shape of the error distribution. Ultimately, to accurately determine the photo-*z* estimation accuracy of an object, its individual photometric error has to be taken into account along with the typical photo-*z* error of its NN galaxies (those with similar colour and magnitude). It follows that any attempt at dealing with the mismatch between spectroscopic and photometric samples must also include their photometric error distribution differences in the calculation of population diagnostics, independently of their feature-space coverage. Otherwise, even adaptations like calculating appropriate weights for the training sample will output too optimistic diagnostics (see Section 4.4). The Happy catalogue provides for the first time an environment where such new approaches can be directly tested.

Secondly, based on the methods tested here, it appears that global model fitting methods perform worse in the presence of large photometric errors than local empirical methods. Neural networks, while in essence fitting an arbitrary functional formula, behave similarly to NN methods in this regard.

4.3.2 *Template fitting methods*

The template fitting based photo-*z* estimation scatterplots for the Happy catalogue are presented in Fig. 8, using the same layout as

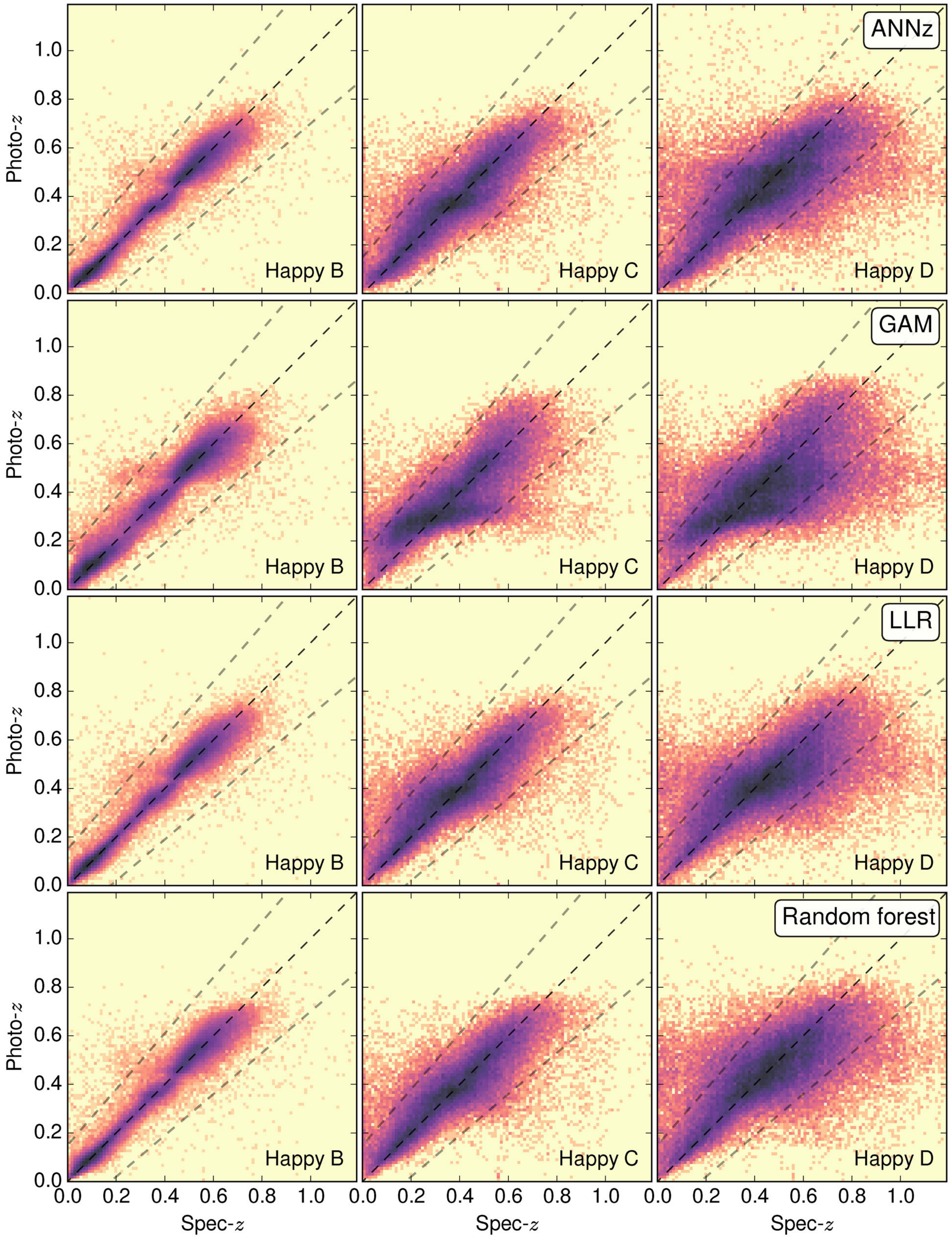

Figure 7. Results from applying empirical photo-z algorithms (lines) to the three testing samples of the Happy catalogue (columns). The colour gradient shows the logarithmic density. The dashed lines define the perfect prediction (middle) and the limits for being considered as outliers. Numerical results for all four diagnostics are shown in Table 3 – left-hand panel.

Table 3. Results for the Happy catalogue.

Method	Set	Diagnostics				Method	Set	Diagnostics			
		Mean	Std	MAD	Outlier rate			Mean	Std	MAD	Outlier rate
		$(\times 10^{-2})$	$(\times 10^{-2})$	$(\times 10^{-2})$	(per cent)			$(\times 10^{-2})$	$(\times 10^{-2})$	$(\times 10^{-2})$	(per cent)
ANNZ	B	0.04	2.87	1.49	0.99	BPZ	B	2.11	3.93	2.81	1.88
	C	0.16	5.41	3.60	5.59		C	0.21	5.81	4.20	7.97
	D	−0.52	6.53	5.44	14.01		D	−1.56	6.66	6.41	20.1
GAM	B	0.09	3.50	1.95	1.36	EAZY	B	−3.66	4.57	4.27	6.31
	C	0.86	6.34	4.84	7.37		C	−4.11	5.48	6.25	17.88
