A&A 608, A39 (2017)
DOI: 10.1051/0004-6361/201731762

Astronomy & Astrophysics

Estimating photometric redshifts for X-ray sources in the X-ATLAS field using machine-learning techniques⋆

G. Mountrichas[1], A. Corral[2,1], V. A. Masoura[1,3], I. Georgantopoulos[1], A. Ruiz[1], A. Georgakakis[1], F. J. Carrera[2], and S. Fotopoulou[4]

[1] National Observatory of Athens, V. Paulou & I. Metaxa, Athens 11532, Greece
e-mail: gmountrichas@gmail.com
[2] Instituto de Fisica de Cantabria (CSIC-Universidad de Cantabria), 39005 Santander, Spain
[3] Section of Astrophysics, Astronomy and Mechanics, Department of Physics, Aristotle University of Thessaloniki, 54 124 Thessaloniki, Greece
[4] Department of Astronomy, University of Geneva, ch. d'Ecogia 16, 1290 Versoix, Switzerland

Received 11 August 2017 / Accepted 3 October 2017

ABSTRACT

We present photometric redshifts for 1031 X-ray sources in the X-ATLAS field using the machine-learning technique TPZ. X-ATLAS covers 7.1 deg^2 observed with *XMM-Newton* within the Science Demonstration Phase of the H-ATLAS field, making it one of the largest contiguous areas of the sky with both *XMM-Newton* and *Herschel* coverage. All of the sources have available SDSS photometry, while 810 additionally have mid-IR and/or near-IR photometry. A spectroscopic sample of 5157 sources primarily in the XMM/XXL field, but also from several X-ray surveys and the SDSS DR13 redshift catalogue, was used to train the algorithm. Our analysis reveals that the algorithm performs best when the sources are split, based on their optical morphology, into point-like and extended sources. Optical photometry alone is not enough to estimate accurate photometric redshifts, but the results greatly improve when at least mid-IR photometry is added in the training process. In particular, our measurements show that the estimated photometric redshifts for the X-ray sources of the training sample have a normalized absolute median deviation, nmad ≈ 0.06, and a percentage of outliers, $\eta = 10$–14%, depending upon whether the sources are extended or point like. Our final catalogue contains photometric redshifts for 933 out of the 1031 X-ray sources with a median redshift of 0.9.

Key words. X-rays: general – galaxies: active – catalogs – techniques: photometric

1. Introduction

Current and future surveys (e.g. XMM, eROSITA, DES, and Euclid) will provide us with large datasets that contain hundreds of thousands of sources. Spectroscopy is expensive in telescope time and challenging to complete for large samples, thus photometric redshift (photo-z) estimations have become a necessity in observational astronomy today. Although photo-z estimations are cheaper and the only means to estimate distances for large samples, they are also subject to systematics and higher uncertainties than spectroscopic redshift estimations (spec-z).

The pursuit of accurate photometric redshifts has led to the development of many photo-z estimation methods that can be divided into two main categories: template-fitting (e.g. Brammer et al. 2008) and machine-learning (e.g. Carrasco Kind & Brunner 2013) techniques, although there are some hybrid methods as well (e.g. Beck et al. 2017). The template-fitting techniques determine the photometric redshifts by fitting synthetic spectral templates, either empirical or synthesized, from stellar population models to observational spectral templates. A number of variations of this technique exist in the literature, such as the Bayesian photometric redshifts (BPZ; Benitez 2000) and Easy and Accurate photo-Z from Yale (EAZY; Brammer et al. 2008). Machine-learning techniques, also known as empirical methods, use a spectroscopic dataset to train an algorithm, which is then applied to a photometric sample to estimate photometric redshifts. Examples of empirical methods include the Artificial Neural Network (ANNz; Collister & Lahav 2004; Lahav & Collister 2012) and random forest techniques, for example, Trees for photo-Z (TPZ; Carrasco Kind & Brunner 2013).

Each of these techniques has its own advantages and disadvantages. Beck et al. (2017) compared the performance of eight photo-z estimation methods (four template-fitting techniques and four machine-learning techniques). Their analysis revealed that all methods perform adequately when the training set coverage is sufficient, but their performance deteriorates when extrapolation is required. Random forest techniques in particular are not expected to perform well beyond the boundaries of the training set. On the other hand, the latter techniques perform better than the other techniques when the photometric measurement errors increase. Beck et al. (2017) concluded that none of the methods is superior to the others and that a trade-off has to be made depending on the available training set, that is to say, its photometric accuracy and coverage.

The machine-learning methods have been successfully applied to derive photometric redshifts for galaxies (e.g. SDSS; Beck et al. 2016) and optical quasi-stellar objects (QSOs) (e.g. Brescia et al. 2015; Cavuoti et al. 2017). However, for X-ray

⋆ The table of the photometric redshifts is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/qcat?J/A+A/608/A39

AGN, only spectral energy distribution (SED) fitting techniques have been used (Salvato et al. 2009; Hsu et al. 2014). AGN SEDs are more complicated than galaxy SEDs, however, because of contamination from the host galaxy, intrinsic obscuration, variability, and dominance of different components in different spectral bands, for instance. Thus, photo-z for AGN through SED fitting is difficult. On the other hand, machine-learning methods require large spectroscopic training samples to perform well, and X-ray datasets that are suitable to be used as training sets are rare.

