Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **482,** 3116–3134 (2019)
Advance Access publication 2018 October 30

doi:10.1093/mnras/sty2922

Statistical analysis of probability density functions for photometric redshifts through the KiDS-ESO-DR3 galaxies

V. Amaro,[1] S. Cavuoti ,[1,2,3]★ M. Brescia ,[2] C. Vellucci,[4] G. Longo,[1] M. Bilicki ,[5,6] J. T. A. de Jong,[7] C. Tortora,[7] M. Radovich,[8] N. R. Napolitano[2] and H. Buddelmeijer [5]

[1]*Department of Physical Sciences, University of Napoli Federico II, via Cinthia 9, I-80126 Napoli, Italy*
[2]*INAF - Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy*
[3]*INFN section of Naples, via Cinthia 6, I-80126, Napoli, Italy*
[4]*DIETI, University of Naples Federico II, Via Claudio, 21, I-80125 Napoli, Italy*
[5]*Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, the Netherlands*
[6]*National Centre for Nuclear Research, Astrophysics Division, PO Box 447, PL-90-950 Łódź, Poland*
[7]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, NL-9700 AV Groningen, the Netherlands*
[8]*INAF - Osservatorio Astronomico di Padova, via dell'Osservatorio 5, I-35122 Padova, Italy*

Accepted 2018 October 23. Received 2018 September 21; in original form 2017 October 26

ABSTRACT

Despite the high accuracy of photometric redshifts (zphot) derived using machine learning (ML) methods, the quantification of errors through reliable and accurate probability density functions (PDFs) is still an open problem. First, because it is difficult to accurately assess the contribution from different sources of errors, namely internal to the method itself and from the photometric features defining the available parameter space. Secondly, because the problem of defining a robust statistical method, always able to quantify and qualify the PDF estimation validity, is still an open issue. We present a comparison among PDFs obtained using three different methods on the same data set: two ML techniques, METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts) and ANNz2 , plus the spectral energy distribution template-fitting method, BPZ (Bayesian photometric redshift). The photometric data were extracted from the Kilo Degree Survey ESO Data Release 3, while the spectroscopy was obtained from the Galaxy and Mass Assembly Data Release 2. The statistical evaluation of both individual and *stacked* PDFs was done through quantitative and qualitative estimators, including a *dummy* PDF, useful to verify whether different statistical estimators can correctly assess PDF quality. We conclude that, in order to quantify the reliability and accuracy of any zphot PDF method, a combined set of statistical estimators is required.

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

1 INTRODUCTION

Redshifts, by allowing the calculation of distances for large samples of galaxies, are at the core of most extragalactic and cosmological studies and are needed for many purposes, such as, to quote just a few, to constrain the dark matter and dark energy contents of the Universe through weak gravitational lensing (Serjeant 2014; Hildebrandt et al. 2017; Fu et al. 2018), to reconstruct the cosmic large-scale structure (LSS; Aragon et al. 2015), to identify galaxy clusters and groups (Capozzi et al. 2009; Annunziatella et al. 2016; Radovich et al. 2017), to disentangle the nature of astronomical sources (Brescia et al. 2012; Tortora et al. 2016), to map the galaxy colour–redshift relationships (Masters et al. 2015), and to measure the baryonic acoustic oscillations spectrum (Gorecki et al. 2014; Ross et al. 2017).

The last few years have seen a proliferation of multiband photometric galaxy surveys, either ongoing (see KiDS – Kilo-Degree Survey, de Jong et al. 2015, 2017; DES – Dark Energy Survey, Annis et al. 2013) or planned (LSST, Ivezic 2009, LSST Science Book 2009 and Euclid, Laureijs et al. 2014, Euclid Red Book 2011). All these surveys require redshift estimates for hundreds of millions or billions of galaxies that cannot be observed spectroscopically and therefore must be obtained via multiband photometry (photometric

★ E-mail: stefano.cavuoti@gmail.com

redshifts or zphot). This is possible due to the existence of a highly non-linear correlation between photometry and redshift, caused by the fact that the stretching introduced by the redshift induces the main spectral features to move through the different filters of a photometric system (Baum 1962; Connolly et al. 1995).

In a broad but widespread oversimplification, there are two main classes of methods commonly used to derive zphot: the spectral energy distribution (SED) template-fitting methods (e.g. Arnouts et al. 1999; Bolzonella, Miralles & Pello 2000; Ilbert et al. 2006; Tanaka 2015) and the empirical (or interpolative) methods (e.g. Firth, Lahav & Somerville 2003; Ball et al. 2008; Carrasco & Brunner 2013b; Brescia et al. 2014b; Graff et al. 2014; Cavuoti et al. 2015a,b; Masters et al. 2015; Sadeh, Abdalla & Lahav 2016; Soo et al. 2017; D'Isanto & Polsterer 2018), both characterized by their advantages and shortcomings. There are also recent experiments that try to combine these two zphot estimation classes, in order to merge their respective capabilities (e.g. Cavuoti et al. 2017b; Duncan et al. 2017; Hoyle & Rau 2018).

SED template-fitting methods are based on a fit (generally a χ^2 minimization) to the multiband photometric observations of the objects. The starting point is a set of template (either synthetic or observed) spectra covering different morphological types and physical properties. Each of these template SEDs is convolved with the transmission functions of any given filters, in order to create synthetic magnitudes as a function of the redshift.

SED fitting methods are capable to derive all at once the zphot, the spectral type, and the probability density function (PDF) of the error distribution of each source. However, these methods suffer from several shortcomings: the potential mismatch between the templates used for the fitting, the properties of the selected sample of galaxies (Abdalla et al. 2011), colour/redshift degeneracies, and template incompleteness. Such issues are stronger at high redshift, where galaxies are fainter and photometric errors larger. Furthermore, at high redshifts there are fewer or no empirical spectra available to build a reliable template library.

Among empirical methods, those based on various Machine Learning (ML) algorithms are the most frequently used. They infer (not analytically) the complex relation existing between the input, mainly multiband photometry (i.e. fluxes, magnitudes, and/or derived colours) and the desired output (the spectroscopic redshift, hereafter zspec). In supervised ML, the learning process is regulated by the spectroscopic information (i.e. redshift) available for a subsample of the objects, whereas in the unsupervised approach, the spectroscopic information is not used in the training phase, but only during the validation phase. There are many ML algorithms that have been used for zphot estimation. To quote just a few: neural networks (Tagliaferri et al. 2002; Collister & Lahav 2004; Brescia et al. 2013; Sadeh, Abdalla & Lahav 2015), boosted decision trees (Gerdes et al. 2010), random forests (Carrasco & Brunner 2013a), and self-organized maps (Carrasco & Brunner 2014a; Masters et al. 2015). ML techniques are endowed with several advantages: (i) high accuracy of predicted zphot within the limits imposed by the spectroscopic knowledge base (hereafter KB); (ii) ability to easily incorporate external information in the training, such as surface brightness, angular sizes, or galaxy profiles (Tagliaferri et al. 2002; Cavuoti et al. 2012; Soo et al. 2017; Bilicki et al. 2018).

On the other hand, ML methods have a very poor capability to extrapolate information outside the regions of the parameter space properly sampled by the training data that, for instance, implies that they cannot be used to estimate redshifts for objects fainter than those present in the spectroscopic sample. Furthermore, supervised methods are viable only if accurate photometry and spectroscopy are available for a quite large (few thousands of objects at least) number of objects. See Hildebrandt et al. (2010), Abdalla et al. (2011), Sánchez et al. (2014) for reviews about the zphot estimation techniques.

Finally, due to their intrinsic nature of self-adaptive learning models, the ML based methods do not naturally provide a PDF estimate of the predicted zphot, unless special procedures are implemented.

In recent years, it has been demonstrated in several studies that PDFs can increase the accuracy of cosmological parameter measurements. For example, Mandelbaum et al. (2008) have shown that most common statistics (bias, outlier rate, standard deviation etc.) are not sufficient to evaluate the accuracy of zphot required by weak-lensing (WL) studies. In particular, the measurement of the critical mass surface density requires a reliable PDF estimation to remove any calibration bias effect.

Over the last few years, particular attention has been paid to develop techniques and procedures able to compute a full zphot PDF for an astronomical source as well as for an entire galaxy sample (Bonnet 2013; Carrasco & Brunner 2013a, 2014a,b). The PDF contains more information than the single-redshift estimate, as it is also confirmed by the improvement in the accuracy of cosmological and WL measurements (Mandelbaum et al. 2008; Viola et al. 2015), when PDFs are used rather than zphot point estimates. However, to the best of our knowledge, the positive role played by the PDFs has been demonstrated only for zphot obtained with SED fitting methods.

In this paper, we perform a comparative analysis of zphot and associated PDF performance among different methods. The data used for this analysis were extracted from the KiDS ESO (European Southern Observatory) Data Release 3 (hereafter, KiDS-ESO-DR3), described in de Jong et al. (2017). In that work, three different methods for photometric redshifts were used and the corresponding catalogues made publicly available:[1] two ML methods, respectively, Machine-learning Estimation Tool for Accurate PHOtometric Redshifts (METAPHOR; Cavuoti et al. 2017a) and ANNz2 (Sadeh et al. 2016; Bilicki et al. 2018), plus one template-fitting method, the Bayesian photometric redshifts (hereafter, BPZ, Benitez 2000). For the purpose of this paper, we also build a *dummy* PDF, independent of method errors and photometric uncertainties, useful to compare and assess the statistical estimators used to evaluate the reliability of PDFs.

The paper is structured as follows: In Section 2, we present the KiDS-ESO-DR3 data used for the analysis. In Section 3, we give a general overview about the calculation of PDFs, and we describe the methods as well as the statistical estimators involved in our analysis. In Section 4, we perform the comparison among the PDF methods and a critical discussion about the statistical estimators. Finally, in Section 5 we draw our conclusions.

2 THE DATA

The sample of galaxies used to estimate zphot and their individual and stacked PDFs was extracted from the third data release of the ESO Public Kilo-Degree Survey (KiDS-ESO-DR3, de Jong et al. 2017). When completed, the KiDS survey will cover 1500 deg^2 (de Jong et al. 2017), distributed over two survey fields, in four broad-band filters (*u, g, r, i*). Compared to the previous data releases

[1] Available at http://kids.strw.leidenuniv.nl/DR3/ml-photoz.php

Table 1. Brighter and fainter limits imposed on the magnitudes and defining the region of the parameter space used for training and test experiments.

Input magnitudes	Brighter limit	Fainter limit
MAG_APER_20_U	16.84	28.55
MAG_APER_30_U	16.81	28.14
MAG_GAAP_U	16.85	28.81
MAG_APER_20_G	16.18	24.45
MAG_APER_30_G	15.86	24.59
MAG_GAAP_G	16.02	24.49
MAG_APER_20_R	15.28	23.24
MAG_APER_30_R	14.98	23.30
MAG_GAAP_R	15.15	23.29
MAG_APER_20_I	14.90	22.84
MAG_APER_30_I	14.56	23.07
MAG_GAAP_I	14.75	22.96

(de Jong et al. 2015), the DR3 not only covers a larger area of the sky, but it also relies on an improved photometric calibration and provides photometric redshifts along with shear catalogues and lensing-optimized image data. The total DR3 data set consists of 440 tiles for a total area covering approximately 450 deg^2, with respect to the 160 deg^2 of the previous releases.

The DR3 provides also an aperture-matched multiband catalogue for more than 48 million sources, including homogenized photometry based on Gaussian Aperture and point spread function (hereafter GAaP) magnitudes (Kuijken 2008). All the measurements (star/galaxy separation, source position, shape parameters) are based on the *r*-band images due to their better quality (see table A.2 of de Jong et al. 2017).

KiDS was primarily designed for WL studies, in order to reconstruct the LSS of the Universe. Indeed, the first 148 tiles of the first two data releases produced their first scientific results on WL for galaxies and groups of galaxies in the Galaxy And Mass Assembly (GAMA, Driver et al. 2011) fields (de Jong et al. 2015), as the reader can find in Viola et al. (2015).

The photometry used in this work consists of the *ugri* GAaP magnitudes, two aperture magnitudes, measured within circular apertures of 4 and 6 arcsec diameter (20 and 30 pixels, referred in Table 1 as $MAG\_APER\_20\_X$ and $MAG\_APER\_30\_X$), respectively, corrected for extinction and zero-point offsets and the derived colours, for a total of 21 photometric parameters for each object.

The original data set was cleaned by removing objects affected by missing information (the performance of ML methods may degrade if data are missing) and by clipping the tails of the magnitude distributions in order to ensure a proper density of training points in the sampled regions of the parameter space. The lower and upper cuts, applied to exclude the tails of the distributions, are reported in Table 1.

Furthermore, as we shall specify in Section 3.1, the fundamental concept of the PDF estimation in METAPHOR is the perturbation of the data photometry, based on a proper fitting function of the flux errors in specifically defined bins of flux. Therefore, in the preparation phase, we excluded from the KB all entries with a photometric error higher than a given threshold (e.g. 1 magnitude) in order to provide a data set used for the polynomial fitting of the errors, as prescribed by the *mixture* perturbation law (see Section 3.1).

