METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts — **Source link** ↗

Stefano Cavuoti, Valeria Amaro, Massimo Brescia, Civita Vellucci ...+2 more authors

Institutions: Astronomical Observatory of Capodimonte, California Institute of Technology

Related papers:

- Bayesian Photometric Redshift Estimation

- ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning

- ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks

- Photometric redshifts for Quasars in multi band Surveys

- EAZY: A Fast, Public Photometric Redshift Code

METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts

S. Cavuoti,[1][*] V. Amaro,[2] M. Brescia,[1] C. Vellucci,[2] C. Tortora[3] and G. Longo[2,4]

[1]*INAF – Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy*
[2]*Department of Physical Sciences, University of Napoli Federico II, via Cinthia 9, I-80126 Napoli, Italy*
[3]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, NL-9700 AV Groningen, the Netherlands*
[4]*California Institute of Technology – Center for Data-Driven Discovery, Pasadena, CA 91125, USA*

ABSTRACT

A variety of fundamental astrophysical science topics require the determination of very accurate photometric redshifts (photo-z). A wide plethora of methods have been developed, based either on template models fitting or on empirical explorations of the photometric parameter space. Machine-learning-based techniques are not explicitly dependent on the physical priors and able to produce accurate photo-z estimations within the photometric ranges derived from the spectroscopic training set. These estimates, however, are not easy to characterize in terms of a photo-z probability density function (PDF), due to the fact that the analytical relation mapping the photometric parameters on to the redshift space is virtually unknown. We present METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts), a method designed to provide a reliable PDF of the error distribution for empirical techniques. The method is implemented as a modular workflow, whose internal engine for photo-z estimation makes use of the MLPQNA neural network (Multi Layer Perceptron with Quasi Newton learning rule), with the possibility to easily replace the specific machine-learning model chosen to predict photo-z. We present a summary of results on SDSS-DR9 galaxy data, used also to perform a direct comparison with PDFs obtained by the LE PHARE spectral energy distribution template fitting. We show that METAPHOR is capable to estimate the precision and reliability of photometric redshifts obtained with three different self-adaptive techniques, i.e. MLPQNA, Random Forest and the standard K-Nearest Neighbors models.

Key words: techniques: photometric – galaxies: distances and redshifts – galaxies: photometry.

1 INTRODUCTION

Redshifts, by being directly correlated to the distance of the sources, lay at the very heart of almost all studies of the extragalactic universe and are used for a wide variety of tasks: to constrain the dark matter and dark energy contents of the Universe through weak gravitational lensing (Serjeant 2014), to understand the cosmic large-scale structure (Aragon Calvo et al. 2015) by identifying galaxy clusters and groups (Capozzi et al. 2009; Annunziatella et al. 2016), to map the galaxy colour–redshift relationships (Masters et al. 2015), and to classify astronomical sources (Brescia et al. 2012; Tortora et al. 2016), to quote just a few. Due to the multitude of ongoing multiband photometric galaxy surveys such as the Kilo-Degree Survey (de Jong et al. 2015), Dark Energy Survey (Annis 2013), Pan-STARRS (Kaiser 2004), and future facilities like the Large Synoptic

Survey Telescope (Ivezic 2009) and *Euclid* (Laureijs et al. 2014), we have entered an era requiring redshift estimates for billions of galaxies. Such a plethora of objects cannot, by any means, be observed spectroscopically, and redshift estimation through multiband photometry (hereafter, photometric redshift or photo-z) has become an indispensable tool in extragalactic astronomy, as the pace of galaxy detection in imaging surveys far outstrips the rate at which follow-up spectroscopy can be performed.

Many methods and techniques for photo-z estimation have been tested on a large variety of all-sky multiband surveys. These methods are broadly split into two large groups: physical template models fitting the spectral energy distributions (SEDs, cf. Arnouts et al. 1999; Bolzonella, Miralles & Pello 2000; Ilbert et al. 2006; Tanaka 2015) or the empirical exploration of the photometric parameter space (defined mainly by fluxes and derived colours). The latter infer the hidden correlation between the photometric data and the redshift using as templates a limited sample of objects with spectroscopic redshifts (cf. Connolly et al. 1995; Carrasco & Brunner 2013;

[*] E-mail: stefano.cavuoti@gmail.com
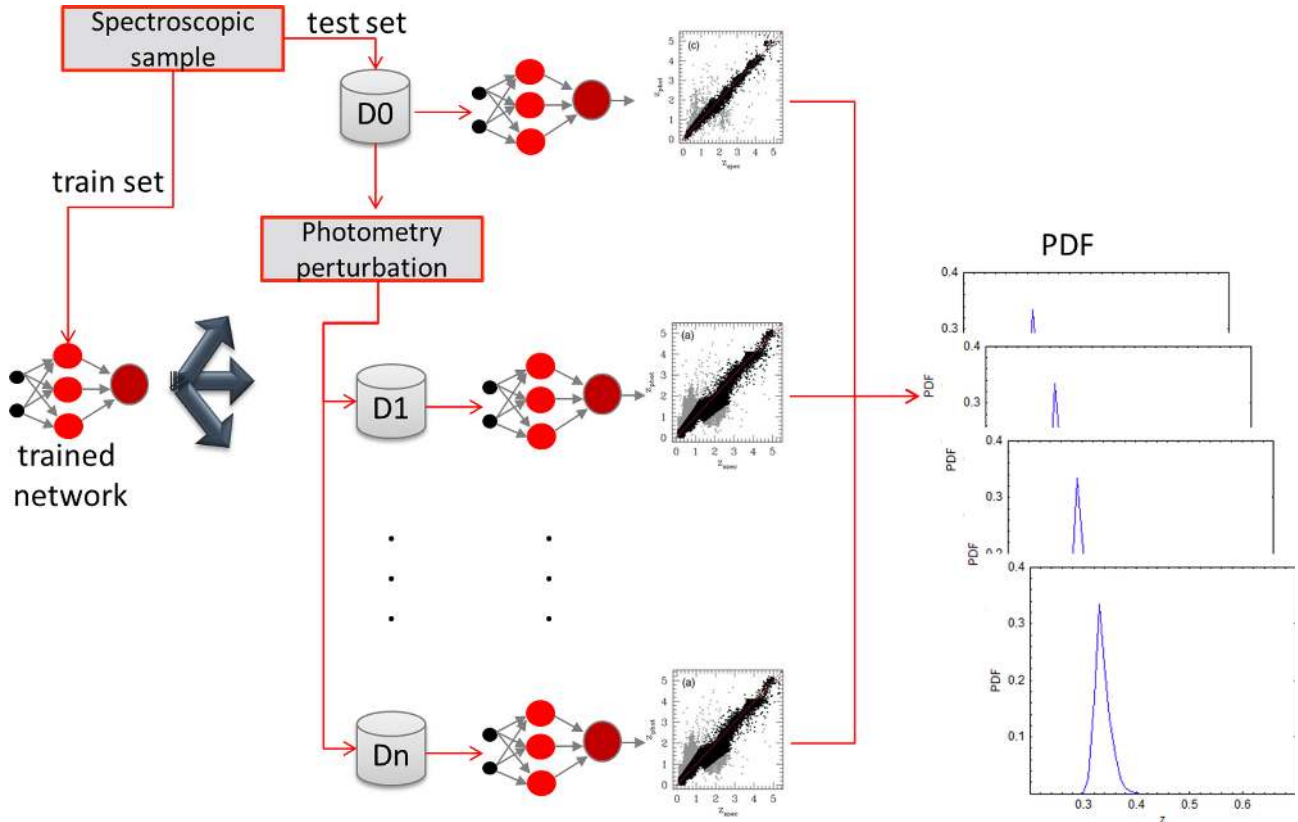
Figure 1. Basic scheme behind the idea of the METAPHOR method. The photo-*z* estimation method shown here is the MLPQNA neural network model, although it could be replaced by an arbitrary interpolation technique.

Brescia et al. 2014b; Cavuoti et al. 2015a). These two approaches, SED template fitting and empirical, present complementary pros and cons, since SED fitting methods are mostly physical prior dependent but able to predict photo-*z* in a wide photometric range, and provide in a rather natural way a likelihood-based estimation of the probability density function (PDF). On the opposite side, empirical, machine-learning (ML)-based techniques are largely independent of the physical priors, and they are able to produce more accurate photo-*z* within the photometric ranges imposed by the spectroscopic Knowledge Base (KB). An additional advantage of ML methods is that they can easily incorporate external information such as surface brightness, galaxy profiles, concentration indexes, angular sizes or environmental properties (cf. Sadeh, Abdalla & Lahav 2015). However, it is important to notice that empirical methods, on one hand, cannot provide accurate estimates outside of the photometric boundaries defined by the KB, and, secondly, they suffer from all selection effects and biases which are present in the KB. These dependencies have been extensively studied in the literature (cf. Ma & Bernstein 2014; Masters et al. 2015; Newman et al. 2015).

The complementarity of these two different approaches has led to several attempts to achieve a virtuous combination between the two approaches, with the purpose of improving the photo-*z* estimation quality (cf. Cavuoti et al. 2016, Fotopoulou et al., in preparation).

Over the years, particular attention has been focused on techniques that compute a full PDF for individual astronomical sources as well as for given galaxy samples. In fact, the single source PDF contains more information with respect to the simple estimate of a redshift together with its error, and it has been shown that PDFs can be effectively used to improve the accuracy of cosmological measurements (Mandelbaum et al. 2008).

From a rigorous statistical point of view, a PDF is an intrinsic property of a certain phenomenon, regardless of the measurement methods that allow us to quantify the phenomenon itself. The PDFs usually produced in the photo-*z* literature do not match this definition since they are just a way to parametrize the uncertainties on the photo-*z* solutions and to provide a robust estimate of their reliability. These pseudo-PDFs are in fact strictly dependent both on the measurement methods (and chosen internal parameters of the methods themselves) and on the underlying physical assumptions. In the absence of systematics, the factors affecting such reliability are photometric errors, internal errors of the methods and statistical biases.

In the framework of ML, a series of methods have been proposed to derive PDFs, not only for single sources, but also to estimate the cumulative or *stacked* PDF for a whole sample of galaxies. This *stacked* PDF describes the probability that a randomly sampled galaxy has a certain redshift. From a more general point of view, the idea is to find the mapping between the input parameters and an associated likelihood function spanning the entire redshift region, properly divided into classes (e.g. redshift bins). Such likelihood is expected to peak in the region where the true redshift actually is, while in the regions where the uncertainty is high, the same likelihood is expected to be flat. Different flavours of PDF determinations can be found in Bonnet (2013), Rau et al. (2015), (Sadeh et al. 2015), Carrasco & Brunner (2013, 2014a) and Carrasco & Brunner (2014b).