	D	−0.51	7.21	6.70	16.38		D	−4.23	6.19	9.17	31.72
LLR	B	0.13	2.81	1.39	1.11	SQL BPZ	B	1.79	4.12	2.75	1.80
	C	0.52	5.45	3.59	6.07		C	0.09	5.94	4.41	8.87
	D	−0.79	6.62	5.62	14.52		D	−1.82	6.77	6.80	21.25
Random forest	B	0.05	2.82	1.41	1.02	SQL LP	B	−0.47	4.15	2.68	3.20
	C	0.34	5.39	3.51	5.58		C	−0.51	5.90	4.61	14.18
	D	−0.28	6.51	5.36	14.2		D	−1.33	6.74	8.63	34.14

in previous such figures. Numerical diagnostics are shown in the right-hand panel of Table 3.

For the Happy B sample, all methods perform reasonably well, but with a notable overall positive bias of ≈0.02 for the two BPZ template cases, and a negative bias of ≈−0.035 for EAZY. The LP template case had the least bias, roughly −0.005. On Happy C, with the photometric errors increasing, the scatter also jumps, but the degraded performance is best illustrated by the outlier rate skyrocketing to 8–18 per cent. On Happy D this trend only continues, with outlier rates becoming unmanageable, between 20 and 34 per cent.

We note that the many extreme outliers of the SQL LP case are a result of overfitting the errors – the LP template set is rather varied (641 templates in this configuration), containing young, dusty starburst galaxies with different dust models, thus more extreme colour combinations can still be fitted. Using a prior, the number of extreme cases could have been greatly mitigated, but that would have increased bias on Happy B, which is the sample we chose to optimize for.

An interesting feature, a populated outlying region appears around $z_{\rm spec} = 0.5$ and $z_{\rm photo} = 0.3$ for most cases in Fig. 8, potentially indicating a systematic effect in the SDSS measurements that leads to erroneous template matches. It also mostly coincides with the elongated linear feature of the GAM method on Happy C (see Section 4.3.1).

A summary of the diagnostics for the most realistic case, depicted in Happy D, is shown in Fig. 9. The configuration of the plot is the same as shown in Fig. 6. The machine-learning methods all have similar performance, with GAM being slightly worse in terms of std, and the template fitting methods clearly lagging behind. However, we note that even the results of the best-performing method leave much to be desired.