In order to perform a zphot comparison through a common spectroscopic base in the work of de Jong et al. (2017), each of the three zphot catalogues (obtained, respectively, by METAPHOR,[2] ANNz2, and BPZ) has been cross-matched in coordinates with the spectroscopic information extracted from the second data release (DR2) of GAMA (Liske et al. 2015), containing spectroscopy in the KiDS-North field (composed of 77 per cent objects from GAMA; 18 per cent from SDSS/BOSS, Ahn et al. 2014; and 5 per cent from 2dFGRS Colless et al. 2001). For what concerns this paper, since the ANNz2 catalogue released with DR3 does not include individual PDFs (Bilicki et al. 2018), these have been derived for the purposes of this work, by uniforming the training and test sets with those used by METAPHOR in de Jong et al. (2017). Finally, since in this work we were interested in performing the zphot PDF comparison among the two mentioned ML methods and BPZ using a uniform data sample, we followed the same approach as de Jong et al. (2017), based on the cross-matching between KiDS-DR3 photometry and SDSS DR9 + GAMA DR2 + 2dFGRS spectroscopy. As described in section 4.2 of de Jong et al. (2017), we performed a random shuffling and split procedure, obtaining a training set of ∼71 000 and a blind validation set of ∼18 000 objects. The final comparison test among methods has been done on a supplementary blind set of ∼64 000 galaxies, as detailed in section 4.4 of de Jong et al. (2017).

3 THE METHODS

In general terms, a PDF is a way to parametrize the uncertainty on the zphot solution. In the context of zphot estimation, a PDF is strictly dependent both on the measurement methods and on the physical assumptions. This simple statistical consideration renders the real meaning of PDFs quite complex to grasp in the case of zphot error evaluation.

Furthermore, a PDF should provide a robust estimate of the reliability of an individual redshift. The factors affecting such reliability are photometric errors, intrinsic errors of the methods, and statistical biases. In fact, under the hypothesis that a perfect reconstruction of the redshift is possible, the PDF would consist of a single Dirac delta. However, since the zphot cannot be perfectly mapped to the true redshift, the corresponding PDF represents the intrinsic uncertainties of the estimate. In other words, as anticipated in Section 1, PDFs are useful to characterize zphot estimates by providing more information than the simple estimation of the error on the individual measurements.

In the following paragraphs, we shortly summarize the characteristics of the methods used in our study. Besides the already mentioned ML methods (METAPHOR, ANNz2) and BPZ, we introduce also a special way to assess the validity of the statistical estimators used to measure the PDF reliability, called *dummy* PDF.

3.1 METAPHOR

The METAPHOR (Cavuoti et al. 2017a) method is a modular workflow, designed to produce both zphot and related PDFs. The internal zphot estimation engine is our model Multi Layer Perceptron trained with Quasi Newton Algorithm (MLPQNA; Brescia et al. 2013, 2014a). METAPHOR makes available a series of functional modules:

[2]In de Jong et al. (2017), it is referred to as MLPQNA, the internal zphot estimation engine of METAPHOR.

(i) Data pre-processing: data preparation, photometric evaluation of the KB, followed by its perturbation based on the given magnitude error distributions,

(ii) zphot prediction: single photometric redshift estimation, based on training/test of the KB with the MLPQNA model;

(iii) PDF estimation: production of the individual zphot PDFs and evaluation of their cumulative statistical properties.

As anticipated in the introduction, in the context of ML techniques, the determination of individual PDFs is a challenging task. This is because we would like to determine a PDF starting from several estimates of zphot, embedding the information on the photometric uncertainties on those estimates. Therefore, we derived an analytical law to perturb the photometry by taking into account the magnitude errors provided in the catalogues.

Indeed, the procedure followed to determine individual source PDFs consists of a single training of the MLPQNA model and the perturbation of the photometry of the given blind test set in order to obtain an arbitrary number N of test sets, each characterized by a variable photometric noise contamination. The decision to perform a single training is mainly motivated by the idea of excluding the contribution of the intrinsic error of the method itself from the PDF calculation. Appendix *A* is dedicated to the analysis of error contributions.

With this goal in mind, we use in this work the following *perturbation* law:

$$\tilde{m}_{ij} = m_{ij} + \alpha_i F_{ij} u_{(\mu=0,\sigma=1)}, \quad (1)$$

where j denotes the jth object's magnitude and i the reference band; α_i is a multiplicative constant, heuristically chosen by the user (generally useful to take into account cases of heterogeneous photometry, i.e. derived from different surveys and in this particular case fixed to 0.9 for all the bands); the term $u_{(\mu=0,\sigma=1)}$ is a random value from a standard normal distribution; finally, F_{ij} is the function used to perturb the magnitudes.

In this work, the selected perturbation function (F_{ij}) is the *mixture*, i.e. a function composed of a constant threshold (in this case heuristically fixed to 0.03) and a polynomial fitting of the average magnitude errors computed in several magnitude bins for each given photometric band. The role of the constant function is to act as a threshold under which the polynomial term is too low to provide a significant noise contribution to the perturbation (see Cavuoti et al. 2017a for further details; in that paper the *mixture* function was called *bimodal*). This choice was made in order to take into account that there are very low average errors for the brighter objects within the catalogues. These perturbations were applied to both GAaP and aperture–magnitude types.

For the calculation of the individual PDFs, we submit the $N + 1$ test sets (i.e. N perturbed sets plus the original one) to the trained model, thus obtaining $N + 1$ zphot estimates. Then, we perform a binning in zphot, thus calculating the probability that a given zphot value belongs to each bin. We selected a binning step of 0.01 for the described experiments and a value of N equal to 1000. The same binning step has been adopted by all three methods compared in this work.

In Fig. 1, we can see the *mixture* functions F_{ij} for the homogenized magnitudes *mag_gaap_x* (with x = *u,g,r,i*).

Concerning the zphot production, the *best-estimate* zphot values are not always corresponding to the given unperturbed catalogue estimate of zphot (hereafter photo-z_0), as calculated by MLPQNA. In particular, it coincides with photo-z_0 if this measurement falls into the interval (or *bin*) representing the *peak* (maximum) of the PDF; otherwise, it coincides with the zphot estimate (among the $N + 1$ zphot estimates mentioned above) closest to photo-z_0 and falling in the bin to which corresponds the PDF *peak*.

3.2 ANNz2

ANNz2 (Sadeh et al. 2016) is a versatile ML package,[3] designed primarily for deriving zphot, but appropriate also for other ML applications such as automated classification. The main ML method used by ANNz2 is based on artificial neural networks (ANNs), but it is also possible to employ boosted decision and regression trees; here we use the ANNs only. We work in the randomized regression mode of ANNz2, in which a (preferably) large number of randomly designed ANNs (100 in our case) are trained on the input spectroscopic calibration data. This ensemble of trained ANNs is used for deriving both zphot point estimates and their PDFs. Here, we provide a brief overview of the PDF generation procedure in the software version employed for this work, referring the reader to Sadeh et al. (2016) and to the online documentation of ANNz2 for more details.[4] Once the desired number of ANNs have been trained, then in the validation phase (called 'optimization' in ANNz2) each source from the spectroscopic validation set[5] is assigned to a distribution of zphot solutions from the individual ANNs. These solutions are then ranked by their performance, and the top one is used to derive the individual zphot estimate, `Z_BEST`, which we use in this paper as the zphot point estimation from ANNz2. In order to derive PDFs, the various ANNs are first folded with their respective single-value uncertainty estimates, derived via the k-nearest neighbour method (Oyaizu et al. 2008). A subset of ranked solutions is combined in different random ways to obtain a set of candidate PDFs. In order to select the final PDF, these candidates are compared using their cumulative distribution functions (CDFs), defined as the integrated PDF for redshifts smaller than the reference value of the true redshift, $z_{\rm spec}$:

$$\mathcal{C}(z_{\rm spec}) = \int_{z_0}^{z_{\rm spec}} p_{\rm reg}(z)\, {\rm d}z. \quad (2)$$

The function $p_{\rm reg}(z)$ is the differential PDF for a given redshift and z_0 is the lower bound of the PDF ($z_0 = 0$ in our case). The final PDF is chosen as the candidate for which the distribution of $\mathcal{C}$ is the closest to uniform (Bordoloi, Lilly & Amara 2010).

ANNz2 may generate two types of PDFs, depending on how the $\mathcal{C}$ function is chosen. In the first case, denoted as `PDF_0`, the CDF is based on $z_{\rm spec}$ from the validation sample; in the second option, `PDF_1`, the results of the best ML solution are used as reference. In this work, we use the `PDF_1` option as we found it to perform generally better than the other one.

3.3 BPZ

The BPZ method (Benitez 2000), as usual for SED fitting techniques, is able to provide a PDF estimation law, based on the equation

$$\chi^2(z, T, A) = \sum_{i=1}^{N_f} \left(\frac{F^f_{\rm obs} - A \times F^f_{\rm pred}(z, T)}{\sigma^f_{\rm obs}} \right)^2, \quad (3)$$

[3] Available from https://github.com/IftachSadeh/ANNZ

[4] Here, we used version 2.2.2 of the ANNz2 software, while some significant changes in the PDF estimation have been introduced since version 2.3.0.

[5] We used the ANNz2 option to randomly split the spectroscopic calibration sample into disjoint training and validation sets in proportion 1:1.

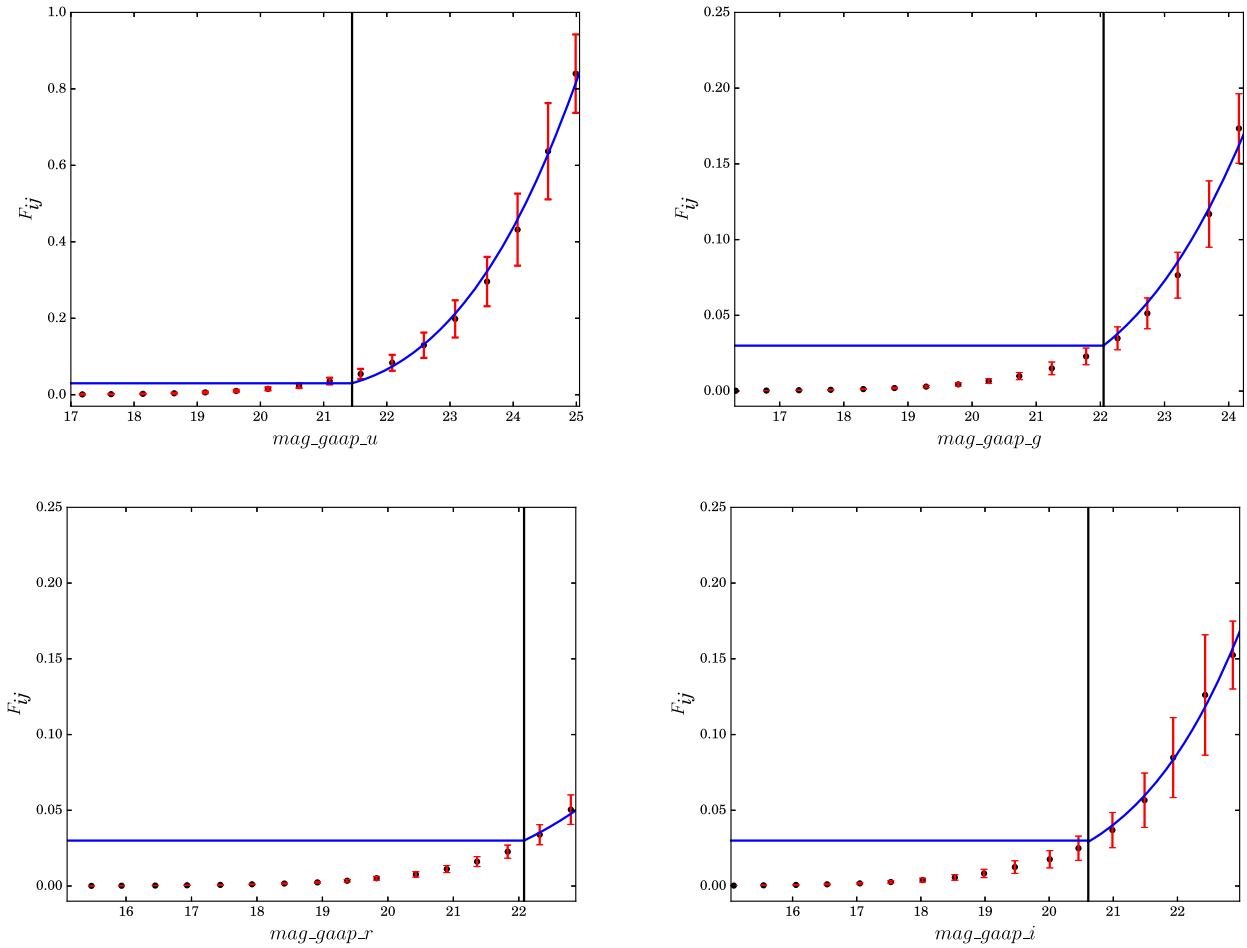

Figure 1. *Mixture* perturbation function F_{ij} in equation (1) for the KiDS GAaP magnitudes, composed of a flat perturbation for magnitudes lower than a selected threshold (black solid lines) and a polynomial perturbation $p_i(m_{ij})$ for higher magnitude values (see Section 3.1). The switching thresholds between the two functions are, respectively, 21.45 in *u* band, 22.05 in *g*, 22.08 in *r*, and 20.61 in *i* band. The black points are the average of the magnitude errors for each magnitude bin. The red lines report the corresponding standard deviation.

Table 2. Statistics of zphot estimation obtained with MLPQNA (zphot estimation engine of METAPHOR), ANNz2, BPZ, on the GAMA DR2: respectively, the bias, the standard deviation, the Normalized Median Absolute deviation, the fraction of outliers outside the 0.15 range, kurtosis, and skewness.