We present here a method which tries to account in a coherent manner for the uncertainties in the photometric data to find a perturbation law of the photometry, which could include not only a special procedure for a fitting of the errors on the attribute
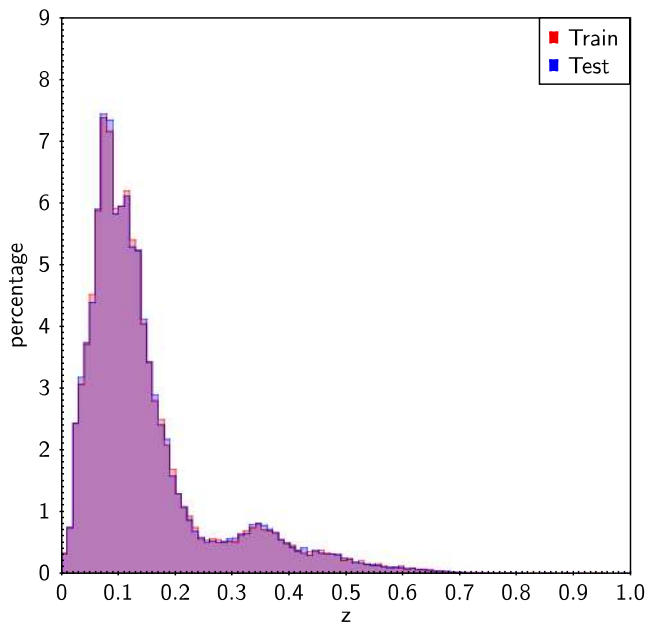
Figure 2. Distribution of SDSS DR9 spectroscopic redshifts used as a KB for the PDF experiments. In blue, the blind test set, and in red, the training set. The values are expressed in percentage, after normalizing the two distributions to the total number of objects.

Table 1. The *psfMag*-type magnitude cuts derived in each band during the KB definition.

Band	Lower limit	Upper limit
u	17.0	26.8
g	16.0	24.9
r	15.4	22.9
i	15.0	23.3
z	14.5	23.0

themselves, but also a level of randomness to be added to the information obtained from the errors. This in order to perform the perturbation of the attributes that have those errors, in a controlled, not biased by systematics, way. A proper error fitting, accounting for the attribute errors, allows us to constrain the perturbation of photometry on the biases of the measurements.

This paper is structured as it follows. In Section 2, we introduce the architecture of the designed processing flow. In Section 3, we describe the data, extracted from the Data Release 9 (DR9) of the Sloan Digital Sky Survey (SDSS), used to analyse the performance of the proposed workflow, while in Section 4, we briefly describe the photo-*z* estimation models used for the experiments. Then, the results are presented and discussed in Section 5. Finally, we draw the conclusions in Section 6.

2 THE METAPHOR PROCESSING FLOW

The complete pipeline processing flow has been named METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts) and includes functionalities which allow us to obtain a PDF from any photo-*z* prediction experiment done with interpolative methods. A layout of the METAPHOR pipeline is shown in Fig. 1 and is based on the following functional macro phases:

(i) *Data preprocessing*. It includes data preparation, photometric evaluation and error estimation of the multiband catalogue used as KB of the photo-*z* experiment. This phase includes also the photometric perturbation of the KB;

(ii) *Photo-z prediction*. It includes the training/test phase to be performed through the selected empirical method;

(iii) *PDF estimation*. This phase is related to the method designed and implemented to furnish a PDF for the produced photo-*z* and to evaluate the statistical performance.

We recall that a KB, in the context of photo-*z* prediction with supervised interpolative methods, is a data set composed of objects for which both photometric and spectroscopic information is available. At the user convenience, such set is randomly divided into several subsets, with arbitrary splitting percentages, in order to build, respectively, the training, validation and test data sets. For instance, a typical rule of thumb is to assign to each subset percentages equal to, respectively, 60 per cent, 20 per cent and 20 per cent of the original data. The training set is used during the learning phase; the validation set can be used to check the training correctness (mainly to avoid overfitting), but in our case, it is embedded into the training phase through the well-known technique of k-fold cross-validation (Geisser 1975); finally, a third set of data (test set) is used to evaluate the prediction performance and error estimate (for instance, the PDF of predicted photo-*z*). Obviously, train and test sets have a null intersection since the test data are never seen by the method at the training and validation stage. Very often this implies the necessity to merge data from different surveys (see e.g. Brescia et al. 2013), in particular by performing a reliable cross-match among different survey catalogues (Riccio et al. 2016) and further cleaning actions on the merged data, either by arbitrary criteria or by taking into account original prescriptions indicated by the survey providers.

METAPHOR is able to estimate a photo-*z* PDF for each single input object of the used data sample.

Given a spectroscopic sample, randomly shuffled and split into training and test sets, a photometry perturbation algorithm (see Section 2.2) and the selected photo-*z* estimation model (see Section 4.1), we proceed by perturbing the photometry of the given test set to obtain an arbitrary number *N* of test sets with a variable photometric noise contamination. Although the procedure foresees to apply perturbation also to the training set, after several tests, we decided to proceed by training the model with the not-perturbed training set and to submit the $N + 1$ test sets (i.e. *N* perturbed with noise sets and the original one) to the trained model, thus obtaining $N + 1$ estimates of photo-*z*. The reason lies in the fact that we found no degradation in performance, ignoring this step due to the large sample size of the training set.

With these $N + 1$ values, we perform a binning in photo-*z*, thus calculating for each one the probability that a given photo-*z* value belongs to each bin. The binning step is an arbitrary decision, to be made taking into account the specific requirements in terms of precision. We selected a step of 0.01 for the experiments described in Section 5.

The pseudo-algorithm, for a given photo-*z* binning step *B*, is the following:

(i) produce *N* photometric perturbations of the given test set, thus obtaining *N* additional test sets;

(ii) perform 1 training (or $N + 1$ train) and $N + 1$ tests;

(iii) derive and store the calculated $N + 1$ photo-*z* values;

(iv) calculate the number of photo-*z* for each bin ($C_{B,i} \in [Z_i, Z_{i+B}[$);
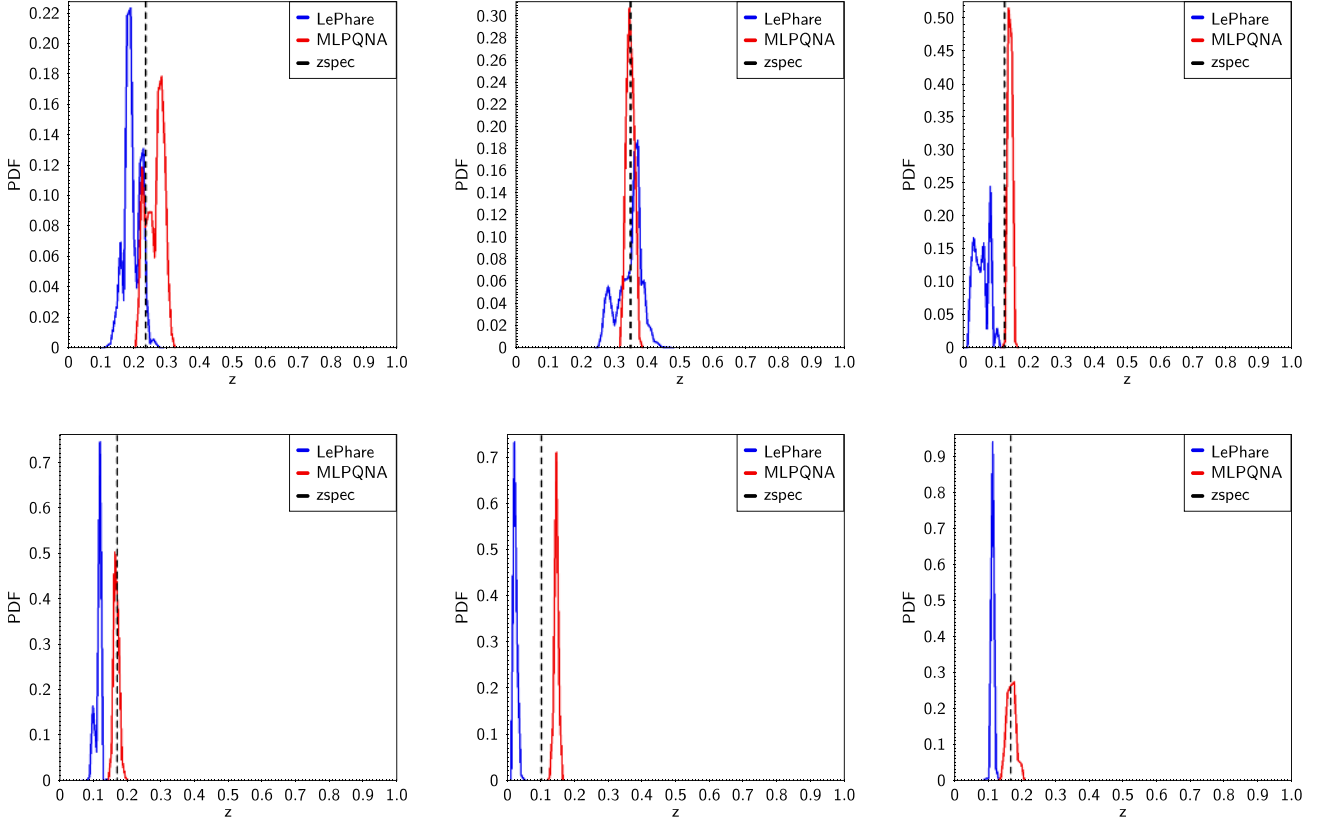
Figure 3. Some examples of photo-z PDF for single objects taken from the test set, obtained by MLPQNA (red) and LE PHARE (blue). The related spectroscopic redshift is indicated by the dotted vertical line. In some cases, the PDF peak appears lowered, due to an effect of a spread over a larger range of the peak (panel in the lower right-hand corner).

Table 2. Results for the various experiments obtained with MLQPNA. Column 1: identification of the experiment; column 2: type of error perturbation; column 3: threshold for the flat component; columns 4 − 10: $f_{0.05}$, $f_{0.15}$, $\langle \Delta z \rangle$, bias, σ, σ_{68}, NMAD (see Section 2.1); column 11: fraction of outliers outside the 0.15 range; column 12: skewness of the Δz; columns 13 − 16: fraction of objects having spectroscopic redshift falling within the peak of the PDF, within 1 bin from the peak, inside the remaining parts of the PDF and outside the PDF, respectively.