4.4 Reshaping photometric feature-space distributions

The results presented above demonstrate how Teddy and Happy allow us to quantify the impact of spectroscopic coverage and photometric errors in photo-*z* methods, respectively. In this section, we show how they can also provide an environment to assess the efficiency of possible solutions. Consider the case of different probability distributions in the feature space as depicted in the C samples from both catalogues. In such cases, when distributions in the feature space between the spectroscopic and photometric samples are significantly different, the mismatch may be incorporated into the learning scheme. In the machine-learning community, such strategies are often classified as *domain adaptation* (DA) techniques (Quionero-Candela et al. 2009).

Importance weighting (Huang et al. 2007) represents one possible DA approach for the learning scenarios at hand. It states that it is possible to reweight the training examples in order to increase the importance of entries that are frequent in the test, but underrepresented in the training sample. This is achieved by assigning higher weights to such test instances. In a similar fashion, samples that are overrepresented in the training set are downweighted. Hence, such a reweighting strategy aims at reducing the shift between training and testing distribution in the feature space. In astronomy, an implementation of similar ideas was presented by Lima et al. (2008) and has been applied to large imaging surveys (e.g. Bonnett et al. 2016), while other forms of DA were also applied to star classification problems (e.g. Vilalta, Gupta & Macri 2013).

Huang et al. (2007) also propose a so-called *kernel mean matching* framework that aims at estimating corresponding weights via quadratic programming. This framework has been applied to the problem of supernova photometric classification leading to encouraging results (Pampana 2016). Albeit technically robust, the approach is limited by the amount of both training and test points it can handle. For the scenarios considered in this work, the data sets can easily consist of hundreds of thousands of objects – too large for standard quadratic programming solvers. For such cases, Kremer et al. (2015) have extended a NN-based technique that scales well for large samples, especially given low-dimensional search spaces. We used this machinery,[8] with the default configuration to calculate the weights for objects in samples A in all the three different scenarios with test samples B, C and D considering both our catalogues. The reader should be aware that the application of this method in sparse test samples, such as those in Happy, might lead to numerical problems. In order to ensure the convergence of the results we constructed larger test samples, following the same procedure described in Section 2.2, for the single purpose of assuring the convergence of the weight coefficients. The extended Happy C catalogue is also publicly available.

Once the weights were calculated, we incorporated them into the GAM and LLR methods, which are the ones allowing an easy implementation of these coefficients without the need of complex modifications in the original code. As expected, results considering sets B as test samples yield the same outcomes as presented in

[8] Code available at https://github.com/kremerj/nnratio.