Estimator	MLPQNA	ANNz2	BPZ
bias	− 0.004	− 0.008	− 0.020
σ	0.065	0.078	0.048
NMAD	0.023	0.019	0.028
outliers	0.98 %	1.60 %	1.13 %
Kurtosis	774.1	356.0	52.2
Skewness	− 21.8	− 15.9	− 2.9

where $F^f_{\rm pred}(z, T)$ is the flux predicted for a template T at redshift z. $F^f_{\rm obs}$ is the observed flux, $\sigma^f_{\rm obs}$ the associated error, while A is a normalization factor. From equation (3), it is clear that the spectroscopic information is not needed, thus implying the possibility to estimate the zphot for all sources.

Individual PDFs are a natural by-product of every SED fitting method. In the case of BPZ for the KiDS-DR3 data, the PDFs are obtained by multiplying the probability by the used priors and then performing a summation over all the templates, in order to obtain the full posterior probability. The theory, implemented in the BPZ code, is expressed by equations (6–12) in the paper of Benitez (2000). This method has been used to obtain BPZ KiDS-DR3 zphot and PDFs, by utilizing the priors specified in Hildebrandt et al. (2012).

Finally, the reference to the selected re-calibrated template set (Capak 2004), as well as more details about the use of BPZ, are provided in de Jong et al. (2017).

3.4 Dummy PDF

In order to have a benchmark tool useful to analyse and compare the statistical validity of previous methods, we set to zero the multiplicative constant parameter α_i of equation (1) for all bands in order to produce a *dummy* perturbation law.

The relative *dummy* PDF obtained by METAPHOR is made by individual source PDFs, for which the one hundred per cent of the zphot estimates (coincident with photo-z_0, i.e. the unperturbed estimate of zphot) fall in the same redshift interval (by fixing the binning step at 0.01, as described in Section 3.1).

The main goal in determining the *dummy* PDFs is to assess the reliability of several statistical estimators used to evaluate an ensemble of PDFs.

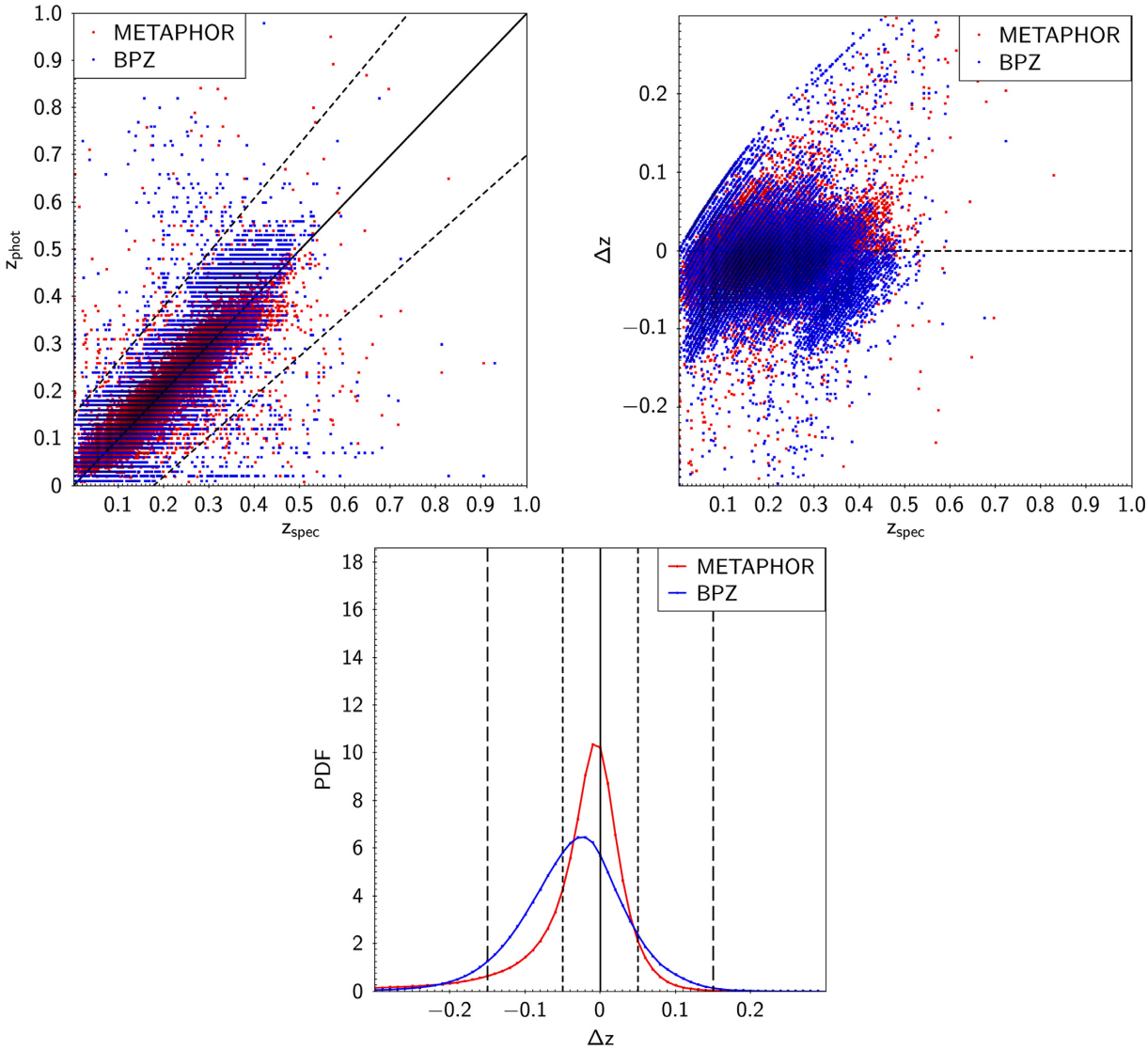

Figure 2. Comparison between METAPHOR (red) and BPZ (blue). Upper row: scatter plot of photometric redshifts as function of spectroscopic redshifts (left-hand panel) and scatter plot of the residuals as function of the spectroscopic redshifts (right-hand panel). Lower row: *stacked* representation of the residuals of PDFs (with redshift bin equal to 0.01).

3.5 Statistical estimators

This section is dedicated to describing the set of statistical estimators adopted to evaluate zphot estimates and relative PDFs performance.

The basic statistics are calculated on the residuals:

$$\Delta z = (z_{\rm spec} - z_{\rm phot})/(1 + z_{\rm spec}). \quad (4)$$

As the individual zphot estimates, in all the presented statistics the following quantities have been considered: the zphot *best-estimates* for METAPHOR and ANNz2 (see, respectively, Sections 3.1 and 3.2); the zphot values *Z_B* provided in the KiDS-DR3 catalogue for BPZ (de Jong et al. 2017); and the photo-z_0 estimates for the *dummy* PDF calculated via METAPHOR (see Section 3.4).

The most common estimators of the zphot accuracy, which we use here, are the standard first four central moments of the residual distribution, respectively, the mean (or bias), standard deviation σ, skewness and kurtosis, the fraction of catastrophic outliers, defined as $|\Delta z| > 0.15$, plus the normalized median absolute deviation (NMAD), defined as

$$NMAD = 1.4826 \times median(|\Delta z - median(\Delta z)|). \quad (5)$$

The cumulative performance of the stacked PDF on the entire sample is evaluated by means of the following three estimators:

(i) $f_{0.05}$: the percentage of residuals Δz within ± 0.05;
(ii) $f_{0.15}$: the percentage of residuals Δz within ± 0.15;
(iii) $\langle \Delta z \rangle$: the average of all the residuals Δz of the stacked PDFs.

Here, by stacked PDFs we mean the individual zphot PDFs transformed into the PDFs of scaled residuals Δz defined in equation (4), and then stacked for the entire sample.

Furthermore, the quality of the individual PDFs is evaluated against the single corresponding zspec from the test set, by defining five categories of occurrences:

(i) *zspecClass* $= 0$: the zspec is within the *bin* (see Section 3.1) containing the peak of the PDF;

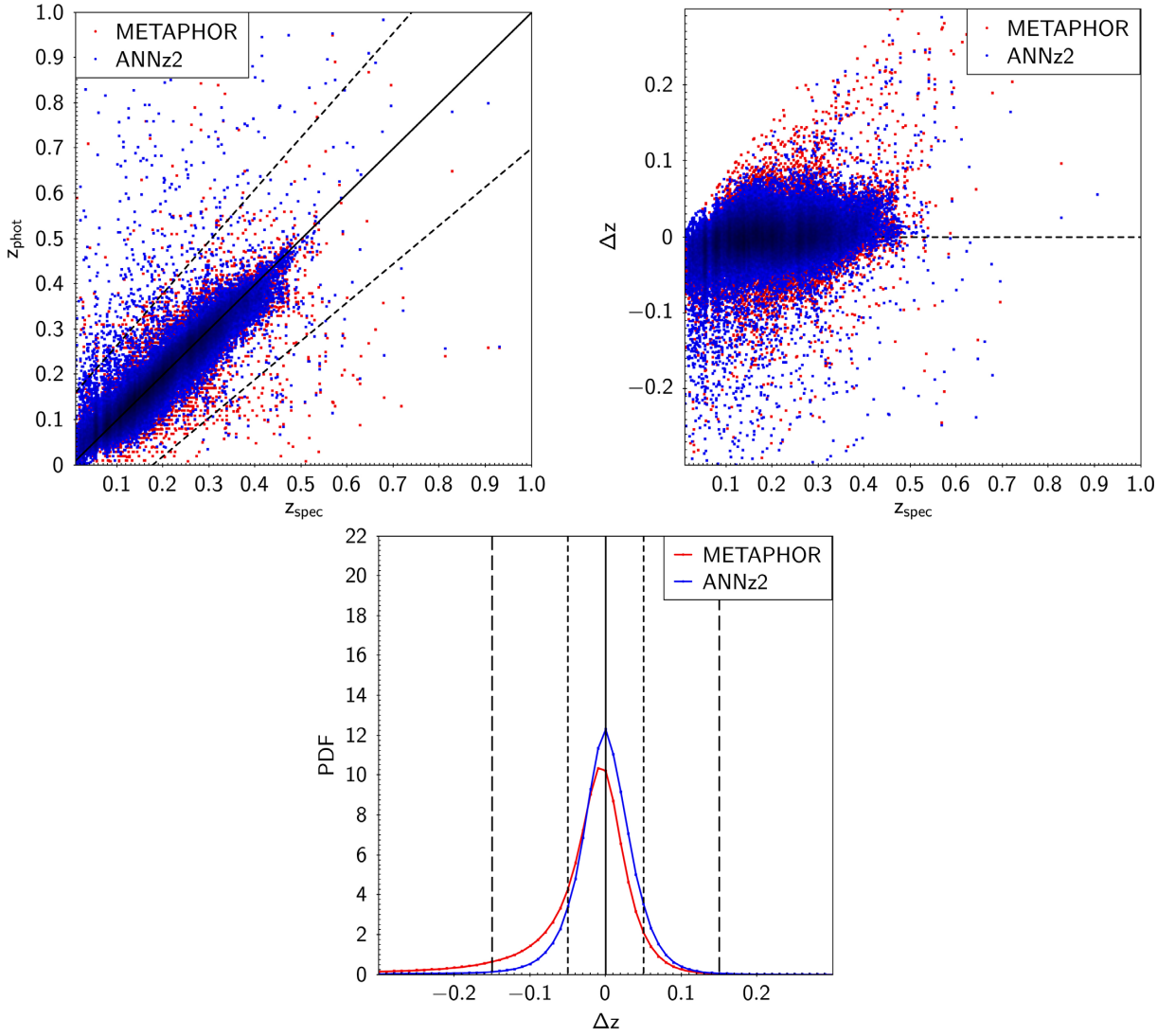

Figure 3. Comparison between METAPHOR (red) and ANNz2 (blue). Upper row: scatter plot of photometric redshifts as function of spectroscopic redshifts (left-hand panel) and scatter plot of the residuals as function of the spectroscopic redshifts (right-hand panel). Lower row: *stacked* representation of the residuals of PDFs (with redshift bin equal to 0.01).

Table 3. Statistics of the zphot error stacked PDFs for METAPHOR, ANNz2, BPZ, and *dummy* obtained by METAPHOR, for the sources cross-matched between KiDS-DR3 photometry and GAMA spectroscopy.

Estimator	METAPHOR	ANNz2	BPZ	dummy
$f_{0.05}$	65.6 %	76.9 %	46.9 %	93.1 %
$f_{0.15}$	91.0 %	97.7 %	92.6 %	99.0 %
$\langle \Delta z \rangle$	− 0.057	0.009	− 0.038	− 0.006

(ii) *zspecClass* = 1: the zspec falls in one bin from the peak of the PDF;

(iii) *zspecClass* = 2: the zspec falls into the PDF, e.g. in a bin in which the PDF is different from zero;

(iv) *zspecClass* = 3: the zspec falls in the first bin outside the limits of the PDF;

(v) *zspecClass* = 4: the zspec falls out of the first bin outside the limits of the PDF.