Id	Type	Threshold	$f_{0.05}$	$f_{0.15}$	$\langle \Delta z \rangle$	|bias|	σ	σ_{68}	NMAD	% of outliers	skewness	% peak	% one bin	% in PDF	% out PDF
1	flat	0.05	92.3	99.8	−2.0E−4	0.0	0.024	0.018	0.017	0.12	−0.12	21.3	32.4	26.9	19.351
2	flat	0.1	87.3	99.7	7.7E−4	0.0	0.019	0.019	0.018	0.11	−0.2	18.0	30.0	44.0	7.0
3	flat	0.2	73.8	98.4	6.5E−4	0.0	0.024	0.024	0.023	0.14	−0.35	14.0	24.0	59.0	2.0
4	flat	0.3	61.4	95.4	−0.0045	0.0	0.03	0.03	0.03	0.17	−0.37	12.0	21.0	66.0	2.0
5	flat	0.4	51.7	90.8	−0.014	0.0	0.039	0.039	0.038	0.31	−0.24	10.0	18.0	69.0	2.0
6	poly.	no	92.9	99.8	−0.0011	0.0	0.024	0.018	0.017	0.11	−0.16	22.1	30.3	13.5	34.13
7	indiv.	no	92.4	99.7	−0.001	0.0	0.024	0.018	0.017	0.12	−0.21	22.0	15.0	31.0	31.0
8	bimod.	0.05	91.8	99.8	−6.1E−4	0.0	0.024	0.019	0.017	0.11	−0.17	21.0	32.0	29.0	18.0
9	bimod.	0.1	87.1	99.6	5.4E−4	0.0	0.025	0.019	0.018	0.11	−0.23	18.0	31.0	44.0	7.0
10	bimod.	0.15	80.6	99.2	0.0012	0.0	0.026	0.021	0.02	0.12	−0.32	16.0	27.0	54.0	3.0
11	bimod.	0.2	73.8	98.4	5.8E−4	0.0	0.028	0.023	0.023	0.13	−0.39	14.0	11.0	73.0	2.0

(v) calculate, for each bin, the probability that the redshift belongs to the bin: $P(Z_i \leq \text{photo} - z < Z_{i+B}) = C_{B,i}/(N+1)$;

(vi) the resulting PDF thus being the set of all probabilities obtained at the previous step;

(vii) calculate the statistics.

2.1 Statistical estimators

The results of the photo-z calculations were evaluated using a standard set of statistical estimators for the quantity

$\Delta z = (z_{\text{spec}} - z_{\text{phot}})/(1 + z_{\text{spec}})$ on the objects in the blind test set, as listed in the following:

(i) bias: defined as the mean value of the residuals Δz;

(ii) σ: the standard deviation of the residuals;

(iii) σ_{68}: the radius of the region that includes 68 per cent of the residuals close to 0;

(iv) NMAD: the normalized median absolute deviation of the residuals, defined as NMAD $(\Delta z) = 1.48 \times \text{Median}(|\Delta z|)$;

(v) fraction of outliers with $|\Delta z| > 0.15$;

(vi) skewness: measurement of the asymmetry of the probability distribution of a real-valued random variable around its mean.

Table 3. Statistics of photo-z estimation performed by the MLPQNA, RF, KNN and LE PHARE models.

Estimator	MLPQNA	KNN	RF	LE PHARE
bias	0.0006	0.0029	0.0035	0.0009
σ	0.024	0.026	0.025	0.060
σ_{68}	0.018	0.020	0.019	0.035
NMAD	0.017	0.018	0.018	0.030
skewness	-0.17	0.330	0.015	-18.076
outliers > 0.15	0.11%	0.15%	0.15%	0.69%

Table 4. Statistics of the *stacked* PDF obtained by LE PHARE and by the three empirical models MLPQNA, KNN and RF through METAPHOR.

Estimator	MLPQNA	KNN	RF	LE PHARE
$f_{0.05}$	91.7%	92.0%	92.1%	71.2%
$f_{0.15}$	99.8%	99.8%	99.7%	99.1%
$\langle \Delta z \rangle$	-0.0006	-0.0018	-0.0016	0.0131

Furthermore, in order to evaluate the cumulative performance of the PDF, we computed the following three estimators on the *stacked* residuals of the PDFs:

(i) $f_{0.05}$: the percentage of residuals within ± 0.05;

(ii) $f_{0.15}$: the percentage of residuals within ± 0.15;

(iii) $\langle \Delta z \rangle$: the weighted average of all the residuals of the *stacked* PDFs.

2.2 The photometry perturbation law

From a theoretical point of view, the characterization of photo-z predicted by empirical methods should disentangle the photometric uncertainties from those intrinsic to the method itself.

The investigation is focused on the random perturbation of the photometry and the consequent estimation of its impact on the photo-z prediction, trying to quantify a PDF of the photo-z error distribution. The photometry perturbation procedure is based on the following expression, which is applied on the given j magnitudes of each band i as many times as the number of perturbations of the test set:

$$m_{ij} = m_{ij} + \alpha_i K_{ij} * \text{gaussRandom}_{(\mu=0, \sigma=1)}, \qquad (1)$$

where α_i is a multiplicative constant, defined by the user; $K_{ij}(x)$ is the weighting associated with each specific band used to weight the Gaussian noise contribution to magnitude values and gaussRandom$_{(\mu=0, \sigma=1)}$ is a random value from a normal distribution. In particular $\alpha_i K_{ij}$ represents the term used to generate the set of perturbed replicates of the blind test set.

We identified the following different cases for the weighting coefficient:

(i) constant weight (flat),
(ii) individual magnitude errors (indiv.),
(iii) polynomial fitting (poly.);
(iv) bimodal function (bimod.).

The constant weight is a floating number between 0 and 1 heuristically chosen. The second choice consists of weighting the Gaussian noise contribution using the individual magnitude error provided for each source. In the case of polynomial fitting, we perform a binning of photometric bands in which a polynomial fitting of the mean error values is used to reproduce the intrinsic trend of the distribution. The last option is a more sophisticated version of the

polynomial fitting coupled with a minimum value chosen through an heuristic series of tests. The binning of the parameter space, while allowing us to better control the photometric uncertainty in different distance regimes, also poses the additional problem of the correct choice of the bin size. This is in order to minimize the risks of information losses, varying between aliasing in the case of high-density binning and masking in the case of an undersampling of the parameter space.

The use of a different multiplicative constant for each band is also considered, in order to customize the photometric error trend on the basis of the specific band photometric quality. This is particularly suitable in the case of photometry obtained by merging different surveys. In the specific case of polynomial fitting, we define an expansion of the error trend aimed to overcome the risks related to mask or aliasing occurrence, due to a wrong choice of the bin size. The impact of such mechanism was analysed, reflecting the necessity to split the perturbation procedure into two steps: first, a preliminary statistical evaluation of the photometric error trend, in order to derive the coefficients of the polynomial noising function, and, secondly, the perturbation of the catalogue photometry. Furthermore, we find helpful the opportunity to extract also the standard deviation from each bin, in order to keep track of the expected error trend and eventually to derive a quality flag.

We performed a comparison amongst the four choices, in order to have a direct evaluation of the impact on the statistical performance between cumulative and individual error trends. In all cases, in order to avoid contamination due to bad-quality photometry, all objects having magnitude errors higher than 1 mag have been excluded from the analysis.

3 A REAL USE CASE: SDSS DATA

In order to evaluate the performance of the METAPHOR processing flow, we used a galaxy spectroscopic catalogue extracted from the SDSS DR9 (York et al. 2000).

The SDSS combines multiband photometry and fibre-based spectroscopy, providing all information required to constrain the fit of a function mapping the photometry into the spectroscopic redshift space. The KB for the presented experiment is composed of objects with specClass *galaxy* together with their photometry (*psfMag*-type magnitudes) and rejecting all objects with non-detected information in any of the five SDSS photometric bands (the original query could be found in Appendix A). From the original query, we extracted $\sim 50\,000$ objects to be used as a train set and $\sim 100\,000$ objects to be used for the blind test set. The redshift distributions for the train and test sets are shown in Fig. 2. The train and the test sets are drawn from the same population distribution in order to minimize the occurrences of biases/mismatch between train and test samples, which could induce degeneracies in the predicted photo-z.

The ranges in terms of magnitudes are reported in Table 1 and detailed in Brescia et al. (2014b), where we also produced a catalogue of photo-z for about 143 million galaxies (Brescia et al. 2014c).

4 THE PHOTO-Z ESTIMATION MODELS

The METAPHOR procedure can be, in principle, applied by making use of any arbitrary empirical photo-z estimation model. Moreover, as it was introduced above, the alternative category of photo-z estimation methods, based on SED template fitting, intrinsically provides PDFs. Therefore, we experimented the METAPHOR procedure with three different empirical methods, for instance, Multi Layer Perceptron with Quasi Newton learning rule (MLPQNA)
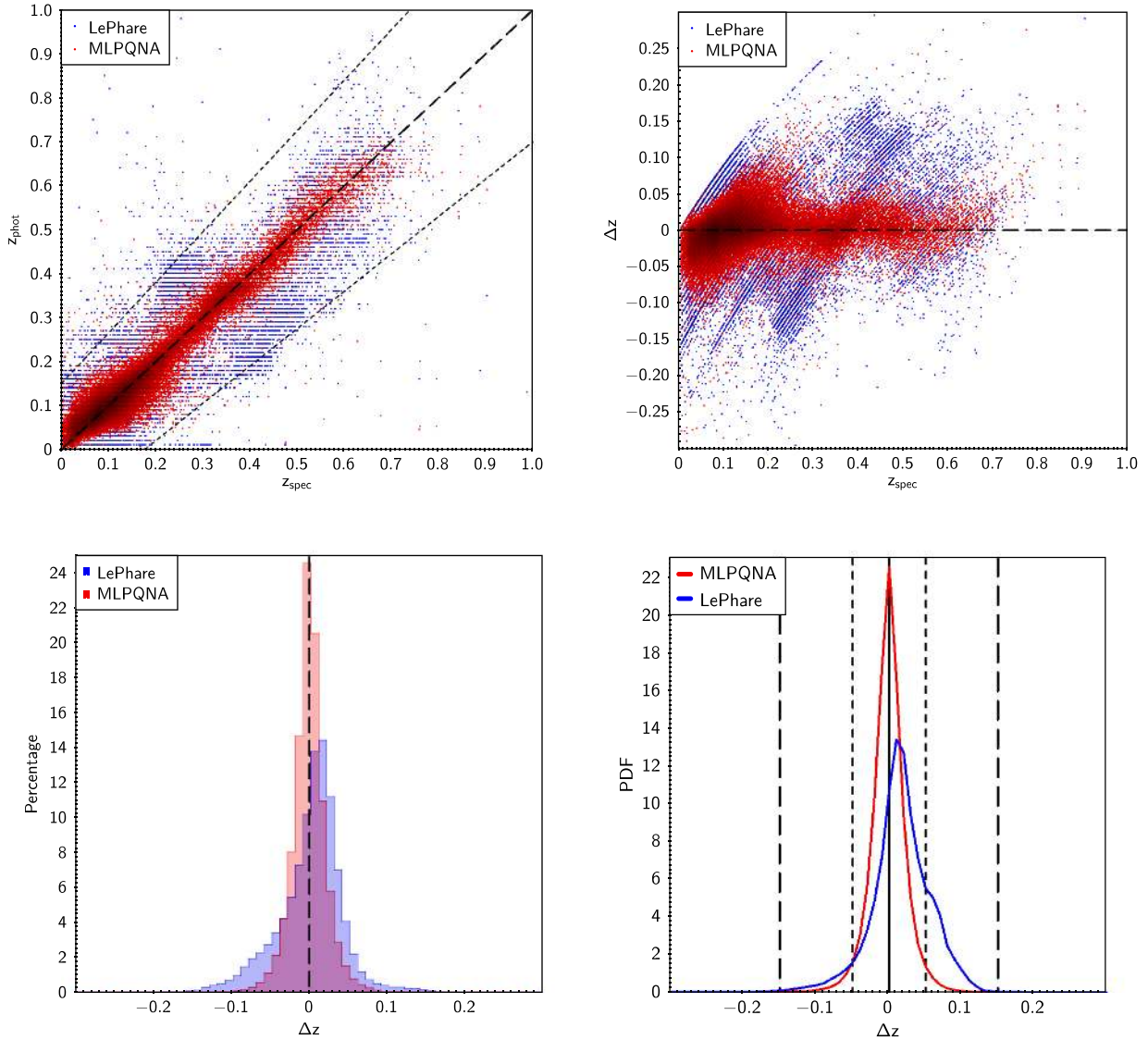
Figure 4. Comparison between MLPQNA (red) and LE PHARE (blue). Left-hand panel of upper row: scatter plot of photometric redshifts as a function of spectroscopic redshifts (z_{spec} versus z_{phot}); right-hand panel of upper row: scatter plot of residuals as a function of spectroscopic redshifts (z_{spec} versus Δz); left-hand panel of lower row: histograms of residuals (Δz); right-hand panel of lower row: *stacked* representation of residuals of the PDFs (the redshift binning is 0.01).

neural network (see Section 4.1), K-Nearest Neighbours (KNN; see Section 4.2) and Random Forest (RF; see Section 4.3), and compared their results with the LE PHARE SED template fitting technique (see Section 4.4).