We here use X-ray sources detected in the XMM-XXL survey (Liu et al. 2016; Georgakakis et al. 2017) to train, for the first time, a machine-learning algorithm (TPZ; Carrasco Kind & Brunner 2013) to estimate photometric redshifts for X-ray AGN in the X-ATLAS field. Our goal is to use these photo-z estimates in a future paper to estimate the star formation rate (SFR) and stellar mass of these sources and study the connection between the AGN activity and the environment of their host galaxy. In this paper, we check the accuracy of the photo-z estimates. The structure of the paper is as follows: in Sect. 2 we describe the X-ray sources for which we estimate photo-z, in Sect. 3 we briefly describe the TPZ algorithm and provide information for the training sample. The results are presented in Sect. 4, while we discuss and summarize the main conclusions of this work in Sect. 5.

2. X-ray sample

The *Herschel* Terahertz Large Area survey (H-ATLAS) is the largest Open Time Key Project carried out with the *Herschel* Space Observatory (Eales et al. 2010), covering an area of 550 deg^2 in five far-infrared and sub-millimeter (submm) bands (100, 160, 250, 350, and 500 μm). 16 deg^2 have been presented in the Science Demonstration Phase (SDP) catalogue (Rigby et al. 2011) and lie within one of the regions observed by the Galaxy And Mass Assembly (GAMA) survey (Driver et al. 2011; Baldry et al. 2010). *XMM-Newton* observed 7.1 deg^2 with a total exposure time of 336 ks (in the MOS1 camera) within the H-ATLAS SDP area, making the XMM-ATLAS one of the largest contiguous areas of the sky with both *XMM-Newton* and *Herschel* coverage. The catalogue contains 1816 unique sources (Ranalli et al. 2015).

To obtain optical, mid-IR, and far-IR photometry for the XMM-ATLAS sources, we cross-matched the X-ray catalogue with the SDSS-DR13 (Albareti et al. 2015), the WISE (Wright et al. 2010), and the VISTA-VIKING catalogues (Emerson et al. 2006; Dalton et al. 2006) with the ARCHES cross-correlation tool xmatch, which symmetrically matches an arbitrary number of catalogues providing a Bayesian probability of association or non-association (Pineau 2016). xmatch associates one or more tuples with each X-ray source, including possible counterparts in VISTA and/or WISE, with the corresponding probability. When a given X-ray source had more than one associate tuple, we selected those with a probability >0.68, of these, those that were included in most catalogues, and finally, those with the highest probability. The cross-match revealed 1031 sources with at least optical photometry. Using the association probabilities derived by xmatch, fewer than 10% of the counterparts in our catalogue are missmatches (≈85 sources). Of the 1031 sources, 848 have mid-IR counterparts, while 589 also have near-infrared (NIR) counterparts (Table 1). Of the 1031 sources, 174 have spectroscopic redshifts from either the SDSS or the GAMA surveys.

Table 1. Number of X-ATLAS X-ray AGN divided based on their available photometry and optical morphology.

Available photometry	Total number of sources	Point-like sources	Extended sources
SDSS	1031 (174)	576(119)	455 (55)
SDSS+WISE	603 (124)	343 (87)	260 (37)
SDSS+WISE+NIR	423 (92)	249 (67)	174 (25)
SDSS+NIR	653 (122)	380 (86)	273 (36)

Notes. In parentheses we quote the number of sources with available spectroscopic redshift from the SDSS and GAMA surveys.

3. Analysis

3.1. Method

To estimate the photometric redshifts for the X-ray AGN in the ATLAS field, we used the publicly available algorithm TPZ. The technique is described in detail in Kind & Brunner (2013). In brief, TPZ is a parallel machine-learning algorithm that uses prediction trees and random forest techniques to generate photometric redshift probability density functions (PDFs) by incorporating measurement errors in the calculation while also efficiently accounting for missing values in the data.

Random forest is an ensemble-learning method for classification, regression, and other tasks. The method generates prediction trees and then combines their predictions. Prediction trees are built by asking questions that split the data until a stopping criterion is met that creates a terminal leaf. The leaf contains a subsample of the data with similar properties, and by applying a model within the leaf, a prediction is made.

TPZ is an empirical technique and therefore required a dataset with spectroscopically measured redshifts to train the algorithm before it was applied to our photometric X-ray sample. The spectroscopic training sample we used in our analysis is described in the following section.