By definition, the *zspecClass* term depends on the chosen bin amplitude (see Section 3.1), which also determines the accuracy level of PDFs. The quality evaluation of the entire PDF can be hence measured in terms of fractions of occurrences of these five categories within the test data set. In particular, these quantities should be regarded as complementary statistical information, useful to complete the PDF reliability analysis. For example, classes 3 and 4 could quantify the amount of objects falling outside the PDF. The distinction between the two classes gives the supplementary information about how far from the PDFs is their zspec, thus contributing to evaluate their reliability.

Finally, we use two additional diagnostics to analyse the *cumulative* performance of the PDFs: the credibility analysis presented in Wittman, Bhaskar & Tobin (2016) and the probability integral transform (PIT), described in Gneiting, Balabdaoui & Raftery (2007).

The credibility test should assess if PDFs have the correct *width* or, in other words, it is a test of the *confidence* of any method used to calculate the PDFs. In particular, the method is considered overconfident if the produced PDFs are too narrow, i.e. too sharply peaked;

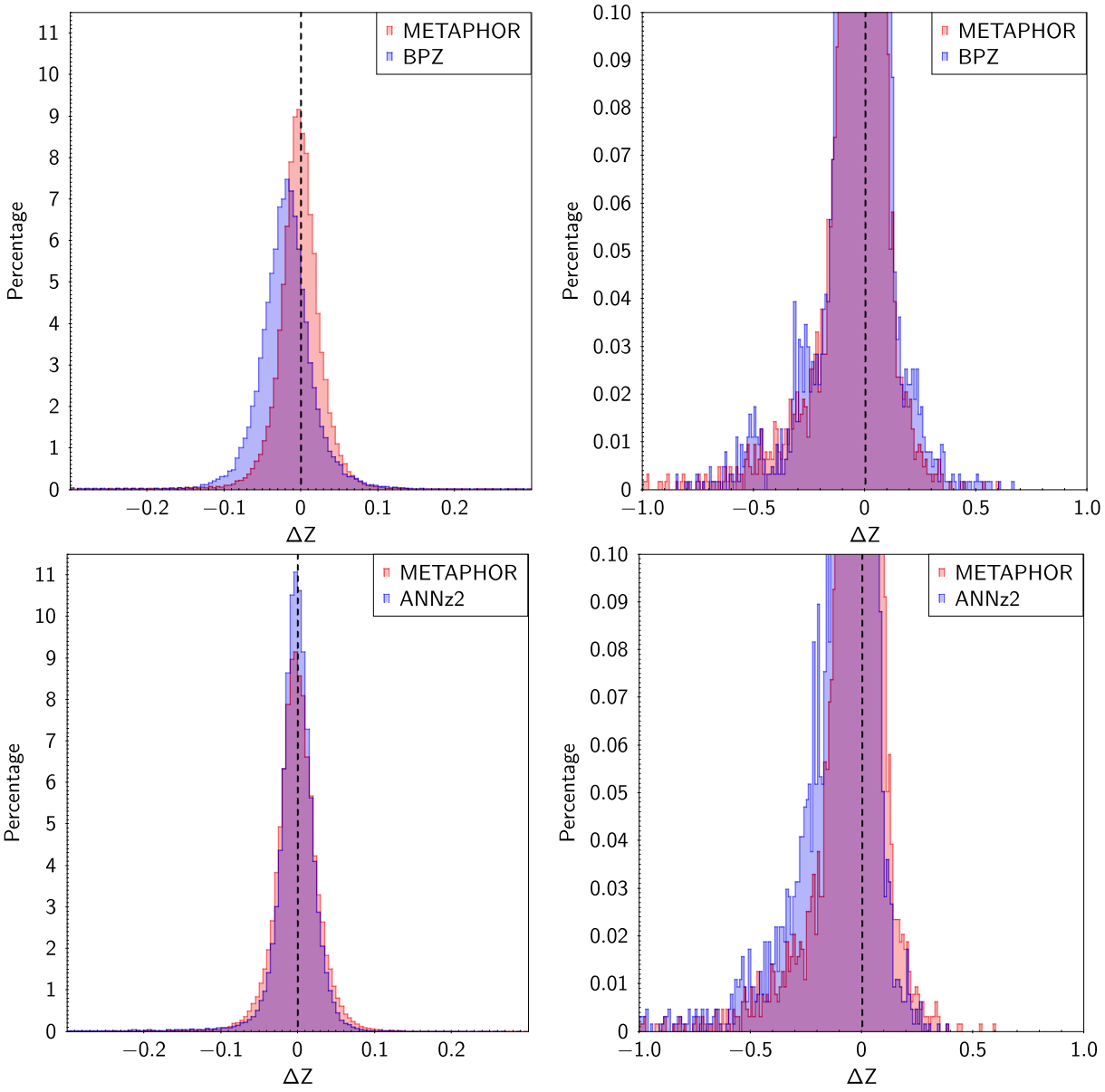

Figure 4. Top panels: comparison between METAPHOR (red) and BPZ (blue); bottom panels: comparison between METAPHOR (red) and ANNz2 (blue). Left-hand panels show the histograms of residual distributions, zoomed in the right-hand panels in order to make more visible the skewness effect. The values are expressed in percentage, after normalizing the distributions to the total number of objects of the blind test set (see Section 2).

Table 4. *zspecClass* fractions for METAPHOR, ANNz2 and BPZ on the GAMA field.

zspecClass	METAPHOR		ANNz2		BPZ	
0	9042	(14.2 %)	12426	(19.4 %)	4889	(7.7 %)
1	16758	(26.3 %)	19040	(29.9 %)	9650	(15.1 %)
2	37233	(58.4 %)	31927	(50.1 %)	49170	(77.15 %)
3	200	(0.3 %)	8	(0.01 %)	0	(0 %)
4	516	(0.8 %)	324	(0.5 %)	31	(0.05)

underconfident otherwise. In order to measure the credibility, rather than the Confidence Intervals (hereafter CI), the highest probability density confidence intervals (HPDCI) are used (Wittman et al. 2016).

The implementation of the credibility method is very straightforward, and it involves the computation of the threshold credibility c_i for the ith galaxy with

$$c_i = \sum_{z \in p_i \geq p_i(z_{\mathrm{spec},i})} p_i(z), \tag{6}$$

where p_i is the normalized PDF for the ith galaxy.

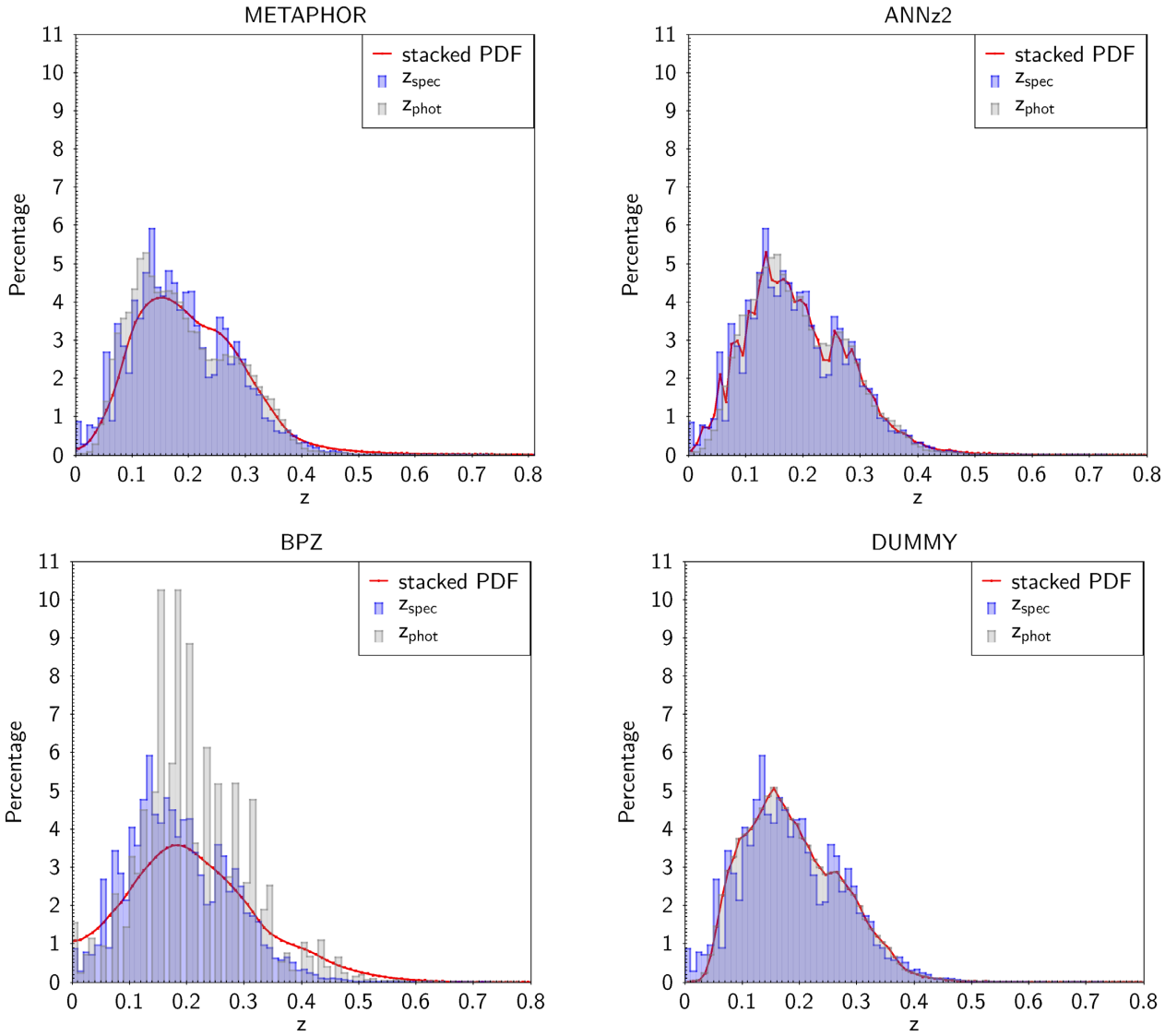

Figure 5. Superposition of the stacked PDF (red) and estimated zphot (grey) distributions obtained by METAPHOR, ANNz2, BPZ, and for the *dummy* (in this last case the zphot distribution corresponds to that of the photo-z_0 estimates, Section 3.4) to the zspec distribution (in blue) of the GAMA field.

The credibility is then tested by calculating the cumulative distribution $F(c)$, which should be equal to c. $F(c)$ resembles a q–q plot, (a typical quantile–quantile plot used for comparing two distributions), in which F is expected to match c, i.e. it follows the bisector in the F and c ranges equal to [0,1]. Therefore, the *overconfidence* corresponds to $F(c)$ falling below the bisector, otherwise the *underconfidence* occurs. In both cases, this method indicates the inaccuracy of the error budget (Wittman et al. 2016).

The PIT histogram measures the predictive capability of a forecast, which is generally probabilistic for continuous or mixed discrete–continuous random variables (Gneiting et al. 2007) and that has been already used to assess the reliability of PDFs in the case of photometric redshifts (see for instance D'Isanto & Polsterer 2018). We can define the PIT as the histogram of the various p_i:

$$p_i = F_i(x_i), \tag{7}$$

where in our case F_i is the CDF of the ith object and $x_i = z_{\mathrm{spec}_i}$. Ideal forecasts produce continuous F_i and PIT with a uniform distribution on the interval (0,1). In other words, we can check the forecast by investigating the uniformity of the PIT: the closer the histogram to the uniform distribution, the better the calibration, i.e. the statistical consistency between the predictive distributions and the validating observations (Baran & Lerch 2016). Nevertheless, it is possible to show that the uniformity of a PIT is a necessary but not sufficient condition for having an ideal forecast (Gneiting et al. 2007).

A strongly U-shaped PIT histogram indicates a highly *underdispersive* character of the predictive distribution (Baran & Lerch 2016).