In particular, the use of different empirical models has been carried out in order to verify the universality of the procedure with respect to different empirical models. It must also be pointed out that, aside from the selection of the RF model, the choice of the KNN method has been driven by its extreme simplicity with respect to the wide family of interpolation techniques. Therefore, by validating the METAPHOR procedure and PDF statistical performance with KNN, it would empirically demonstrate its general applicability to any other empirical method. All these methods are briefly described in the following sections.

According to the traditional supervised paradigm of ML, the KB used is split into different subsets, dedicated to training and test steps, respectively. The training set is used to learn the hidden relationship between photometric and spectroscopic information,

while the blind test set allows the evaluation and validation of the trained model on objects never submitted before to the network. In order to analyse the results on the test objects, a series of statistical estimators is then derived (see Section 2.1).

4.1 MLP with quasi-Newton algorithm

The MLPQNA model belongs to Newton's methods aimed at finding the stationary point of a function by means of an approximation of the Hessian of the training error through a cyclic gradient calculation. The implementation makes use of the known L-BFGS algorithm (Limited memory – Broyden Fletcher Goldfarb Shanno; Byrd, Nocedal & Schnabel 1994), originally designed to solve optimization problems characterized by a wide parameter space. The description details of the MLPQNA model have already been extensively discussed elsewhere (cf. Cavuoti, Brescia & Longo 2012b; Cavuoti et al. 2012a, 2014; Brescia et al. 2013, 2014a; Cavuoti et al. 2015b).
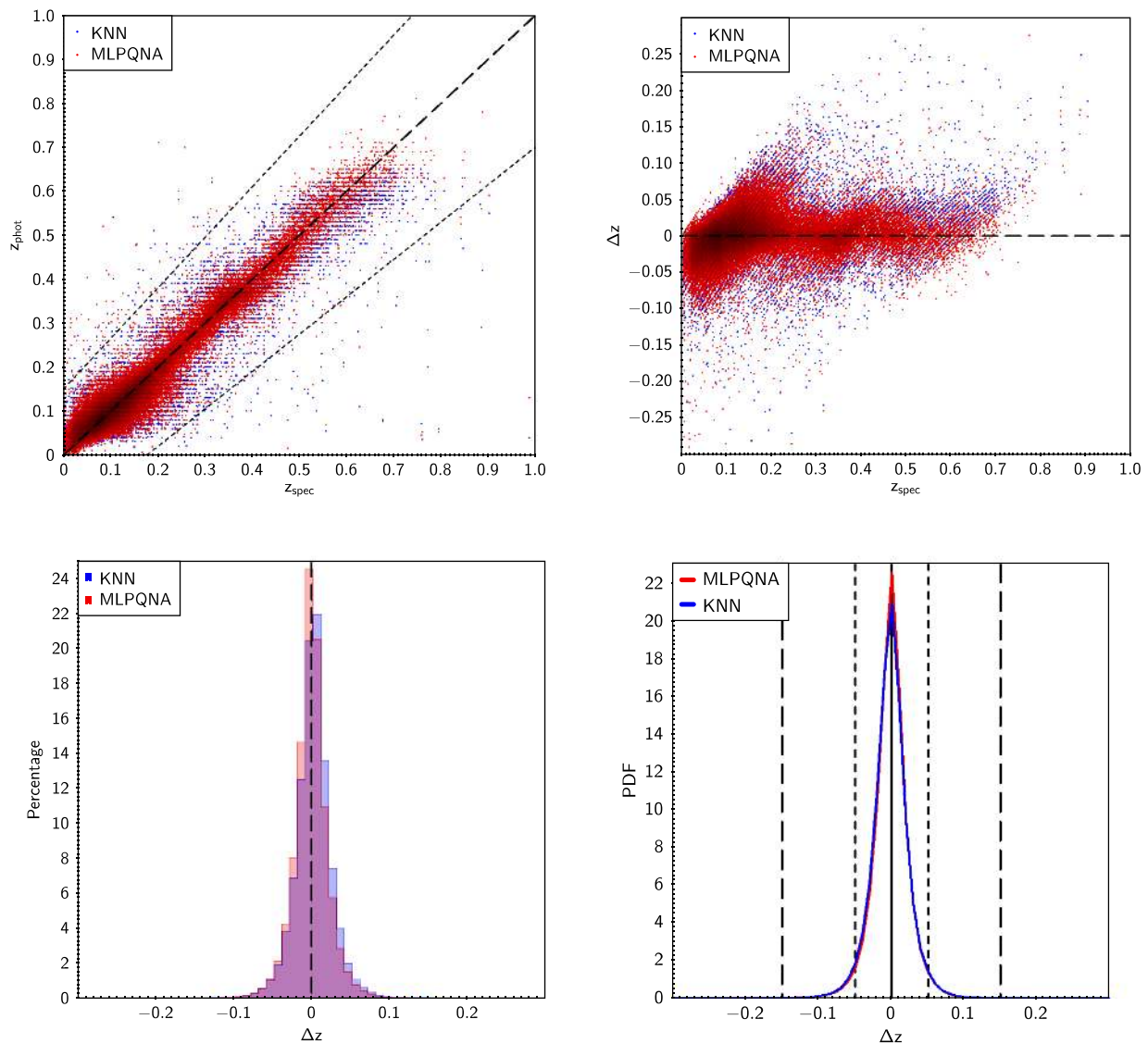
Figure 5. Comparison between MLPQNA (red) and KNN (blue). Left-hand panel of upper row: scatter plot of photometric redshifts as a function of spectroscopic redshifts ($z_{\rm spec}$ versus $z_{\rm phot}$); right-hand panel of upper row: scatter plot of residuals as a function of spectroscopic redshifts ($z_{\rm spec}$ versus Δz); left-hand panel of lower row: histograms of residuals (Δz); right-hand panel of lower row: *stacked* representation of residuals of the PDFs (the redshift binning is 0.01).

4.2 K-Nearest Neighbors

In a KNN model (Cover & Hart 1967), the input consists of the K closest training examples in the parameter space. A photo-z is estimated by averaging the targets of its neighbours. The KNN method is based on the selection of the N training objects closest to the object currently analysed. Here, closest has to be intended in terms of Euclidean distance among all photometric features of the objects. Our implementation makes use of the public library SCIKIT-LEARN (Pedregosa et al. 2011).

4.3 Random Forest

RF (Breiman 2001) is a supervised model which learns by generating a forest of random decision trees, dynamically built on the base of the variations in the parameter space of the training sample objects. Each single or group of such trees is assumed to become representative of specific types of data objects, i.e. the

best candidate to provide the right answer for a subset of data having similarities from the parameter space point of view. In the case of photometric redshifts, it has already been validated as a good estimator (Cavuoti et al. 2015b; Hoyle et al. 2015).

4.4 LE PHARE SED fitting

To test the METAPHOR workflow, we used the LE PHARE code as a benchmark to perform an SED template fitting experiment (Arnouts et al. 1999; Ilbert et al. 2006). SDSS observed magnitudes were matched with those predicted from a set of SEDs. Each SED template was redshifted in steps of $\Delta z = 0.01$ and convolved with the five SDSS filter transmission curves. The following merit function was then minimized:

$$\chi^2(z, T, A) = \sum_{i=1}^{N_{\rm f}} \left(\frac{F_{\rm obs}^{\rm f} - A \times F_{\rm pred}^{\rm f}(z, T)}{\sigma_{\rm obs}^{\rm f}} \right)^2, \tag{2}$$
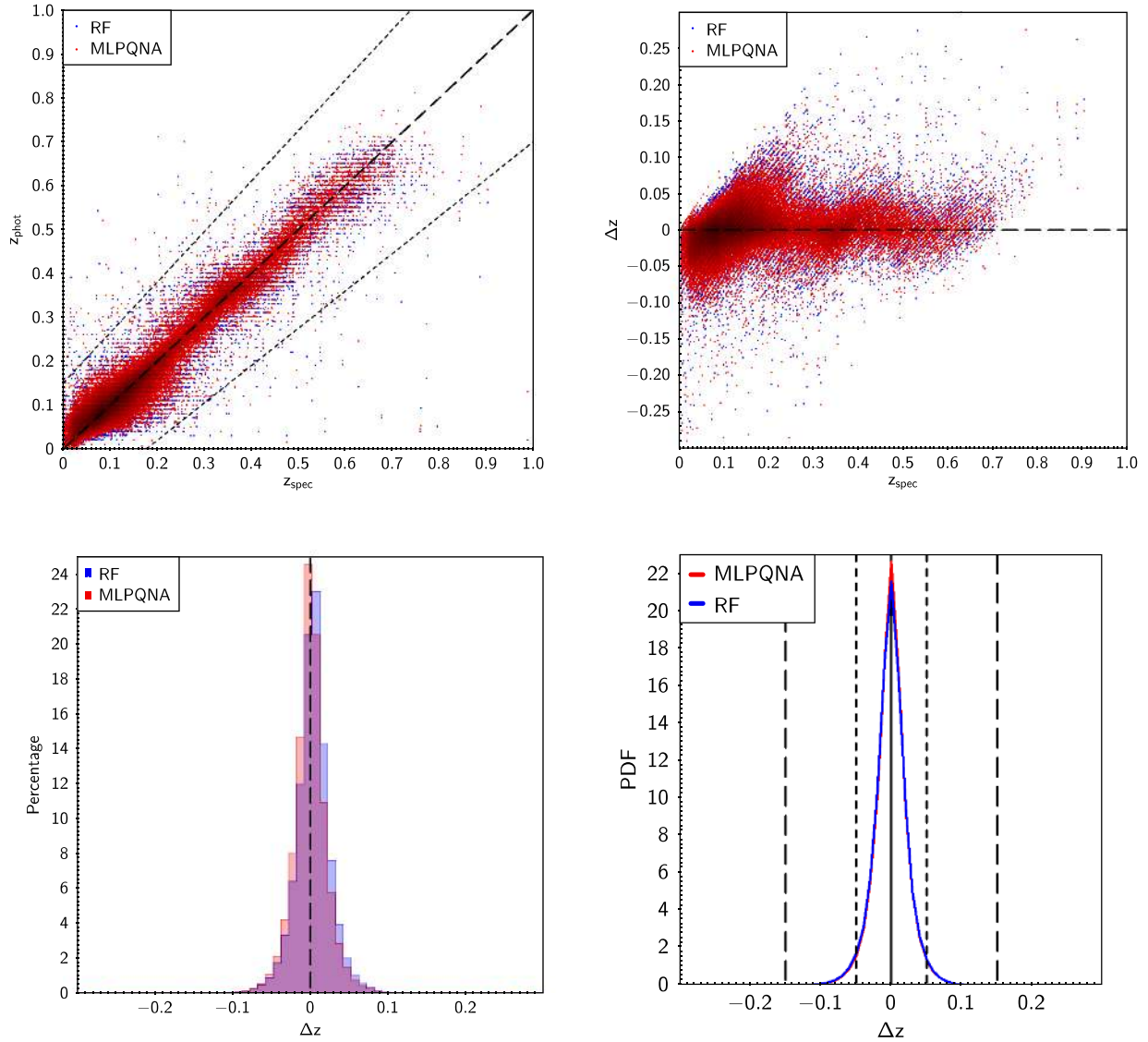
Figure 6. Comparison between MLPQNA (red) and RF (blue). Left-hand panel of upper row: scatter plot of photometric redshifts as a function of spectroscopic redshifts (z_{spec} versus z_{phot}); right-hand panel of upper row: scatter plot of residuals as a function of spectroscopic redshifts (z_{spec} versus Δz); left-hand panel of lower row: histograms of residuals (Δz); right-hand panel of lower row: *stacked* representation of residuals of the PDFs (the redshift binning is 0.01).