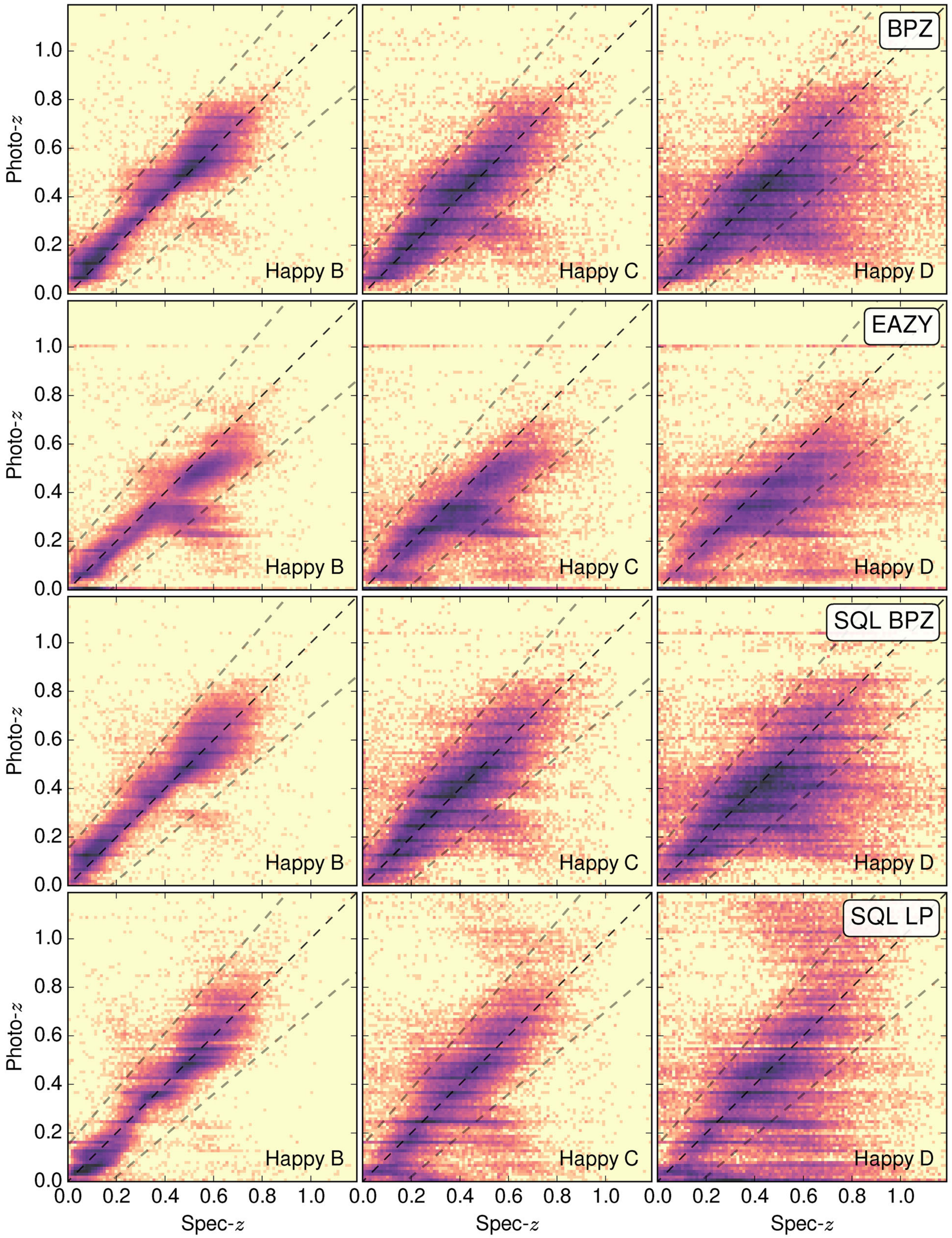

Figure 8. Template fitting photo-z results obtained from the four methods described in Section 3.2 (lines) for the three testing subsamples of the Happy catalogue (columns). The colour gradient shows the logarithmic density. The dashed lines define the perfect prediction (middle) and the limits for being considered as outliers. Numerical results for all four diagnostics are shown in Table 3 – right-hand panel.

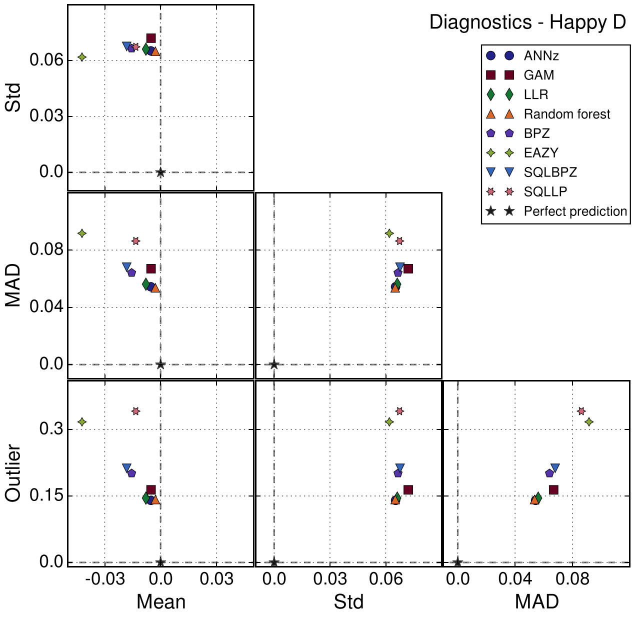

Figure 9. Comparison of numerical diagnostics obtained from applying different photo-z methods to Happy D.

the left-hand panels of Figs 3 and 7. Similarly, results from sets D were not that different from the original case. They were built to emphasize the gap in coverage between training and test sets – a situation where DA is not expected to have a large effect. As stated before, supervised learning algorithms learn by example and their results should not be extrapolated beyond the range covered by the training sample.