4 COMPARISON AMONG METHODS

A preliminary comparison among the three methods METAPHOR, ANNz2, and BPZ, only in terms of zphot prediction performance, has been already given in de Jong et al. (2017). That comparison was based on statistics applied to the residuals defined by the equation (4), reported in table 8 and fig. 11 of de Jong et al. (2017). In that figure, the upper panel shows the plots of zphot versus

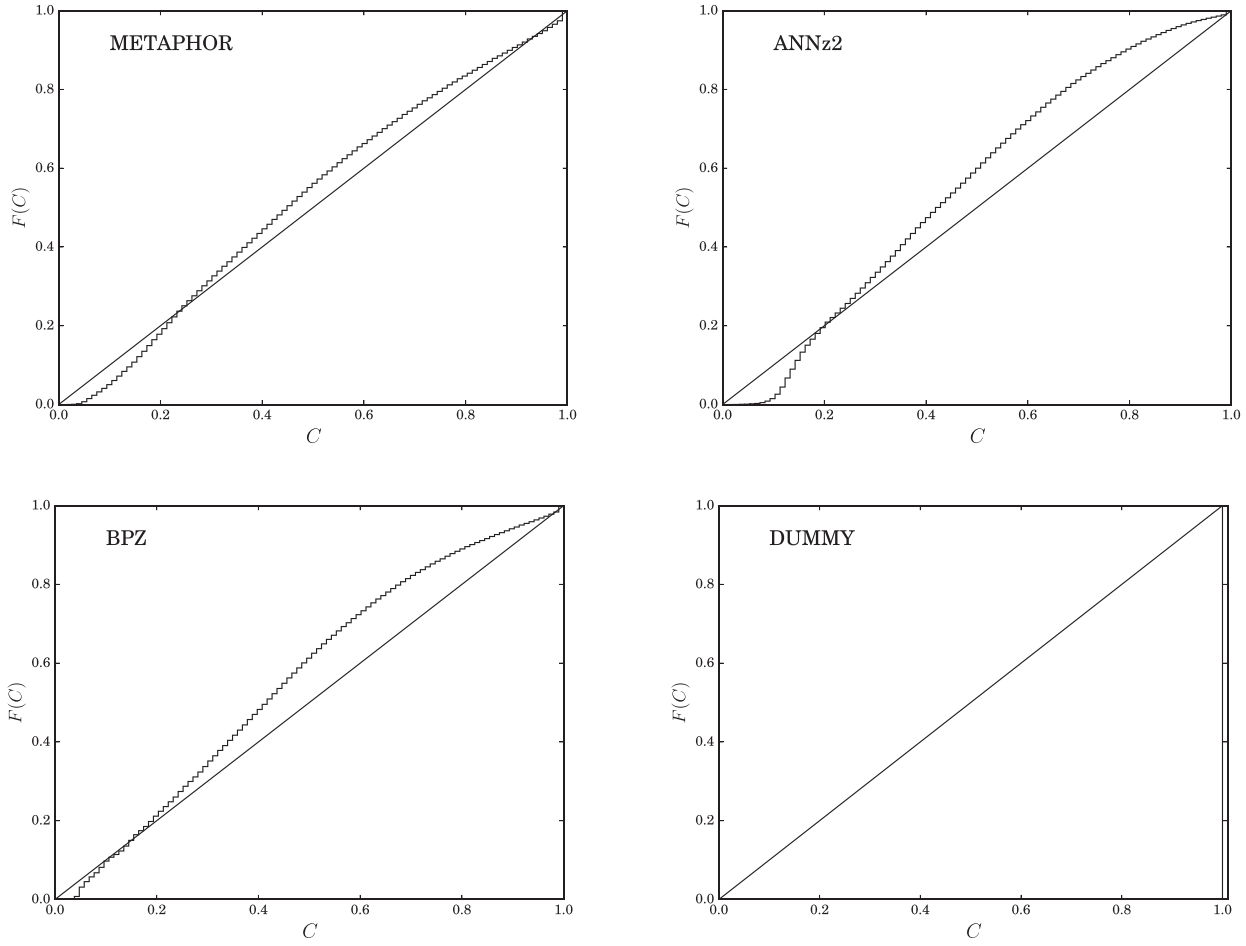

Figure 6. Credibility analysis (see Section 3.5) obtained for METAPHOR, ANNz2, BPZ, and the *dummy* PDF.

GAMA-DR2 spectroscopy, while in the bottom panel residuals versus *r*-magnitude are shown for the three methods.

More recently, in Bilicki et al. (2018) a comparison among the three methods has also been presented on KiDS-DR3 data, more in terms of zphot estimation quality at the full spectroscopic depth available, confirming the better behaviour of ML methods at bright end of KiDS data sample ($z < 0.5$), as well as comparable quality of ML methods and BPZ at higher redshift ($z \sim 1$).

4.1 Statistics on zphot and stacked PDFs

The statistical comparison among the three methods on the data set obtained by cross-matching KiDS-DR3 and GAMA data (see Section 2) summarized in Table 2. It shows a better performance in terms of bias and fraction of outliers for METAPHOR, while BPZ and ANNz2 obtain, respectively, a lower σ and *NMAD* of the errors.

In Figs 2 and 3, we show the comparison on the GAMA field between METAPHOR and, respectively, BPZ and ANNz2, in terms of graphical distributions of predicted zphot and stacked PDFs of the residuals.

From Fig. 2, it is apparent that the correlation between zphot and zspec is tighter for METAPHOR than for BPZ. In terms of stacked PDF, the distributions are in agreement with statistics of Table 3 since the BPZ PDF is more enclosed within the ± 0.15 residual range.

Fig. 3 shows a tighter photospectro redshift correlation for ANNz2 as well as a better symmetry of the stacked PDF.

The effects of kurtosis and skewness are evident from Fig. 4. The kurtosis is a measure of the shape of the residual distribution, particularly suitable for characterizing its tails. From Fig. 4 and Table 2, all three methods show a leptokurtic behaviour. This means that the distributions asymptotically approach zero faster than the Gaussian distribution therefore indicating a small amount of outliers with respect to the Gaussian limit at 2σ ($\sim$0.2 per cent for METAPHOR, $\sim$0.0005 per cent for BPZ, and $\sim$1.5 per cent in the case of ANNz2). This also implies that in this case the standard deviation could be considered a poor estimator for the zphot prediction performance.

The skewness is a measure of the symmetry around zero of the Δz distribution. All the three compared methods show a negative value (see Table 2), mostly due to a longer tail towards negative than to positive Δz. This is more pronounced in the case of METAPHOR and ANNz2 (right-hand panels of Fig. 4), but a negative skewness is expected in zphot residual distributions because of an inherent tendency to overestimate the redshift. By calculating the residuals through equation (4), all methods naturally tend towards negative *zspec−zphot* in the low-redshift regime because negative photometric redshifts are removed (meaningless), introducing the above negative bias in *zspec−zphot*.

In Table 3, we report the fraction of residuals in the two ranges $[-0.05, 0.05]$ and $[-0.15, 0.15]$ and the average of residuals for all

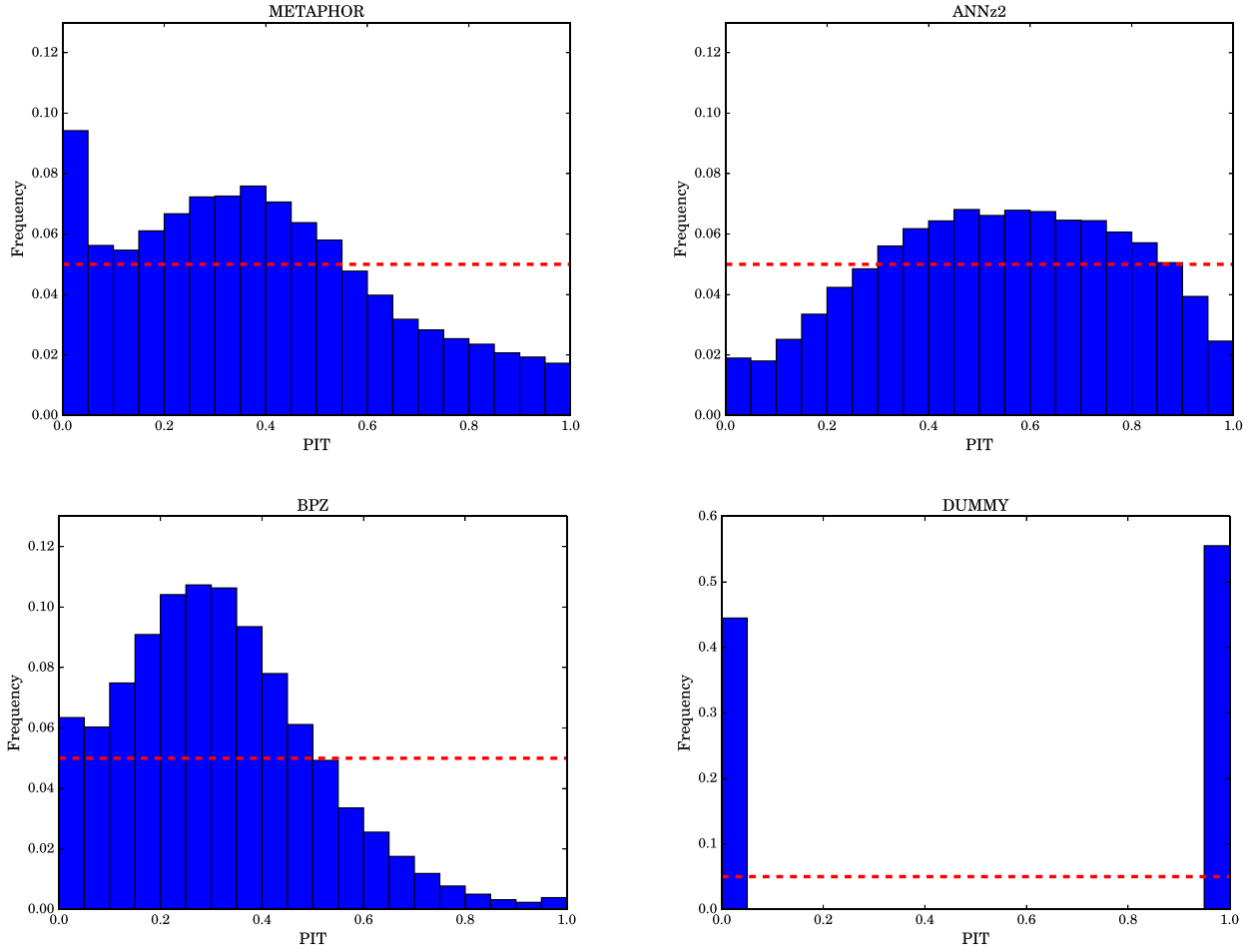

Figure 7. PIT obtained for METAPHOR (top left-hand panel), ANNz2 (top right-hand panel), BPZ (bottom left-hand panel), and for the *dummy* PDF (bottom right-hand panel).

Table 5. Tomographic analysis of the stacked PDFs for METAPHOR, ANNz2, BPZ, and *dummy* PDF calculated by METAPHOR, respectively, in 10 bins of the homogenized magnitude mag_gaap_r.

Bin	r band	Amount	METAPHOR			ANNz2			BPZ			dummy		
			$f_{0.05}$ (%)	$f_{0.15}$ (%)	$\langle\Delta z\rangle$	$f_{0.05}$ (%)	$f_{0.15}$ (%)	$\langle\Delta z\rangle$	$f_{0.05}$ (%)	$f_{0.15}$ (%)	$\langle\Delta z\rangle$	$f_{0.05}$ (%)	$f_{0.15}$ (%)	$\langle\Delta z\rangle$
1	]16.0,16.5]	122	16.3	37.2	−0.330	80.9	99.5	−0.016	26.5	87.0	−0.080	97.5	100	−0.015
2	]16.5,17.0]	290	23.9	49.0	−0.249	81.7	99.2	−0.015	28.5	86.7	−0.080	97.9	99.3	−0.009
3	]17.0,17.5]	858	34.2	62.4	−0.185	82.0	98.4	−0.016	36.4	89.7	−0.068	95.1	98.7	−0.006
4	]17.5,18.0]	1,873	48.0	75.7	−0.132	81.6	97.4	−0.017	41.0	90.7	−0.060	94.2	97.8	−0.010
5	]18.0,18.5]	4,427	59.0	84.6	−0.086	82.2	98.2	−0.011	45.4	92.5	−0.050	95.3	98.7	−0.006
6	]18.5,19.0]	8,230	64.9	89.4	−0.067	81.1	98.0	−0.008	47.6	93.1	−0.043	94.3	98.8	−0.008
7	]19.0,19.5]	15,388	68.9	92.6	−0.051	79.2	97.9	−0.008	48.5	93.2	−0.037	93.7	98.9	−0.007
8	]19.5,20.0]	22,952	68.5	93.8	−0.043	75.9	98.0	−0.006	47.8	92.9	−0.033	93.4	99.2	−0.003
9	]20.0,20.5]	9,178	65.8	94.2	−0.040	61.4	97.0	−0.010	45.4	91.6	−0.033	89.9	98.9	−0.007
10	]20.5,21.0]	367	55.5	88.4	−0.061	44.5	80.4	−0.104	43.1	88.9	−0.033	74.6	94.0	−0.025

the probed methods. The last column shows such statistics also for the *dummy* PDF. Table 4 summarizes the distribution of fractions of samples among the five categories of individual PDFs, obtained by the evaluation of their spectroscopic redshift position with respect to the PDF.

From Table 3, it appears evident that in terms of PDFs, ANNz2 performs quantitatively better than the other two methods, while the *dummy* PDF, derived from METAPHOR, obtains the best estimates. This demonstrates that the statistical estimators adopted for the stacked PDF show low robustness in terms of quality assessment of zphot errors and that there is a need for a deeper understanding of the real meaning of a PDF in the context of zphot quality estimation as well as a careful investigation of the statistical evaluation criteria.

The former statement about ANNz2 performance is also supported by Table 4, where ANNz2 shows a percentage of 49.4 per cent of samples falling within one bin from the PDF peak (the sum of fractions for *zspecClass* 0 and 1) against, respectively, the 40.5 per cent and 22.8 per cent, of the other two methods.

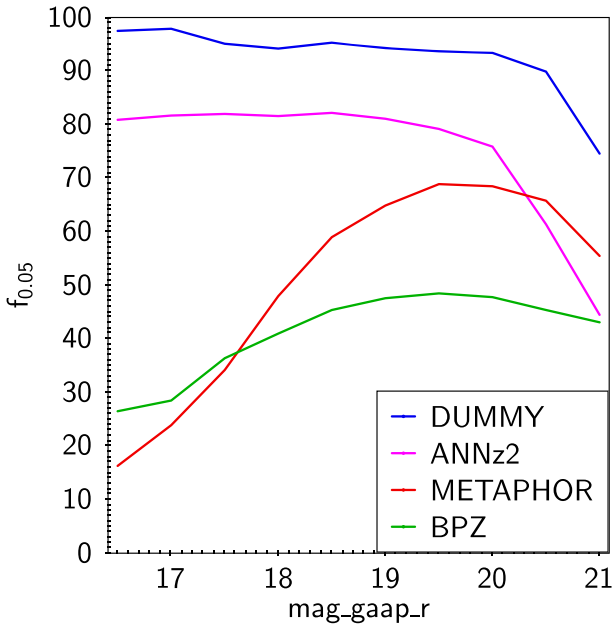

Figure 8. Residuals fraction in the range [−0.05, 0.05] of the PDFs versus magnitude *mag_gaap_r* in the range [16.0, 21.0], used for the tomographic analysis shown in Table. 5 From top to bottom, *dummy* (blue), ANNz2 (violet), METAPHOR (red), and BPZ (green).