where $F_{pred}^{f}(z, T)$ is the flux predicted for a template T at redshift z. F_{obs}^{f} is the observed flux and σ_{obs}^{f} the associated error derived from the observed magnitudes and errors. The index 'f' refers to the considered filter and $N_f = 5$ is the number of filters. The photometric redshift is determined from the minimization of $\chi^2(z, T, A)$ varying the three free parameters: the photometric redshift, $z = z_{phot}$, the galaxy spectral type T, and the normalization factor A.

For the SED fitting experiments with LE PHARE, we used the SDSS *Modelmag* magnitudes in the *u, g, r, i* and *z* bands (and related 1σ uncertainties), corrected for galactic extinction using the map in Schlafly & Finkbeiner (2011). As a reference template set, we adopted the 31 SED models used for COSMOS photo-*z* (Ilbert et al. 2009). The basic COSMOS library is composed of galaxy templates from Polletta et al. (2007), which includes three SEDs of elliptical galaxies (E) and five templates of spiral galaxies (S0, Sa, Sb, Sc, Sd). These models are generated using the code GRASIL (Silva et al. 1998), providing a better joining of ultraviolet and mid-infrared than those by Coleman, Wu & Weedman (1980) used in

Ilbert et al. (2006). Moreover, to reproduce very blue colours not accounted for by the Polletta et al. (2007) models, 12 additional templates using Bruzual & Charlot (2003) models with starburst ages ranging from 3 to 0.03 Gyr have been added. In order to improve the sampling of the redshift–colour space and therefore the accuracy of the redshift measurements, the final set of 31 spectra was obtained by linearly interpolating the original templates. We have finally imposed the flat prior on absolute magnitudes, by forcing the galaxies to have absolute *i*-band magnitudes in the range of (−10, −26).

LE PHARE , as it is usual in the case of SED template-fitting techniques, provides the PDF for the estimated photo-*z* through the $\chi^2(z)$ distribution, which is defined as

$$PDF(z) \propto \exp\left(-\frac{\chi^2(z) - \chi_{min}^2}{2}\right), \qquad (3)$$

where χ_{min}^2 is the minimum of $\chi^2(z)$, corresponding to the best-fitting redshift.
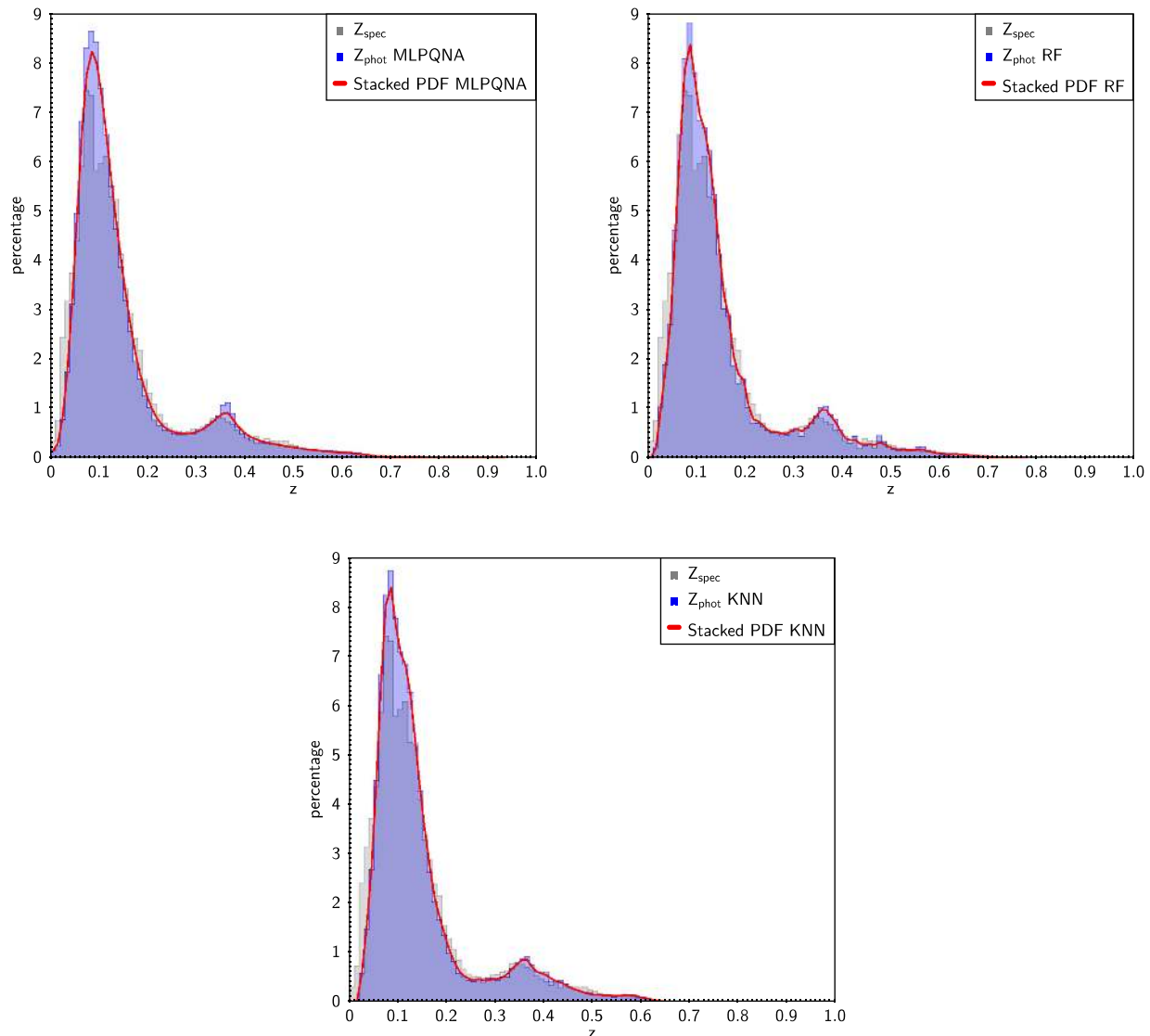
Figure 7. Superposition of the *stacked* PDF (red) and estimated photo-z (blue) distributions obtained by METAPHOR with, respectively, MLPQNA, RF and KNN on the z_{spec} distribution (in grey) of the blind test set.

We wish to stress that our main interest was to check the consistency of our ML-based results with PDFs from standard SED-fitting procedures, without running any competition among different methods. For this reason, we used a basic implementation of the LE PHARE code, not taking into account the systematics in the templates, data sets and optimizations (Brammer 2008; Ilbert et al. 2009; Tanaka 2015), and only imposing a flat prior on the absolute magnitudes. In literature, most of such systematics are taken into account introducing zero-point offsets and a template error function.

Zero-point offsets in the photometric bands due to a bad calibration and uncertainties in the model templates (e.g. stellar tracks, extinction law and other features not included in the spectra) can produce shifts between the predictions and real data. These average shifts are usually determined by means of an iterative process which minimizes the χ^2 for the spectroscopic sample with the redshift set to the z_{spec} value. Then, these shifts were applied to the magnitudes and used for the redshift determination (Ilbert et al. 2009). We have done some tests, and except for the more uncertain u band, for which the shift can also reach values of 0.1 mag or more, for the other bands the shifts are less than 0.01 mag; thus, for the

sample under analysis and for the main objectives of the paper, the contribution from zero-point shifts was negligible.

Since no template is immune to these systematics, in general it is also possible to introduce an error budget in the χ^2 minimization to account for them. However, this error budget would be less than $\sim$0.05 and varies a little across the wavelengths probed by SDSS bands (see e.g. Brammer 2008). Tanaka (2015) generalized the error function in Brammer (2008), adding a systematic flux stretch to the random flux uncertainty, used to reduce the mismatch between data and models. Both the terms account for systematics at a few per cent level in the optical wavelengths. The calculation of this error function could be coupled with zero-point shifts.

5 RESULTS AND DISCUSSION

In a previous paper (Brescia et al. 2014c), we already used the MLPQNA method to derive photometric redshifts for the SDSS-DR9, obtaining an accuracy better than the one presented here ($\sigma = 0.023$, bias $\sim$5 $\times$ 10^{-4} and $\sim$0.04 per cent of outliers against, respectively, 0.024, 0.0063 and 0.12 per cent). This apparent

Table 5. Tomographic analysis of the photo-z estimation performed by the MLPQNA on the blind test set.

| Estimator	Overall	]0, 0.1]	]0.1, 0.2]	]0.2, 0.3]	]0.3, 0.4]	]0.4, 0.5]	]0.5, 0.6]	]0.6, 1]
bias	−0.0006	−0.0002	−0.0002	−0.0008	−0.0010	0.0017	−0.0028	−0.0054
σ	0.024	0.022	0.024	0.029	0.027	0.027	0.031	0.040
σ_{68}	0.018	0.018	0.019	0.018	0.019	0.019	0.021	0.028
NMAD	0.017	0.017	0.016	0.016	0.017	0.016	0.019	0.027
skewness	−0.17	1.39	0.048	−1.26	−1.75	−2.58	−1.56	−3.30
outliers > 0.15	0.11%	0.04%	0.04%	0.60%	0.40%	0.40%	0.80%	0.60%

Table 6. Tomographic analysis of the PDF obtained by MLPQNA on the blind test set. Statistics of the *stacked* PDF obtained by MLPQNA.