Samples C on the other hand, are a good testing ground, since in their case the spectroscopic sample provides at least a few examples in all regions of the parameter space occupied by the photometric sample. Using GAM in Teddy C we obtained a bias of 3 × 10^{-4} – half of the value achieved without considering the weights (see Table 2), while other diagnostics were not significantly changed. On the other hand, LLR did not show noticeable deviations between the weighted and unweighted results. This was also expected given the local characteristic of the method. Since LLR performs the regression exercise considering only a certain region of the parameter space, the weights calculated based on density estimates were not much different for a given set of neighbours.

Interesting qualitative results were encountered in Happy C. The weighting method was able to remove the horizontal trend mentioned in Section 4.3.1, resulting in a much more homogeneous distribution in $z_{\rm spec}$ × $z_{\rm photo}$ space (see Fig. 10), although numerical diagnostics were not significantly changed. The latter is a direct consequence of the different underlying error distributions between training and test samples, as noted in Section 4.3.1. The horizontal feature on the figure, which was associated with different underlying photometric feature distributions, was minimized, but in order to make a difference in the population diagnostics it is imperative to also take the error distributions into account.

5 DISCUSSION

In this paper, we present a comprehensive discussion of two main differences between spectroscopic and photometric samples which must be addressed by photo-z analyses: the lack of complete spectroscopic coverage in the colour/magnitude space and the presence of distinct photometric error distributions and ranges.

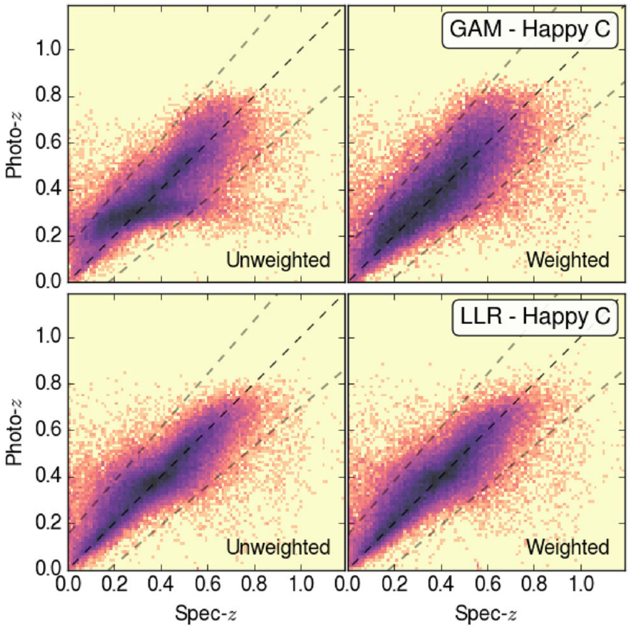

Figure 10. Comparison between results from LLR and GAM methods in the traditional approach (left) and after reweighting the spectroscopic sample (right) for set Happy C.

These problems are well known within the photo-z community but up to now there was no standard environment enabling a quantitative analysis of their influence on the final photo-z estimates. We developed such an environment through the construction of two catalogues: Teddy aims at isolating the effect of incomplete spectroscopic coverage and Happy probes the effect of photometric error distributions and ranges. Both catalogues are composed of four samples, with sample A being representative of a spectroscopic sample – and thus should be used for training purposes – and samples B, C and D depicting increasingly complex data situations. Teddy was built completely from the SDSS-DR12 spectroscopic sample while Happy uses photometry from SDSS and spectroscopy from many different sources (Section 2) – this allowed us to reproduce the statistical properties of the real SDSS-DR12 photometric sample while still possessing spectroscopic measurements for all objects in Happy samples. Both catalogues were submitted to the scrutiny of various machine learning and template fitting photo-z methods – both established and new, such as GAM.