However, for all the *stacked* PDF estimators, the *dummy* PDF obtains better statistical results than all other methods. By construction, the *dummy* PDFs are non-zero only at a single value therefore it is not worth to report its statistics regarding the *zspecClass* estimator (see Section 3.5) since, as expected, most of the spectroscopic redshifts fall outside the PDF. Furthermore, the *zspecClass* estimator for the *dummy* PDF is equal to 0 and 4, i.e. the zspec falls either in the bin which corresponds to the PDF peak or outside the PDF. The *dummy* PDF method is then particularly suitable to verify that the residual fractions reported in Table 3 are not sufficient to quantify the performance of a PDF. In Fig. 5, we superimpose the stacked distribution of PDFs, derived by the three methods plus the *dummy* PDF, on the photometric and spectroscopic redshift distributions. The stacked trend of the *dummy* PDF method reproduces the photometric distribution since it does not take into account the redshift error contribution arising from the photometric uncertainties introduced through the perturbation law in equation (1). Very close to the spectroscopic redshift distribution is the stacked PDF of *dummy* and ANNz2, while BPZ and METAPHOR, although still able to follow the spectroscopic distribution, differ from the first two methods. Nevertheless, METAPHOR and ANNz2 PDFs show a better agreement with the individual photometric redshift distributions.

4.2 Credibility analysis and PIT

We also show in a graphical form the two estimators introduced in Section 3.5, namely the credibility analysis on the cumulative PDFs and the PIT. Figs 6 and 7 show these two respective diagrams for the three methods and the *dummy* PDF. The credibility analysis trend of METAPHOR (top left-hand panel of Fig. 6) reveals a higher degree of credibility with respect to ANNz2 and BPZ (respectively, top right-hand and bottom left-hand panels of Fig. 6), the latter being characterized by a higher *underconfidence*. However, the credibility diagram of the *dummy* PDF (bottom right-hand panel of Fig. 6) is identically unitary for each galaxy of the data set. This is evidence of the inability to evaluate the credibility of a zphot error PDF in an objective way. In other words, according to the construction of the HPDCI for the credibility analysis (see Section 3.5), the *dummy* PDF method shows that the 100 per cent of the photo-z_0's fall in the 100 per cent of the HPDCI, thus the predictions are entirely *overconfident*.

The statistical evaluation of the three methods and the *dummy* PDF based on the PIT diagram is shown in Fig. 7. We observe a better behaviour of ANNz2 (top right-hand panel) than the two other methods, METAPHOR (top left-hand panel) and BPZ (bottom left-hand panel). For ANNz2, the *overdispersive* and *underdispersive* trends appear less pronounced than for the other cases, especially BPZ. However, the PIT histogram for *dummy* PDFs shows an entirely degraded (i.e. *underdispersive*) behaviour of the zphot distribution (bottom right-hand panel of Fig. 7). This result was expected since by definition its CDF is a step function, thus allowing only values 0 or 1, corresponding to the two bars in Fig. 7. This is in some contradiction to the previous statistics, shown for the quantitative estimators for the *dummy* PDFs (Tables 3 and 5), which were indicating the best behaviour for the dummy stacked PDF.

4.3 PDF tomography

Finally, in order to analyse the stacked PDFs obtained by the four estimation methods in different ranges of magnitude, we performed a binning in *mag_gaap_r* in the range [16.0, 21.0] with a step $\Delta mag = 0.5$, resulting in a tomography of 10 bins. The range has been chosen in order to ensure a minimum amount of objects per bin to calculate the statistics. The results in terms of the fraction of residuals and the overall average for the stacked PDFs are reported in Table 5, while the fraction of residuals $f_{0.05}$ is shown as a function of *r*-band magnitude in Fig. 8.

Given the statistics in Tables 5 and 6, we observe that for BPZ the zspec falls within the PDF in practically all bins and that the highest concentration of PDFs is within $f_{0.15}$. This behaviour indicates a broad shape of the PDFs, also confirmed by the *underconfidence* shown in Fig. 6 and by the *overdispersion* in Fig. 7. The latter figure also shows the presence of a high bias, visible from the unbalanced trend. Furthermore, the PIT tomography, reported in Figs 9 and 10, shows a high variability and confirms the general *overdispersion* and bias of the PDFs. Turning to the HPDCI tomography, the overall trend of Fig. 6 indicates a general *underconfidence*, but Figs 11 and 12 show an inversion, from a high *underconfidence* to a lower *overconfidence*, compatible with the general variability of BPZ PDFs.

ANNz2 shows a similar behaviour as BPZ for $f_{0.15}$ but has a higher percentage of $f_{0.05}$, which indicates PDFs more centred around zspec. Furthermore, the values of *zspecClass* equal to 3 and 4 show that zspec mostly falls within the PDFs, thus indicating that also in the case of ANNz2 the PDFs have a broad shape, albeit to a lesser extent. This is also confirmed by the *overdispersive* trend of the PIT diagram in Fig. 7 as well as by the *underconfidence* in Fig. 6. In terms of PIT and HPDCI tomography, ANNz2 shows a more regular behaviour than BPZ.

METAPHOR shows a stacked PDF with a more pronounced average $\langle \Delta z \rangle$ than BPZ and ANNz2 in Table 5, due to the larger tails of its Δz distribution. Moreover, both Tables 5 and 6 indicate that the METAPHOR and especially the BPZ PDFs have a broader shape than those of ANNz2, for example by looking at the percentages for *zspecClass* = 2. This is also reflected by the PIT diagram of Fig. 7, which reveals highly biased and *overdispersive* PDFs. In contrast with previous statistics, the HPDCI diagram indicates that

Table 6. *zspecClass* fractions for METAPHOR, ANNz2, and BPZ in tomographic bins of the homogenized magnitude *mag_gaap_r*.

Bin	*r*-band	Amount	METAPHOR (%)					ANNz2 (%)					BPZ (%)				
			0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
1	]16.0,16.5]	122	4.9	13.9	80.3	0.8	0.0	38.5	29.5	32.0	0.0	0.0	0.0	1.6	98.4	0.0	0.0
2	]16.5,17.0]	290	10.3	22.7	66.5	0.3	0.0	38.4	22.1	39.4	0.0	0.0	0.7	1.7	97.6	0.0	0.0
3	]17.0,17.5]	858	19.0	32.3	47.8	0.6	0.3	31.3	27.9	40.5	0.0	0.3	1.2	3.4	95.3	0.0	0.1
4	]17.5,18.0]	1,873	18.7	33.4	46.1	0.6	1.2	29.6	31.0	38.9	0.0	0.5	1.9	6.0	92.0	0.0	0.05
5	]18.0,18.5]	4,427	16.7	31.0	50.3	0.5	0.8	24.9	35.2	39.2	0.0	0.6	4.0	10.0	86.0	0.0	0.09
6	]18.5,19.0]	8,230	18.6	30.9	49.2	0.2	1.0	23.3	33.7	42.3	0.0	0.7	6.5	14.7	78.7	0.0	0.1
7	]19.0,19.5]	15,388	14.7	27.6	56.7	0.2	0.8	19.5	31.3	48.6	0.02	0.6	8.4	16.3	75.2	0.0	0.07
8	]19.5,20.0]	22,952	12.7	24.3	66.5	0.4	0.9	17.3	28.6	53.7	0.01	0.4	9.0	16.9	74.2	0.0	0.03
9	]20.0,20.5]	9,178	10.7	21.2	66.5	0.4	0.9	15.4	25.5	58.6	0.02	0.4	8.4	15.3	76.3	0.0	0.0
10	]20.5,21.0]	367	10.3	16.3	70.3	0.8	2.2	10.9	17.4	70.3	0.0	1.4	9.0	12.8	77.9	0.0	0.0

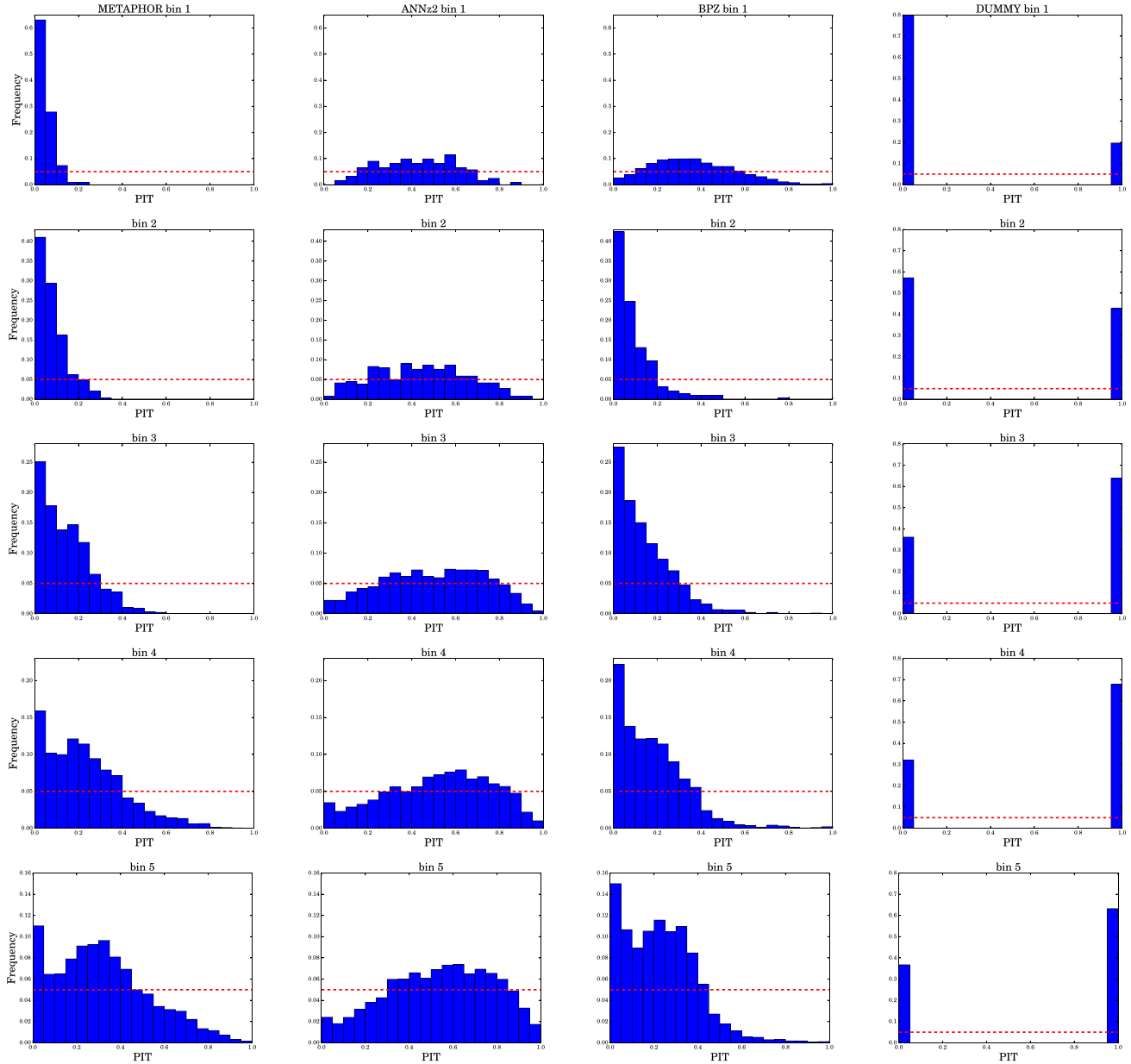

Figure 9. PIT obtained for METAPHOR (first column panels), ANNz2 (second column panels), BPZ (third column panels), and for the *dummy* PDF (fourth column panels) in the first five magnitude tomographic bins from Table 5.