Estimator	Overall	]0, 0.1]	]0.1, 0.2]	]0.2, 0.3]	]0.3, 0.4]	]0.4, 0.5]	]0.5, 0.6]	]0.6, 1]
$f_{0.05}$	91.7%	93.4%	91.2%	89.9%	90.2%	87.2%	83.8%	76.8%
$f_{0.15}$	99.8%	99.9%	99.9%	99.2%	99.5%	99.5%	99.2%	98.9%
$\langle \Delta z \rangle$	−0.0006	−0.0011	−0.0001	0.0005	−0.0018	0.0025	−0.0015	−0.0015
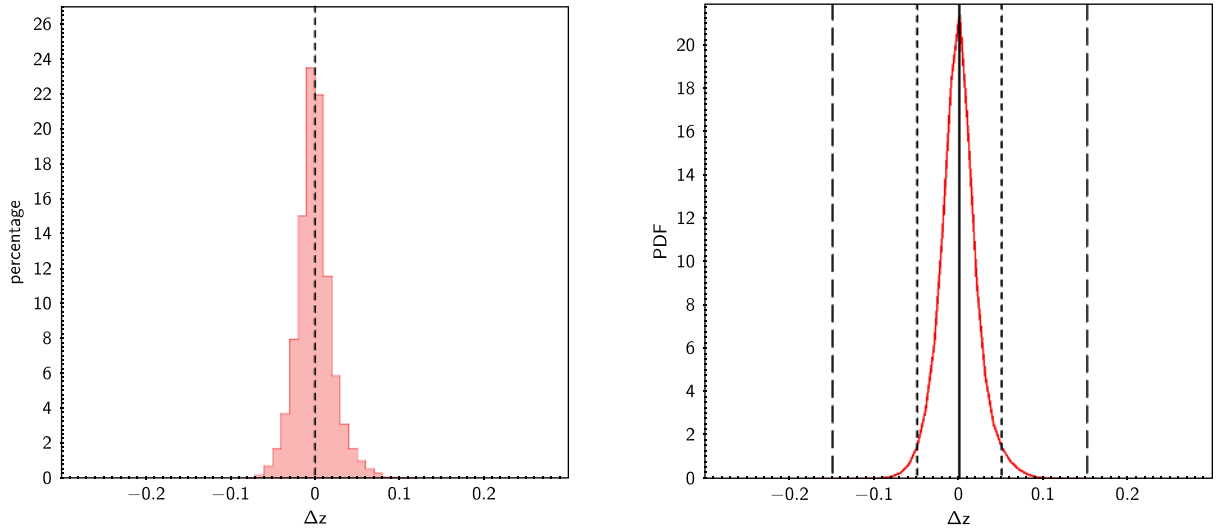
Figure 8. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0, 0.1]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.
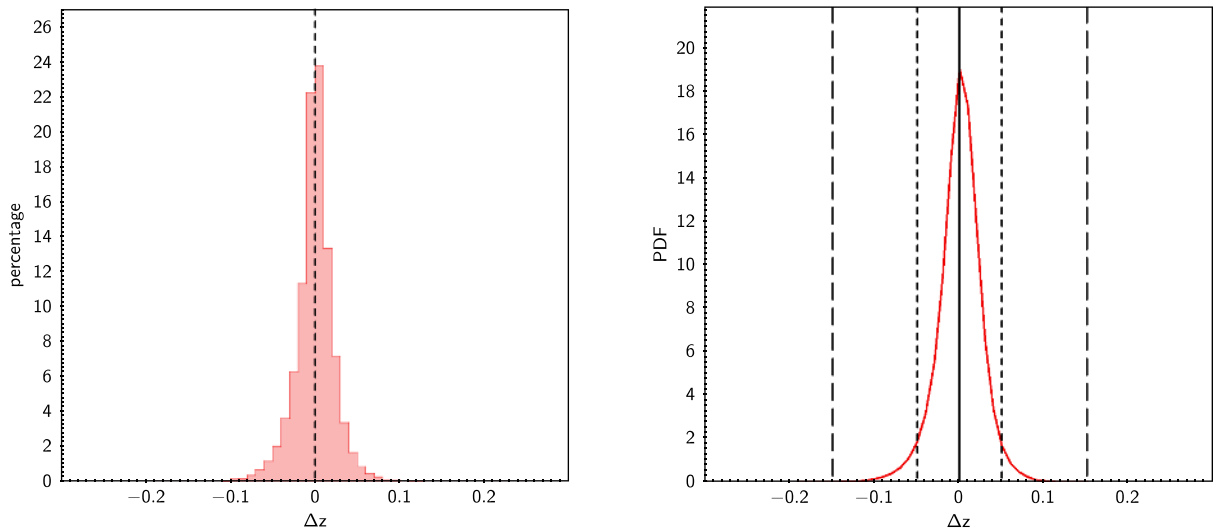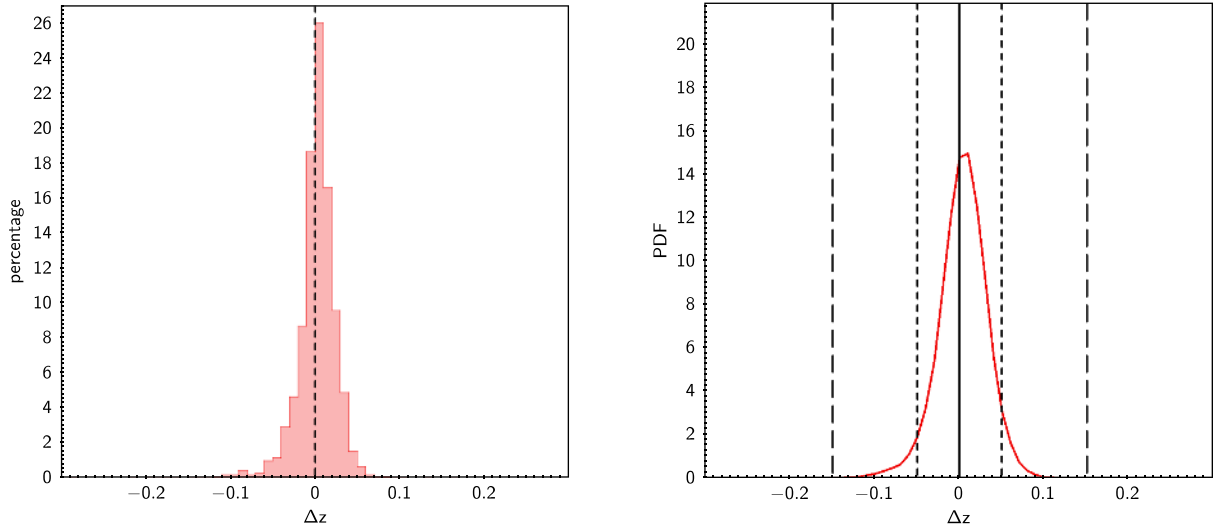

Figure 9. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0.1, 0.2]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.
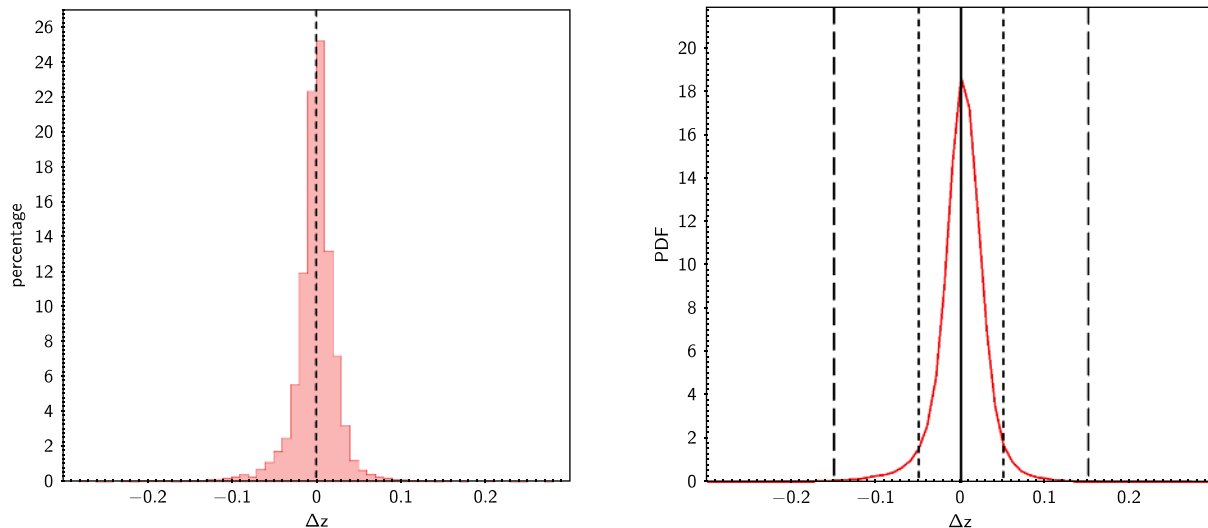
Figure 10. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0.2, 0.3]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.
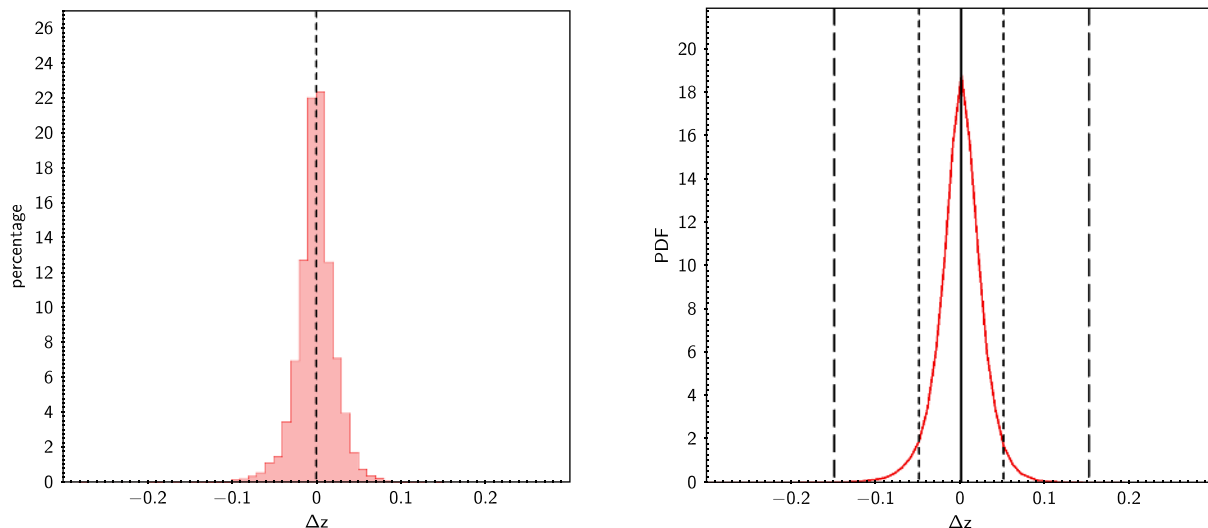
Figure 11. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0.3, 0.4]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.

discrepancy can be easily understood if we take into account the fact that the spectroscopic KB used in the previous work was much larger than the one used here (in Brescia et al. 2014c, ∼150 000 objects were used for the train and ∼348 000 for the test set, while in this work, only ∼50 000 objects were used for the training phase). The smaller KB used here is justified by the different purpose of this work which aims at assessing the quality of PDF derived by METAPHOR rather than at deriving a new catalogue of photo-z for the SDSS-DR9. The training phase of MLPQNA is in fact computationally intensive and the reduction of the training sample was imposed by the need to perform a large number of experiments.