We confirmed that most methods can adequately handle a difference in distribution shape as long as there is sufficient coverage in the training sample. In this set-up, template fitting methods perform worse, especially in terms of bias. However, when training set coverage is not available and thus extrapolation must be performed, local machine-learning methods fail, while global methods achieve reasonable results (see Section 4.2.1). Also, in the extrapolation case template fitting methods with the proper configuration can perform comparably to the better machine-learning methods. It should be noted that our extrapolation test sample (Teddy D) still has a relatively large number of training set points, and although their colour coverage is rather limited, we do not have to extrapolate very far in terms of redshift. More extensive extrapolation might prove to be difficult for all kinds of machine-learning methods.

We demonstrated that even with a photometric error cut in the photometric sample, a differing error distribution can lead to significantly worse results than what is observed on a traditional spectroscopic validation set. Increasing measurement error impairs performance for all methods, with template fitting methods being affected the most, global machine-learning methods behaving better, and finally local ML algorithms proving the most error resistant. Ultimately, the choice of photo-z method in applications will have to be optimized for the data at hand, specifically the colour–magnitude *and* error distributions of both the available training set and the target photometric sample. When choosing between various template fitting and machine-learning methods, a trade-off has to be made between extrapolation capability, performance within the coverage of the training set, and error resistance. With advances in instrumentation, the more accurate photometry of upcoming surveys (Stubbs et al. 2007; Li et al. 2016) and the relative lack of corresponding extensive spectroscopic samples might well tip the focus of such photo-z method optimization towards extrapolation, favouring template fitting or global machine-learning methods. However, ground-based observations do have strong physical limits regarding achievable photometric accuracy (Hartman et al. 2005; Stubbs et al. 2007), therefore the error resistance property cannot be ignored, either.

In order to illustrate how the catalogues presented here can be used to test alternative techniques aimed at dealing with the issues described above, we apply a density ratio weighting of the training sample which is designed to deal with differences in feature distributions (Section 4.4). Our results show that the method is able to deal with biases in photo-z determination but it does not numerically improve the overall diagnostics. This was expected, since such reweighting schemes can only reduce potential differences in the feature-space distribution, and cannot help with coverage or error distribution. In our context, the most realistic scenario with adequate coverage, depicted in Happy C, has a distinct photometric error distribution which was not taken into account. This is a clear example that simply reweighting the training sample is not enough to provide a realistic estimation of photo-z accuracy. Also, from a real-world perspective, in astronomical catalogues there could be shifts other than the ones mentioned so far, e.g. shifts in conditional probability distributions. Dealing with such additional shifts is a serious challenge yet to be overcome.

We note that most methods have quality cuts that could be used to filter out the worst photo-z failures. For example, template fitting methods can recognize bad or improbable template matches, and wide or multipeaked posterior redshift PDFs, while machine-learning methods can signal extrapolation, large NN bounding boxes, large deviations among neighbours, or sparse training set coverage. However, we saw no way to make the same quality cut across all the methods tested here, therefore in the spirit of fairness we chose not to do *no* quality cuts. Moreover, our goal is to emphasize that results grow worse through samples B to D, and that unless the issues addressed here are accounted for, we cannot explore the full potential of our data sets.

Finally, the fact that external, better quality spectroscopy was needed for Happy to properly mimic the real photometric sample characteristics while also possessing spectroscopic measurements is crucial – when one only has spectroscopic observations from the same instrument, it is impossible to test results on the poorer quality photometric observations. Thus, Teddy will never be Happy, or in other words, it is not possible to realistically address the photo-z challenges with only the uniform spectroscopic-quality measurements of a given survey, and without taking more than the feature-space coverage into account.

Teddy and Happy are publicly available.[9] They were built to be used as test benches in proving the fitness of current and future photo-z approaches. We strongly recommend using such difficult test cases as opposed to a simple cross-validation within a spectroscopic sample, since the latter was shown not to give representative results for the realistic use case. We also hope that providing such a user-friendly environment to test these issues will encourage not only astronomers, but researchers from other related areas to approach this problem.

ACKNOWLEDGEMENTS

This work is a product of the 3rd COIN Residence Program. We thank Zsolt Frei for encouraging the realization of this edition. The programme was held in Budapest, Hungary in August/2016 and supported by the Hungarian Academy of Sciences and Astrophysics group at Eötvos Lorand University.