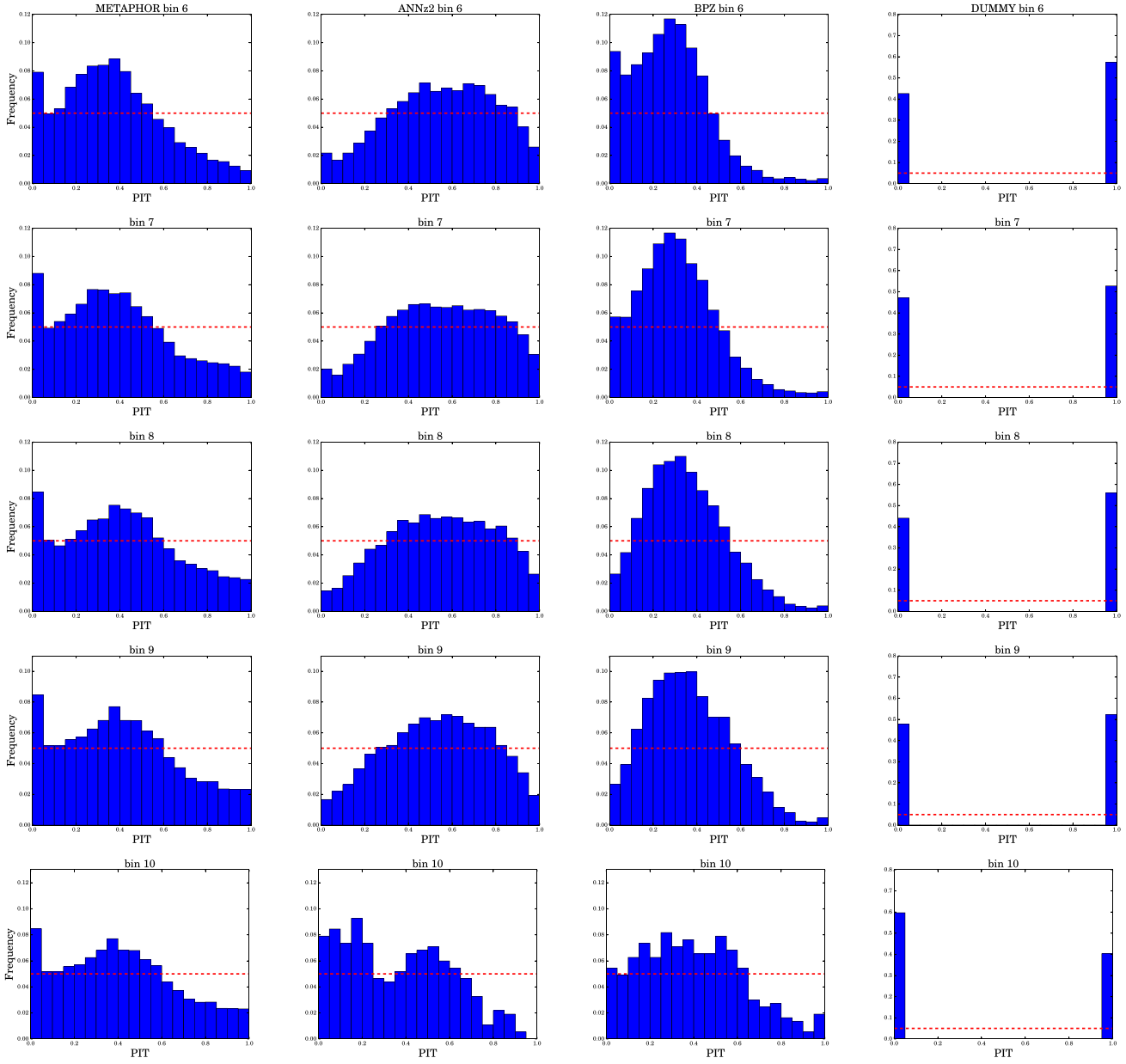

Figure 10. PIT obtained for METAPHOR (first column panels), ANNz2 (second column panels), BPZ (third column panels), and for the *dummy* PDF, calculated by METAPHOR (fourth column panels) in the second five magnitude tomographic bins from Table 5.

METAPHOR is less *underconfident* than BPZ and ANNz2. The tomographic analysis of the PIT diagram reports highly biased PDFs for the bins of Fig. 9, while in the other bins of Fig. 10 METAPHOR shows similar characteristics as ANNz2, albeit with different types of defects. Finally, in terms of HPDCI tomography, Figs 11 and 12 reveal a general coherence in the behaviour of METAPHOR, except in the first and last bin, which are least populated.

5 CONCLUSIONS

Due to the increasing demand for reliable zphot and the intrinsic difficulty to provide reliable error PDF estimation for ML methods, a plethora of solutions has been proposed. The derivation of PDFs with ML models is in fact conditioned by the mechanism used to infer the hidden flux–redshift relationship. In fact, this mechanism imposes the necessity to disentangle the contributions to the zphot estimation error budget, by distinguishing the intrinsic method error from the photometric uncertainties. Furthermore, due to the large variety of methods proposed, there is also the problem of finding objective and robust statistical estimators of the quality and reliability of the derived PDFs.

We believe that it is extremely useful to estimate the zphot error through the intrinsic photometric uncertainties, by considering that the observable photometry cannot be perfectly mapped to the true redshift. Furthermore, the evaluation of a statistically meaningful PDF should consider the effective contribution of the intrinsic error of the method.

In Cavuoti et al. (2017a), we presented METAPHOR, a method designed to provide a PDF of photometric redshifts calculated by ML methods. METAPHOR has already been successfully tested on

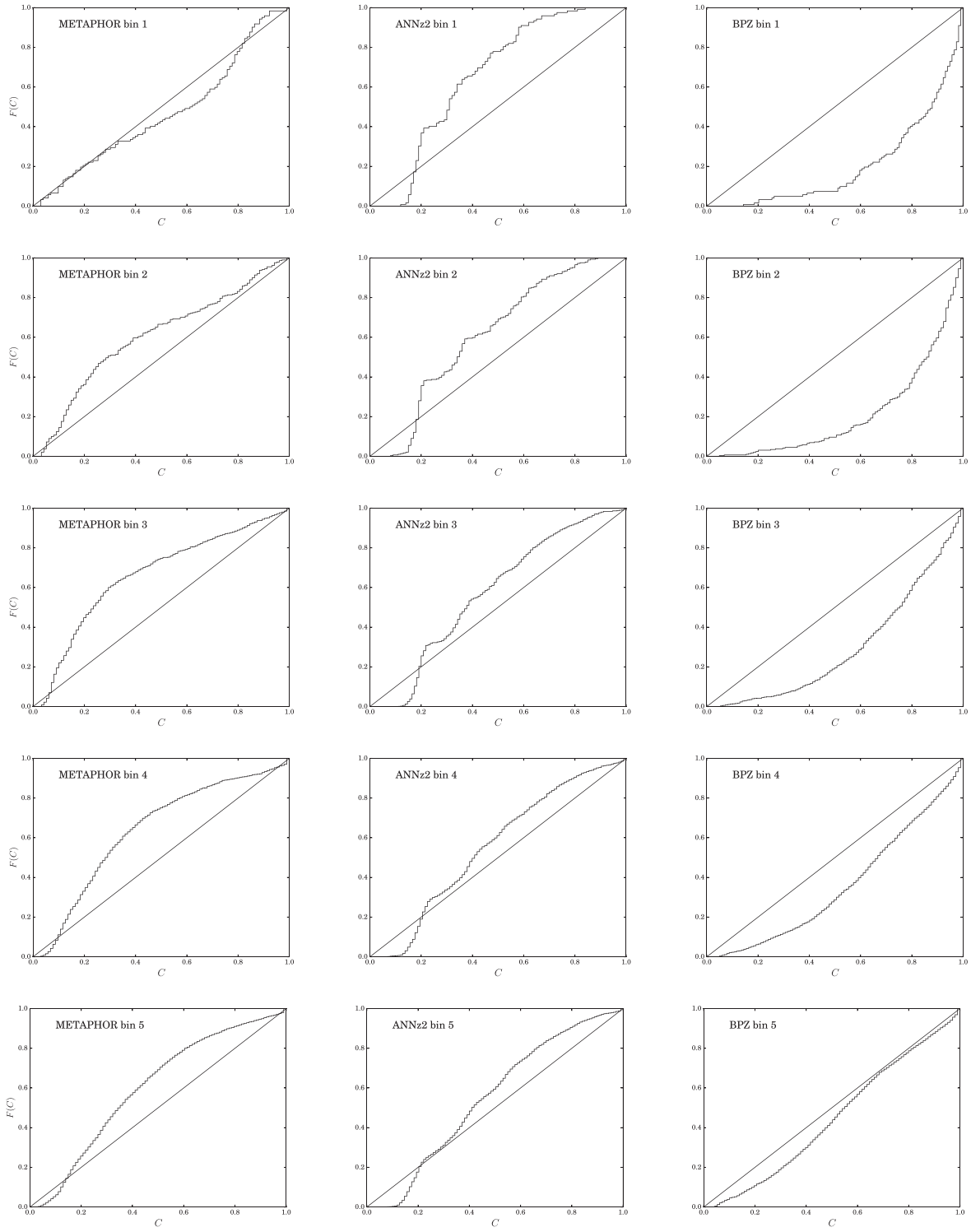

Figure 11. Credibility analysis (see Section 3.5) obtained for METAPHOR, ANNz2, and BPZ for the first five magnitude tomographic bins from Table 5. The credibility plots for the *dummy* PDF are the same as the bottom right-hand panel of Fig. 6 in all the bins.

SDSS (Cavuoti et al. 2017a) and KiDS-DR3 (de Jong et al. 2017) data, and uses the neural network MLPQNA (Brescia et al. 2013, 2014a) as the internal zphot estimation engine.

Main goal of this work is a deeper analysis of zphot PDFs obtained by different methods: two ML models (METAPHOR and ANNz2) and one based on SED fitting techniques (BPZ), through a direct comparison among such methods. The investigation was focused on both cumulative (*stacked*) and individual PDF reliability. Moreover, the methods were subjected to a comparative analysis using different kinds of statistical estimators to evaluate their degree of coherence. Exactly for this reason, by modifying the METAPHOR internal mechanism, we also derived a *dummy* PDF method (see

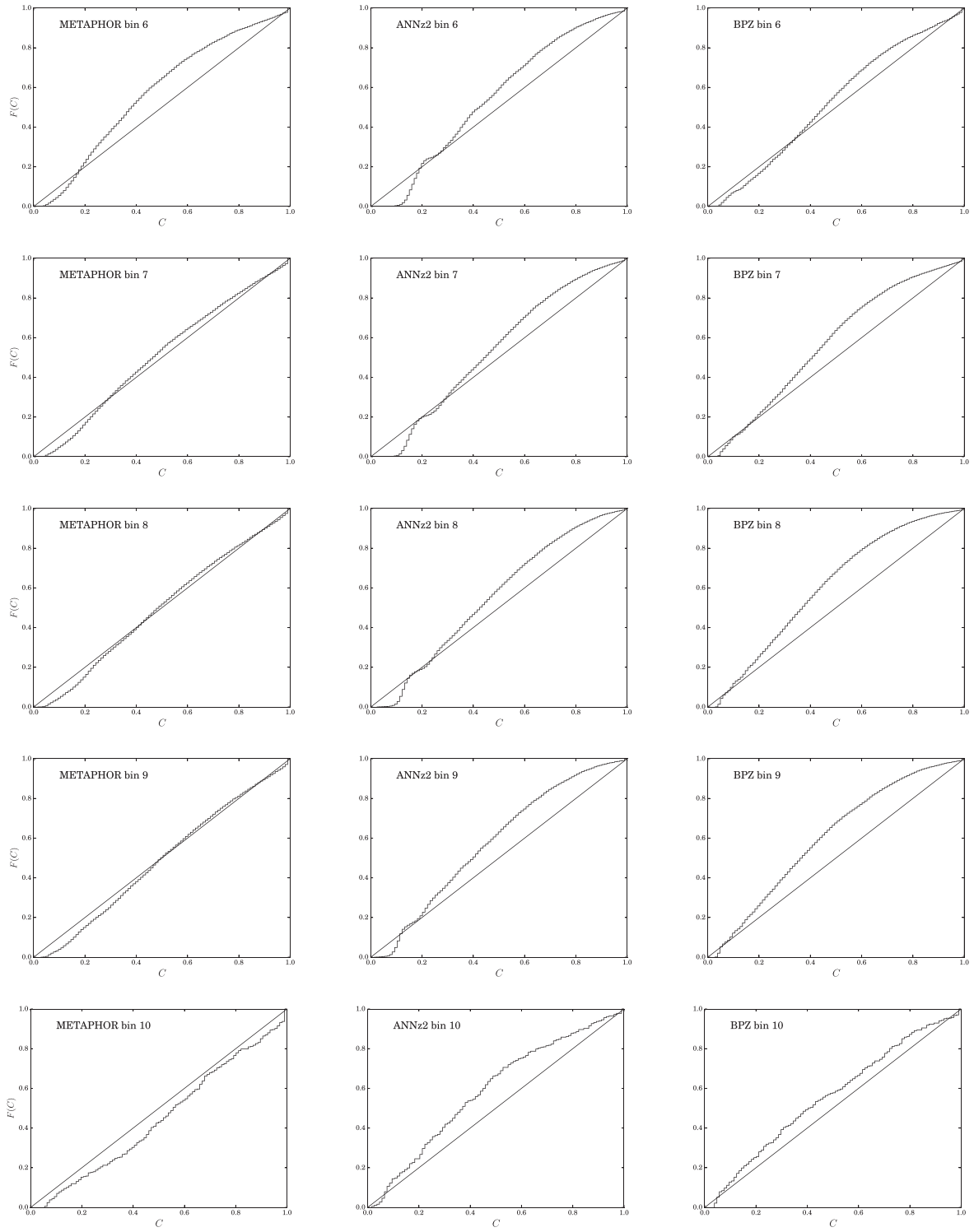

Figure 12. Credibility analysis (see Section 3.5) obtained for METAPHOR, ANNz2 and BPZ for the second five magnitude tomographic bins from Table 5. The credibility plots for the *dummy* PDF are the same as the bottom right-hand panel of Fig. 6 in all the bins.

Section 3.4), helpful to obtain a benchmark tool to evaluate the objectivity of the various statistical estimators applied on the presented methods.