The stacked PDF has been obtained by considering bin by bin the average values of the single PDFs. The cumulative statistics used to evaluate the stacked PDF quality have been derived by calculating the stacked PDF of the residuals Δz. In this way, aside from the evaluation of PDFs for single objects (a sub-sample is shown in Fig. 3), it is possible to obtain a cumulative evaluation within the most interesting regions of the error distribution.

In order to compare the different perturbation laws described in Section 2.2, we performed a variety of experiments with MLPQNA using 100 photometric perturbations. Results are summarized in Table 2. The most performing experiment turns out to be number 8, where we made use of a bimodal perturbation law with threshold 0.05 and a multiplicative constant $\alpha = 0.9$ (see equation 1). This experiment leads to a stacked PDF with ∼92 per cent within [−0.05, 0.05], $\sigma_{68} = 0.019$, ∼21 per cent of the objects falling within the peak of the PDF, ∼53 per cent falling within one bin from the peak and ∼82 per cent falling within the PDF. We therefore run an additional experiment using the same configuration as in number 8 but improving the error representation using 1000 perturbations. This experiment led to an increase in the performances: $\sigma_{68} = 0.018$ and ∼21.8 per cent within the peak of the PDF, ∼54.4 per cent within one bin from the peak and ∼89.6 per cent inside the PDF.

In order to verify the universality of the procedure with respect to the multitude of methods that could be used to estimate photo-z, the use of three different empirical models (for instance, MLPQNA,

Figure 12. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0.4, 0.5]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.
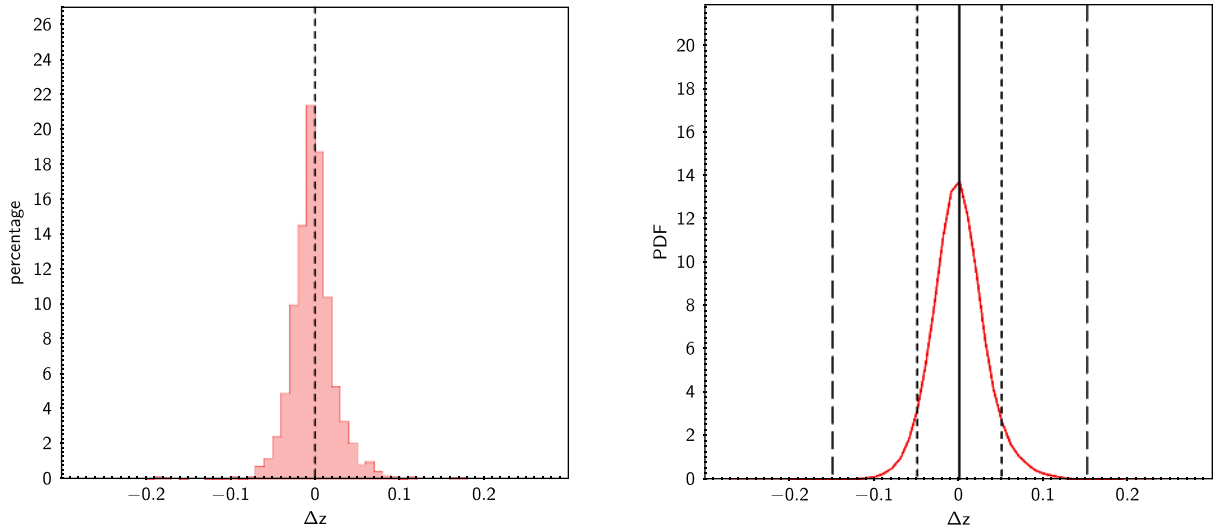
Figure 13. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0.5, 0.6]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.

RF and KNN) has been carried out. We also derived PDFs with the LE PHARE method, in order to evaluate the quality of the produced PDFs using a classical SED template fitting model as a benchmark. In Table 3, we report the results in terms of the standard set of statistical estimators used to evaluate the quality of predicted photo-z for all methods.

The results about the statistics of the stacked PDFs are shown in Table 4.

5.1 Comparison between METAPHOR and SED template fitting

Although there is a great difference in terms of statistical estimators between LE PHARE and MLPQNA, as it can be seen from Table 3 and first three panels of Fig. 4, the results of the PDFs in terms of $f_{0.15}$ are comparable (see Table 4 and the right-hand panel in the lower row of Fig. 4). But the greater efficiency of MLPQNA induces an improvement in the range within $f_{0.05}$, where we find

$\sim$92 per cent of the objects against the $\sim$72 per cent for LE PHARE. Both individual and *stacked* PDFs are more symmetric in the case of empirical methods presented here than for LE PHARE. This is particularly evident by observing the skewness (see Table 4), which is $\sim$100 times greater for the SED template fitting method; this can also be seen by looking at panels in the lower row of Fig. 4.

5.2 METAPHOR as general provider of PDF for empirical models

The model KNN performs slightly worse than MLPQNA in terms of σ and outliers rate (Table 3), as it can be seen by looking at the first three panels of Fig. 5, while RF obtains results which pose this model between KNN and MLPQNA in terms of statistical performance, as visible from Table 3 and panels of Fig. 6. The higher accuracy of MLPQNA causes a better performance of PDFs in terms of $\langle \Delta z \rangle$. However, also in the case of KNN and RF, METAPHOR is capable to produce reliable PDFs, comparable with those produced
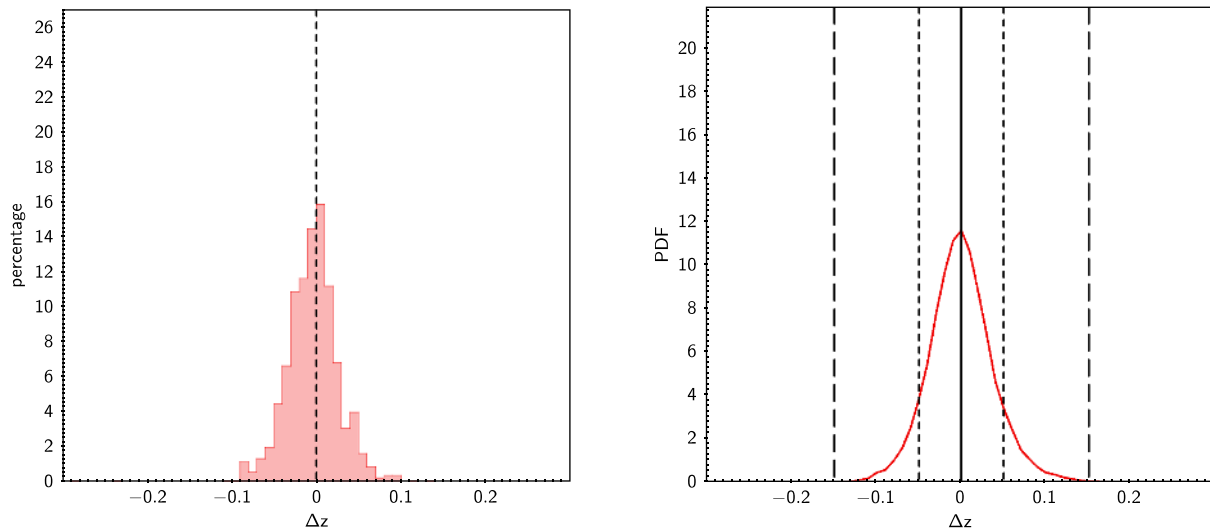
Figure 14. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0.6, 0.7]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.
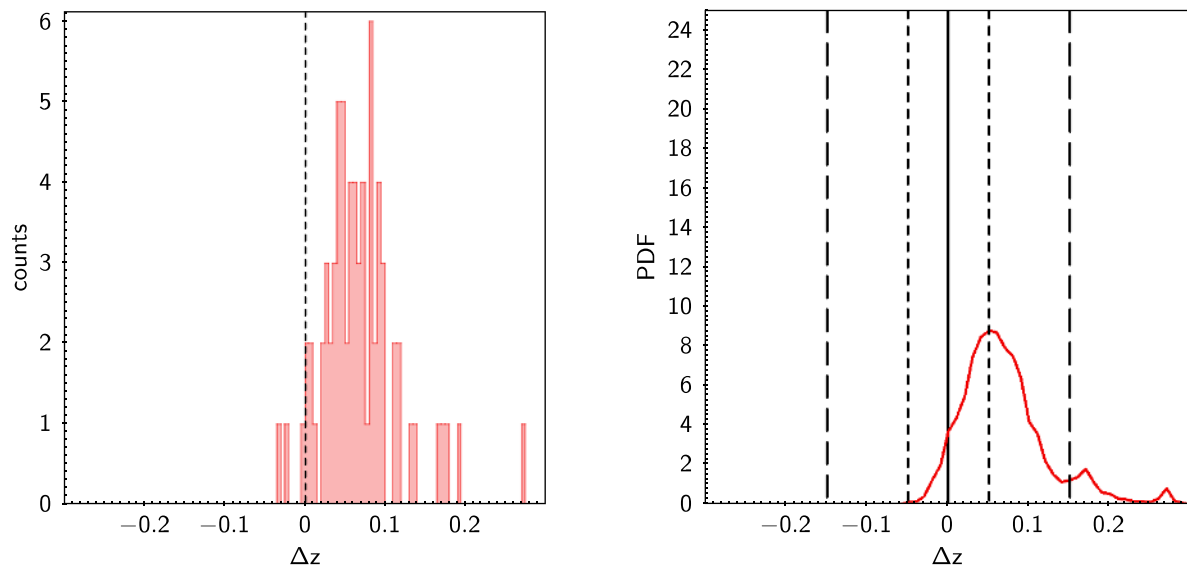

Figure 15. Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin]0.7, 1.0]. Upper panel: histogram of residuals (Δz); lower panel: *stacked* representation of residuals of the PDFs.

for MLPQNA (see Table 4 and right-hand panel in the lower row of Figs 5 and 6). This confirms the capability of METAPHOR to work efficiently with different empirical methods, regardless of their nature since even a very simple empirical model like KNN is able to produce high-quality PDFs.

The efficiency of the METAPHOR with the three empirical methods is particularly evident by looking at Fig. 7, where we show the *stacked* PDF and the estimated photo-z distributions, obtained by METAPHOR with each of the three models, superposed over the distribution of spectroscopic redshifts. The *stacked* distribution of PDFs, derived with the three empirical methods, results almost undistinguishable from the distribution of spectroscopic redshifts, with the exception of two regions: one in the peak of the distribution at around $z \simeq 0.1$ and the other at $z \simeq 0.4$. The first one can be understood in terms of a mild overfitting induced by the uneven distribution of objects in the training set. In fact around $z \simeq 0.1$ there are a large number of objects in the training set which induce

a bias causing a small reduction in the generalization capability. The second one ($z \simeq 0.4$) can be explained by the fact that the break at 4000 Å enters in the r band at this redshift. It induces an edge effect in the parameter space, which leads our methods to generate predictions biased away from the edges. However, biases in colour-space (averaging over/between degeneracies) specific to the SDSS filters clearly play a role as well.