We thank Laszlo Dobos, Madhura Killedar, Maria Luiza Linhares Dantas, Nikolay Kolesnikov and Pierre-Yves Lablanche for enlightening discussions during the programme. EEOI thanks Renuka Pampana and Ricardo Vilalta for valuable discussions on DA. We also thank T. Plüssmaci for helping us establish an unorthodox working environment.

RSS thanks FAPESP (projects 2016/13470-3, 2012/00800-4) for financial support. MVCD thanks his scholarship from FAPESP (processes 2014/18632-6 and 2016/05254-9). RB was supported through the New National Excellence Program of the Ministry of Human Capacities, Hungary, and the grant NKFIH OTKA NN 114560 of Hungary.

The IAA Cosmostatistics Initiative[10] (COIN) is a non-profit organization whose aim is to nourish the synergy between astrophysics, cosmology, statistics and machine-learning communities. We thank John Hammersley and the whole Overleaf team for providing collaborative tools with which this paper was written.[11] This work made use of the GitHub,[12] a web-based hosting service, the GIT version control software, and Slack,[13] a team collaboration platform.

In memoriam of Joseph M. Hilbe (1944 December 30 to 2017 March 12).

REFERENCES

Abdalla F. B., Amara A., Capak P., Cypriano E. S., Lahav O., Rhodes J., 2008, MNRAS, 387, 969
Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, MNRAS, 417, 1891
Alam S. et al., 2015, ApJS, 219, 12
Andreon S., Hurn M. A., 2010, MNRAS, 404, 1922
Antolini E., Heyl J. S., 2016, MNRAS, 462, 1085
Baldry I. K. et al., 2014, MNRAS, 441, 2440
Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2016, MNRAS, 460, 1371
Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2017, Astron. Comput., 19, 34
Benítez N., 2000, ApJ, 536, 571
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, 476
Bonnett C. et al., 2016, Phys. Rev. D, 94, 042005

[9] https://github.com/COINtoolbox/photoz_catalogues
[10] http://cointoolbox.github.io/
[11] https://www.overleaf.com/org/coin
[12] www.github.com
[13] https://slack.com

Boris N. V., Sodré L., Jr., Cypriano E. S., Santos W. A., de Oliveira C. M., West M., 2007, ApJ, 666, 747

Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503

Breiman L., 2001, Mach. Learn., 45, 5

Brescia M., Cavuoti S., Longo G., De Stefano V., 2014, A&A, 568, A126

Bruzual G., Charlot S., 2003, MNRAS, 344, 1000

Budavári T., 2009, ApJ, 695, 747

Budavári T., Szalay A. S., 2008, ApJ, 679, 301

Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511

Cavuoti S. et al., 2015, MNRAS, 452, 3100

Christensen R., 2011, Plane Answers to Complex Questions: The Theory of Linear Models. Springer Texts in Statistics. Springer-Verlag, New York

Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, AJ, 132, 926

Coil A. L. et al., 2011, ApJ, 741, 8

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

Colless M. et al., 2001, MNRAS, 328, 1039

Colless M. et al., 2003, preprint (astro-ph/0306581)

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

Cool R. J. et al., 2013, ApJ, 767, 118

Csabai I., Connolly A. J., Szalay A. S., Budavári T., 2000, AJ, 119, 69

Csabai I. et al., 2003, AJ, 125, 580

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

Davis M. et al., 2003, in Guhathakurta P., ed., Proc. SPIE Conf. Ser. Vol. 4834, Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II. SPIE, Bellingham, p. 161

De Souza R. S., Maio U., Biffi V., Ciardi B., 2014, MNRAS, 440, 240

De Souza R. S. et al., 2015a, Astron. Comput., 12, 21

De Souza R. S., Hilbe J. M., Buelens B., Riggs J. D., Cameron E., Ishida E. E. O., Chies-Santos A. L., Killedar M., 2015b, MNRAS, 453, 1928

De Souza R. S. et al., 2016, MNRAS, 461, 2115

Dobson A., 2001, An Introduction to Generalized Linear Models, 2nd edn. Taylor & Francis, USA