Regarding the *dummy* PDF (Table 3), the more the PDF is representative of an almost perfect mapping of the parameter space on the true redshifts, the better are the performances in terms of stacked PDF estimators. However, we have shown that the PIT histogram and the credibility analysis provide important complementary statistical information, the first showing the total *underdispersive* trend of the reconstructed photometric redshift distribution; the second

reporting an *overconfidence* of all zphot estimates. Both the *underdispersion* and the *overconfidence* are related to the narrowness of the PDFs: the narrower they are, the more the PIT histogram is *underdispersed* and the results, as determined by the credibility analysis, overconfident.

Thus, it appears clear that the statistical estimators used for the stacked PDF (for instance $f_{0.05}$, $f_{0.15}$, and $\langle \Delta z \rangle$), are not self-consistent and should be combined with other statistical estimators, such as the PIT diagrams and credibility analysis.

Although the credibility analyses of the different methods, based on the Wittman diagram (Fig. 6) and the PIT diagram (Fig. 7), appear comparable in terms of overall results, their tomography (Figs 9, 10, 11, and 12) shows different behaviours at different redshift regimes.

Summarizing the results for the three PDF estimation methods analysed, considering the combination of statistical estimators, ANNz2 is favoured by the $f_{0.05}$, $f_{0.15}$, $\langle \Delta z \rangle$, and the PIT diagram. However, METAPHOR is more competitive, in particular when considering the confidence analysis. BPZ has the best PDFs in the faintest magnitude bin. Moreover, all three methods show a generally broad shape of their PDFs, albeit to a different extent, with also a bias in the case of BPZ and METAPHOR. However, they show occasional fluctuations in their tomographic analysis. For instance, BPZ reverses its *overconfidence* trend at fainter magnitudes, while METAPHOR and BPZ show a high level of variability along the magnitude bins in terms of *underdispersion* and bias. In the specific case of our method METAPHOR, all mentioned defects require further investigation in terms of the photometric perturbation function.

It should be noted that the current comparison is preliminary since the methods explored in this paper deal with different sources of errors. In fact, ANNz2 takes into account only the internal errors of the method, METAPHOR only those induced by the photometry, BPZ includes both these error sources and, finally, the benchmark (*dummy* PDF) does not include either of these two.

All considerations together lead us to affirm that a detailed analysis of the performances, based on a combination of independent statistical estimators, is key to unraveling the nature of the estimated zphot PDFs and to assess the objective validity of the method employed to derive them.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for all the comments and suggestions that improved the manuscript. Based on data products from observations made with European Southern Observatory (ESO) telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017, and 177.A-3018, and on data products produced by Target/OmegaCEN, Istituto Nazionale di AstroFisica (INAF)-Osservatorio Astronomico di Capodimonte Napoli (OACN), INAF-Osservatorio Astronomico di Padova (OAPD) and the KiDS production team, on behalf of the KiDS consortium. OmegaCEN and the KiDS production team acknowledge support by NOVA and NWO-M grants. MBr acknowledges financial contribution from the agreement ASI/INAF I/023/12/1. MBr acknowledges also the *INAF PRIN-SKA 2017 program 1.05.01.88.04* and the funding from *MIUR Premiale 2016: MITIC*. MBi is supported by the Netherlands Organization for Scientific Research, NWO, through grant number 614.001.451. GL and MBr acknowledge partial support from the H2020 Marie Curie ITN - SUNDIAL. CT is supported through an NWO-VICI grant (project number 639.043.308). SC acknowledges support from the project 'Quasars at high redshift: physics and Cosmology' financed by the ASI/INAF agreement 2017-14-H.0. JTAdJ is supported by the Netherlands Organisation for Scientific Research (NWO) through grant 621.016.402. DAMEWARE has been used for this work (Brescia et al. 2014a).

REFERENCES

Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, MNRAS, 417, 1891
Ahn C. P. et al., 2014, ApJS, 211, 17
Annis J. T., 2013, AAS Meeting, 221, 335.05
Annunziatella M. et al., 2016, A&A, 585, A160
Aragon-Calvo M. A., van de Weygaert R., Jones B. 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
Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12
Baran S., Lerch S., 2016, Mixture EMOS model for calibrating ensemble forecasts of wind speed. Environmetrics, 27, 116
Baum W. A., 1962, in McVittie G. C., ed., Proc. IAU Symp. 15, Problems of Extra-Galactic Research. Macmillan Press, New York, p. 390
Benítez N., 2000, ApJ, 536, 571
Bilicki M. et al., 2018, A&A, 616, A69
Bolzonella M., Miralles J. M., Pelló R., 2000, A&A, 363, 476
Bonnet C., 2013, MNRAS, 449, 1043
Bordoloi R., Lilly S. J., Amara A., 2010, MNRAS, 406, 881
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
Capak P. L., 2004, PhD thesis, Univ. Hawai'i
Capozzi D., de Filippis E., Paolillo M., D'Abrusco R., Longo G., 2009, MNRAS, 396, 900
Carrasco K., Brunner R. J., 2013a, MNRAS, 432, 1483
Carrasco K., Brunner R. J., 2013b, ASP Conf. Ser., Astronomical Data Analysis Software and Systems XXII. Vol. 475. 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., 2012, A&A, 546, 13
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., 2017a, MNRAS, 465, 1959
Cavuoti S. et al., 2017b, MNRAS, 466, 2039
Colless M. et al., 2001, MNRAS, 328, 1039
Collister A. A., Lahav O., 2004, PASP, 116, 345
Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655
D'Isanto A., Polsterer K. L., 2018, A&A, 609, A111
de Jong J. T. A. et al., 2015, A&A, 582, A62
de Jong J. T. A. et al., 2017, A&A, 604, A134
Driver S. P. et al., 2011, MNRAS, 413, 971
Duncan K. J., Jarvis M. J., Brown M. J. I., Röttgering H. J. A., 2017, MNRAS, 477, 5177
Euclid Red Book, 2011, ESA Technical Document, ESA/SRE(2011)12, Issue 1.1. preprint (arXiv:1110.3193)
Firth A. E., Lahav O., Somerville R. S., 2003, MNRAS, 339, 1195
Fu L. et al., 2018, MNRAS, 479, 3558
Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823
Gneiting T., Balabdaoui F., Raftery A. E., 2007, Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc.: Series B (Statistical Methodology), 69, 243

Gorecki A., Abate A., Ansari R., Barrau A., Baumont S., Moniez M., Ricol J. S., 2014, A&A, 561, A128
Graff A., Feroz F., Hobson M. P., Lasenby A., 2014, MNRAS, 441, 1741
Hildebrandt H. et al., 2010, A&A, 523, A31
Hildebrandt H. et al., 2012, MNRAS, 421, 2355
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Hoyle B., Rau M. M., 2018, preprint (arXiv:1802.02581)
Ilbert O. et al., 2006, A&A, 457, 841
Ivezic Z., 2009, American Physical Society, APS April Meeting, May 2-5, W4.003
Kuijken K., 2008, A&A, 482, 1053
Laureijs R. et al., 2014, in Proc. SPIE Conf. Ser. Vol. 9143. Space Telescopes and Instrumentation 2014: Optical, Infrared and Millimeter Wave. SPIE, Bellingham, p. 91430H
Liske J. et al., 2015, MNRAS, 452, 2087
Mandelbaum R. et al., 2008, MNRAS, 386, 781
Masters D. et al., 2015, ApJ, 813, 53
Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008, ApJ, 689, 709
Radovich M. et al., 2017, A&A, 598, A107
Ross A. J. et al., 2017, MNRAS, 472, 4456
Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502
Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502
Sánchez C. et al., 2014, MNRAS, 445, 1482
Science Collaborations and LSST Project, 2009, LSST Science Book, Version 2.0, preprint (arxiv:0912.0201)
Serjeant S., 2014, ApJ, 793, L10
Soo J. Y. H. et al., 2017, MNRAS, submitted
Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, Lecture Notes in Computer Science, Vol. 2859, Neural Networks for Photometric Redshifts Evaluation. Springer-Verlag , Berlin, p. 226
Tanaka M., 2015, ApJ, 801, 20
Tortora C. et al., 2016, MNRAS, 457, 2845
Viola M. et al., 2015, MNRAS, 452, 3529
Wittman D., Bhaskar R., Tobin R., 2016, MNRAS, 457, 4005

APPENDIX A: ANALYSIS OF METAPHOR ERROR SOURCES

In this appendix, we investigated the possibility to quantify the contribution of the method error to the zphot estimation. For instance, such error, in the case of METAPHOR, mostly depends on the random initialization of the neural connection weights in the MLPQNA neural network, used as internal engine to determine the zphot point estimates.

Through a test performed on the SDSS DR9 data (Cavuoti et al. 2017a), we already showed that N different trainings did not degrade the PDF performance: de facto the error introduced by the method appears negligible. On the other hand, N network trainings are very time consuming. Here, we deepened this exercise with METAPHOR pipeline for the KiDS-DR3 data, by performing two different experiments described next.

We created 100 training samples, namely 100 random extractions from the training set used to obtain the KiDS-DR3 PDFs (see Section 2). Each of the 100 training sets contains 10 000 objects. The experiments are the following:

(1) Experiment (i): 100 training + test executions by keeping unchanged both training and test sets. The single training set has been randomly selected from among the 100 sets available and the test set corresponds to the sample obtained by cross-matching the KiDS-DR3 photometry with GAMA DR2 spectroscopy (see Section 2);

Table A1. Statistics of the zphot error stacked PDFs obtained by METAPHOR, for the experiments (i) and (ii).

Estimator	exp (i)	exp (ii)
$f_{0.05}$	92.2 %	92.1 %
$f_{0.15}$	98.4 %	98.4 %
$\langle \Delta z \rangle$	− 0.008	− 0.008

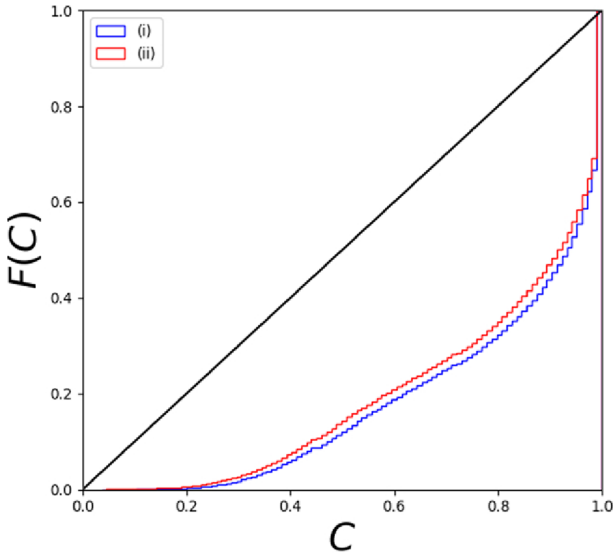

Figure A1. Credibility analysis obtained for the experiments (i) (blue) and *(ii)* (red).

(2) Experiment (ii): 100 training + test executions by varying the training set each time and by keeping unchanged the test set (same set as previous experiment).

In both experiments, all other set-up parameters of the full METAPHOR pipeline have been left unchanged. Therefore, the difference between the two experiments is only the training set-up of the internal engine MLPQNA (weights initialization is left random and photometry is fixed at each training of the experiment (i), while weights initialization is left random and training photometry is variable in the experiment (ii). In other words, in the experiment (i) we isolated the effect of the random weights initialization, while in the second experiment we kept the sum of the two effects (weights initialization and variable training photometry). Both experiments lasted ∼11 on a 8-core pentium *i*7.

In Table A1, we report the stacked PDF statistics for the two experiments, while the credibility and PIT analyses are shown in Figs A1 and A2, respectively. The results, as expected, are comparable to those obtained for the *dummy* PDF (see Table 3) since in both cases we did not introduce any photometry perturbation. The degradation of the stacked PDF performance (compared to that of the *dummy* PDF) is of the order of 1 per cent and 0.6 per cent for the residual fractions, respectively, $f_{0.05}$ and $f_{0.15}$, while 0.002 for $\langle \Delta z \rangle$. Also, the comparison with the credibility and PIT diagrams of the *dummy* PDF (right-bottom diagrams of Figs 6 and 7, respectively) reveals only the small differences induced by the 100 trainings in experiments (i) and (ii), instead of the single training in the *dummy* case. Such statistical variations can be considered negligible if compared to those obtained by the photometry perturbation of the test

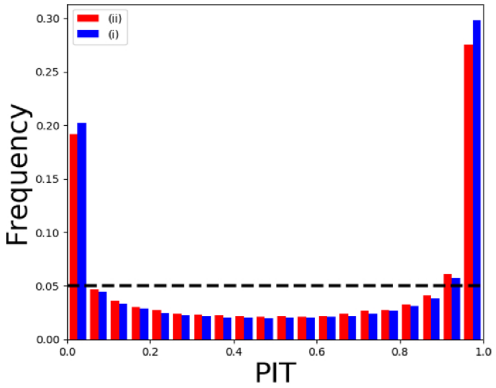

Figure A2. PIT obtained for the experiments (i) (blue) and *(ii)* (red).

set (see Section 4 and Table 3), where instead the computational cost for experiments (i) and (ii) becomes prohibitive (increasing computing time by ∼70 per cent).

A comparison between experiments (i) and (ii) in terms of credibility and PIT diagrams shows very similar results. In particular, by overlapping the two kinds of diagrams (Figs A1 and A2), it appears evident that the sum of contributions of the variable photometry within the training set plus the random weights initialization differ very little from the case in which the photometry is kept unchanged.

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