By analysing the relation between the spectroscopic redshift and the produced PDFs, we find that about $\sim$22 per cent of z_{spec} falls in the bin PDF peak, but we emphasize that a further $\sim$33 per cent of z_{spec} falls one bin far from the peak (in our exercise, this means a distance of 0.01 from the peak). Finally, $\sim$10 per cent of the z_{spec} falls outside the PDF. We analysed the results in a tomographic way in order to verify whether there is a different behaviour in different regions. This has been done by cutting the output in bins of z_{phot} (the best guess of our method) and deriving the whole statistics bin by bin. Results are shown in Tables 5 and 6 and in Figs 8–15.
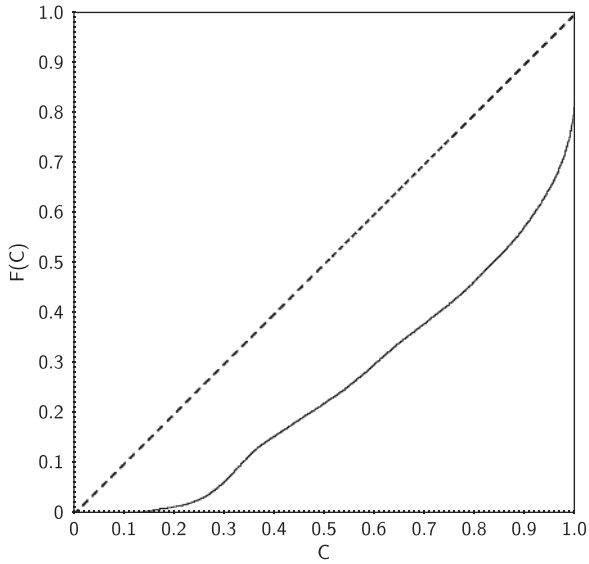
Figure 16. Credibility analysis (Wittman et al. 2016) of the PDFs.

In order to analyse the level of confidence of our PDFs, we performed a test using the credibility analysis presented in Wittman, Bhaskar & Tobin (2016). The diagram shown in Fig. 16 indicates an overconfidence of our method. We notice, however, that this test is more suitable for continuous distribution functions and in our case is likely to introduce some artefacts in the low-credibility region.

6 CONCLUSIONS

In the general scenario of the photometric redshift (photo-z) estimation, a probability density estimation (PDF) should provide a robust estimate of the reliability of any individual redshift, and it is strictly dependent on the measurement methods and on the physical assumptions done. In the absence of systematics, the main factors which affect the photo-z reliability are photometric errors, internal errors of the methods and statistical biases. The redshift inference has intrinsic uncertainties due to the fact that the available observables cannot be perfectly mapped to the true redshift. Therefore, the PDF is an effective way to parametrize the uncertainty on the solution for photo-z estimation.

SED template fitting methods intrinsically provide a photo-z PDF for each data object. In contrast, the PDF characterization for empirical methods is a challenging problem, widely discussed in the recent literature. In fact, it is much harder to obtain a PDF for photo-z predicted by empirical methods, in particular for those based on ML techniques, due to their hidden way to find the flux–redshift correlations in the parameter space. From a theoretical point of view, the characterization of photo-z predicted by empirical methods should be based on the real capability to evaluate the distribution of the photometric errors, to identify the correlation between photometric and spectroscopic error contributions and to disentangle the photometric uncertainty contribution from that one internal to the method itself.

In this work, we introduce METAPHOR, a method designed to provide a reliable PDF of the error distribution of photometric redshifts predicted by empirical methods. The method is implemented as a modular workflow, whose internal engine for photo-z estimation is based on the MLPQNA neural network . The METAPHOR procedure can, however, be applied by making use of any arbitrary empirical photo-z estimation model. One of the most important goals of

this work was to verify the universality of the procedure with respect to different interpolative models. For this reason, we experimented the METAPHOR processing flow on three alternative empirical methods. Besides the canonical choice of MLPQNA, a powerful neural network that we developed and tested on many photo-z estimation experiments, the alternative models selected were RF and KNN. In particular, the choice of KNN has been mainly driven by its extreme simplicity with respect to the wide family of interpolation techniques. We tested the METAPHOR strategy and the photo-z estimation models on a sample of the SDSS DR9 public galaxy catalogue.

The presented photo-z estimation results and the statistical performance of the cumulative PDFs, achieved by MLPQNA, RF and KNN through the proposed procedure, demonstrate the validity and reliability of the METAPHOR strategy, despite its simplicity, as well as its general applicability to any other empirical method.

REFERENCES

Annis J. T., 2013, AAS Meeting, 221, 335.05
Annunziatella M. et al., 2016, A&A, 585, A160
Aragon Calvo M. A., Weygaert R. van de, Jones Bernard J. T., Mobasher B., 2015, MNRAS, 454, 463
Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
Bolzonella M., Miralles J. M., Pello R., 2000, A&A, 363, 476
Bonnet C., 2013, MNRAS, 449, 1043
Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503
Breiman L., 2001, Mach. Learn., 45, 25
Brescia M., Cavuoti S., Paolillo M., Longo G., Puzia T., 2012, MNRAS, 421, 1155
Brescia M., Cavuoti S., D'Abrusco R., Mercurio A., Longo G., 2013, ApJ, 772, 140
Brescia M. et al., 2014a, PASP, 126, 783
Brescia M., Cavuoti S., Longo G., De Stefano V., 2014b, A&A, 568, A126
Brescia M., Cavuoti S., Longo G., De Stefano V., 2014c, VizieR Online Data Catalog, 568, A126
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Byrd R. H., Nocedal J., Schnabel R. B., 1994, Math. Program., 63, 129
Capozzi D., de Filippis E., Paolillo M., D'Abrusco R., Longo G., 2009, MNRAS, 396, 900
Carrasco K., Brunner R. J., 2013, in Friedel D. N., ed., ASP Conf. Ser. Vol. 475, Astronomical Data Analysis Software and Systems XXII. Astron. Soc. Pac., San Francisco, p. 69
Carrasco K., Brunner R. J., 2014a, MNRAS, 438, 3409
Carrasco K., Brunner R. J., 2014b, MNRAS, 442, 3380
Cavuoti S., Brescia M., Longo G., Mercurio A., 2012a, A&A, 546, 13
Cavuoti S., Brescia M., Longo G., 2012b, in Radziwill N. M., Chiozzi G., eds, Proc. SPIE Conf. Ser. Vol. 8451, Software and Cyberinfrastructure for Astronomy II. SPIE, Bellingham, p. 845103
Cavuoti S., Brescia M., D'Abrusco R., Longo G., Paolillo M., 2014, MNRAS 437, 968
Cavuoti S. et al., 2015a, MNRAS, 452, 3100
Cavuoti S., Brescia M., De Stefano V., Longo G., 2015b, Exp. Astron., 39, 45

Cavuoti S. et al., 2016, MNRAS, in press
Coleman G. D., Wu C.-C., Weedman D. W., 1980, ApJS, 43, 393
Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655
Cover T. M., Hart P. E., 1967, IEEE Trans. Inf. Theory, 13, 21
de Jong J. T. A. et al., 2015, A&A, 582, A62
Geisser S., 1975, J. Am. Stat. Assoc., 70, 320
Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015, MNRAS, 449, 1275
Ilbert O. et al., 2006, A&A, 457, 841
Ilbert O. et al., 2009, ApJ, 690, 1236
Ivezic Z., 2009, APS April Meeting Abstracts, May 2–5, # W4.003
Kaiser N., 2004, Proc. SPIE, 5489, 11
Laureijs R. et al., 2014, Proc. SPIE, 9143, 8
Ma Z., Bernstein G., 2008, ApJ, 682, 39
Mandelbaum R. et al., 2008, MNRAS, 386, 781

Masters D. et al., 2015, ApJ, 813, 53
Newman J. A. et al., 2015, Astropart. Phys., 63, 81
Pedregosa et al., 2011, J. Mach. Learn. Res., 12, 2825
Polletta M. et al., 2007, ApJ, 663, 81
Rau M. M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, MNRAS, 452, 3710
Riccio G., Brescia M., Cavuoti S., Mercurio A., di Giorgio A. M., Molinari S., 2016, PASP, preprint (arXiv:1611.04431)
Sadeh I., Abdalla F. B., Lahav O., 2015, PASP, 128, 104502
Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103
Serjeant S., 2014, AJ, 793, L10
Silva L., Granato G. L., Bressan A., Danese L., 1998, ApJ, 509, 103
Tanaka M., 2015, AJ, 801, 20
Tortora C. et al., 2016, MNRAS, 457, 2845
Wittman D., Bhaskar R., Tobin R., 2016, MNRAS, 457, 4005
York D. G. et al., 2000, AJ, 120, 1579

APPENDIX A: SPECTROSCOPIC QUERY

The following SQL code has been used to obtain the spectroscopic KB to train and test the model.

```sql
SELECT
  p.objid, s.specObjID, p.ra, p.dec,
  p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i,
  p.psfMag_z, p.psfmagerr_u, p.psfmagerr_g,
  p.psfmagerr_r, p.psfmagerr_i, p.psfmagerr_z,
  p.fiberMag_u, p.fiberMag_g, p.fiberMag_r,
  p.fiberMag_i, p.fiberMag_z, p.fibermagerr_u,
  p.fibermagerr_g, p.fibermagerr_r,
  p.fibermagerr_i, p.fibermagerr_z,
  p.petroMag_u, p.petroMag_g, p.petroMag_r,
  p.petroMag_i, p.petroMag_z, p.petromagerr_u,
  p.petromagerr_g, p.petromagerr_r,
  p.petromagerr_i, p.petromagerr_z,
  p.modelMag_u, p.modelMag_g, p.modelMag_r,
  p.modelMag_i, p.modelMag_z,
  p.modelmagerr_u, p.modelmagerr_g,
  p.modelmagerr_r, p.modelmagerr_i,
  p.modelmagerr_z,
  p.extinction_u, p.extinction_g,
  p.extinction_r, p.extinction_i,
  p.extinction_z, s.z as zspec,
  s.zErr as zspec_err, s.zWarning,
  s.class, s.subclass, s.primTarget
INTO
  mydb.galaxies_spec
FROM
  PhotoObjAll as p,
  SpecObj as s
WHERE
  s.class = 'GALAXY' AND s.zWarning = 0 AND
  p.mode = 1 AND p.SpecObjID = s.SpecObjID AND
  dbo.fPhotoFlags('PEAKCENTER') != 0 AND
  dbo.fPhotoFlags('NOTCHECKED') != 0 AND
  dbo.fPhotoFlags('DEBLEND_NOPEAK') != 0 AND
  dbo.fPhotoFlags('PSF_FLUX_INTERP') != 0 AND
  dbo.fPhotoFlags('BAD_COUNTS_ERROR') != 0 AND
  dbo.fPhotoFlags('INTERP_CENTER') != 0
```

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