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

Driver S. P. et al., 2011, MNRAS, 413, 971

Elliott J., de Souza R. S., Krone-Martins A., Cameron E., Ishida E. E. O., Hilbe J., 2015, Astron. Comput., 10, 61

Garilli B. et al., 2008, A&A, 486, 683

Garilli B. et al., 2014, A&A, 562, A23

Guzzo L. et al., 2014, A&A, 566, A108

Hartman J. D., Stanek K. Z., Gaudi B. S., Holman M. J., McLeod B. A., 2005, AJ, 130, 2241

Hastie T., Tibshirani R., 1990, Generalized Additive Models. Chapman & Hall/CRC, London

Hildebrandt H., Wolf C., Benítez N., 2008, A&A, 480, 703

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

Hogan R., Fairbairn M., Seeburn N., 2015, MNRAS, 449, 2040

Huang J., Smola A. J., Gretton A., Borgwardt K. M., Schölkopf B., 2007, in Schölkopf B., Platt J., Hofmann T., eds, Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA, USA p. 601

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

Ilbert O. et al., 2009, ApJ, 690, 1236

Ishida E. E. O., de Souza R. S., 2011, A&A, 527, A49

Ishida E. E. O., de Souza R. S., 2013, MNRAS, 430, 509

Isobe T., Feigelson E. D., Akritas M. G., Babu G. J., 1990, ApJ, 364, 104

Jolliffe I., 2013, Principal Component Analysis. Springer-Verlag, New York

Jones D. H. et al., 2004, MNRAS, 355, 747

Jones D. H. et al., 2009, MNRAS, 399, 683

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

Kremer J., Gieseke F., Steenstrup Pedersen K., Igel C., 2015, Astron. Comput., 12, 67

Krone-Martins A., Ishida E. E. O., de Souza R. S., 2014, MNRAS, 443, L34

Kutner M., 2005, Applied Linear Statistical Models. McGraw-Hill, New York

Laureijs R. et al., 2011, preprint (arXiv:1110.3193)

Le Fèvre O. et al., 2004, A&A, 417, 839

Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258

Li T. S. et al., 2016, AJ, 151, 157

Lilly S. J. et al., 2007, ApJS, 172, 70

Lilly S. J. et al., 2009, ApJS, 184, 218

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

Lupton R. H., Gunn J. E., Szalay A. S., 1999, AJ, 118, 1406

MacDonald C. J., Bernstein G., 2010, PASP, 122, 485

Malavasi N., Pozzetti L., Cucciati O., Bardelli S., Cimatti A., 2016, A&A, 585, A116

Miles N., Freitas A., Serjeant S., 2007, in Ellis R., Allen T., Tuson A., eds, Applications and Innovations in Intelligent Systems XIV. Springer-Verlag, London, p. 75

Myers R., Montgomery D., Vining G., Robinson T., 2012, Generalized Linear Models: with Applications in Engineering and the Sciences. Wiley, New York

Natarajan A., Zentner A. R., Battaglia N., Trac H., 2014, Phys. Rev. D, 90, 063516

Nelder J. A., Wedderburn R. W. M., 1972, J. R. Stat. Soc. A, 135, 370

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

O'Mill A. L., Duplancic F., García Lambas D., Sodré L., Jr, 2011, MNRAS, 413, 1395

Pampana R., 2016, Master's thesis, Univ. Houston, USA

Parkinson D. et al., 2012, Phys. Rev. D, 86, 103518

Quionero-Candela J., Sugiyama M., Schwaighofer A., Lawrence N. D., 2009, Dataset Shift in Machine Learning. The MIT Press, Cambridge, MA, USA

Ruppert D., Wand M., Carroll R., 2003, Semiparametric Regression. Cambridge Univ. Press, Cambridge

Sánchez C. et al., 2014, MNRAS, 445, 1482

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

Stensbo-Smidt K., Gieseke F., Igel C., Zirm A., Steenstrup Pedersen K., 2017, MNRAS, 464, 2577

Stubbs C. W. et al., 2007, PASP, 119, 1163

Vilalta R., Gupta K. D., Macri L., 2013, Astron. Comput., 2, 46

Wadadekar Y., 2005, PASP, 117, 79

Wood S., 2006, Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC, Boca Raton, FL, USA

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