3.2. Training sample

The X-ray catalogue we used to train the TPZ algorithm comes from the XXM-XXL survey. XMM-XXL covers a total area of about 50 deg^2 with an exposure time of about 10 ks per XMM pointing (Liu et al. 2016; Georgakakis et al. 2017). In the north, 8445 X-ray sources are detected (XXL-N). This region extends to about 25 deg^2. Of these sources, 5294 have optical (SDSS) photometry. Reliable spectroscopy from SDSS-III/BOSS is available for 2512 AGN (Menzel et al. 2016). To increase the size of our training sample, we also included sources from the XWAS (*XMM-Newton* Wide Angle Survey; Esquej et al. 2013), XBS (Della Ceca et al. 2004), XMS (Barcons et al. 2007), and COSMOS (Brusa et al. 2010) surveys. We also added ~1500 optically selected X-ray AGN with spectroscopic redshifts from the SDSS-DR13 dataset by cross-matching the 3XMM-DR5 catalogue with SDSS, UKIDSS (Hambly et al. 2008; Irwin 2008), 2MASS (Skrutskie et al. 2006), and WISE. This increased the total number of sources in our training sample to 5157 (Table 2). Testing the performance of TPZ (see next section) with and without the optically selected X-ray AGN revealed that the inclusion of these extra sources marginally but systematically improved the training process of the TPZ code. Specifically, the outlier percentage (see next section) decreased by 2–3% in all cases. Therefore, the results we present next were estimated using the training sample described above.

Table 2. Number of sources used to train TPZ, with the corresponding available photometry.

Available photometry	Total number of sources	Point-like sources	Extended sources
SDSS	5157	2703 (1900)	2454 (1200)
SDSS+WISE	4781	2473 (1500)	2308 (1400)
SDSS+WISE+NIR	3212	1613 (1000)	1599 (1000)
SDSS+NIR	3313	1679 (1000)	1634 (1100)

Notes. The second column presents the total number of the sources, while the third and fourth columns show the numbers of sources divided into point like and extended. In parentheses we quote the number of sources we used to train TPZ during the validation process (see text for more details).

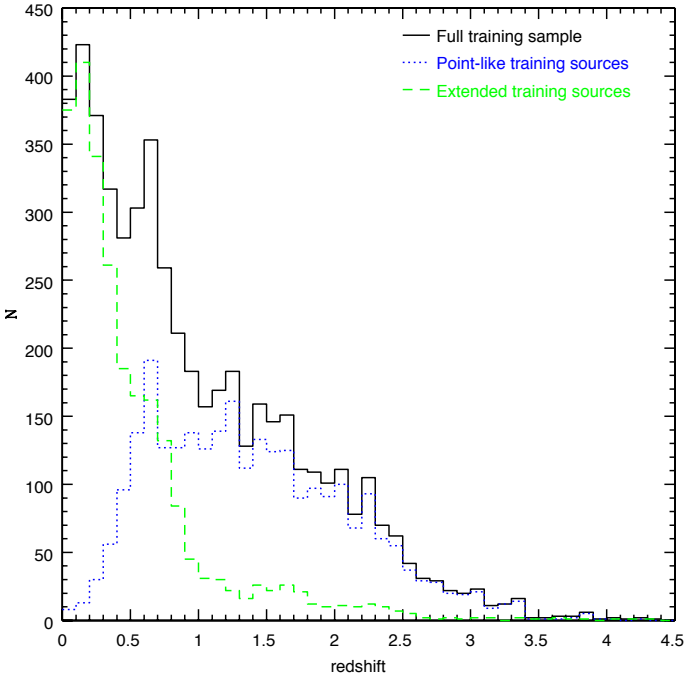

Fig. 1. Redshift distribution of the 5157 sources used to train the TPZ algorithm (black solid line). The dashed and dotted lines present the redshift distribution when we split the training sources into extended and point like based on their optical classification.

In addition to the photometric bands of SDSS (u, g, r, i, and z), we included mid-IR (W1, W2) and near-IR (J, H, K) bands in the training process of TPZ to determine whether its performance improved. For this purpose we cross-matched the 5157 sources with the WISE catalogue and near-IR catalogues, that is, with VISTA, UKIDSS, or 2MASS. The cross-match was performed using the xmatch cross-correlation tool and following the same analysis as described in the previous section for the ATLAS sources. The number of sources we obtained and the available photometry is presented in Table 2. Although TPZ can infer missing photometry, in our validation tests and the estimation of the photometric redshifts of the X-ATLAS sources, only the available photometric bands were used for each subsample.

The redshift distribution of the training set is presented in Fig. 1.

3.3. Checking the performance of TPZ using the training set

To check the performance of TPZ in estimating accurate photometric redshifts, we split our training set into two subsamples. One was used to train the algorithm, and the other subsample was used as a test case for which we estimated photometric sources. This is an ideal scenario since both subsamples share the same region of the parameter space and the same quality of (spectroscopic) data, that is, the same distribution in redshift and magnitude as well as the same photometric errors. To account for the fainter magnitudes of our photometric X-ATLAS sources compared to the spectroscopic training sample and to facilitate a more accurate check of the TPZ performance, in this test we trained TPZ using colours instead of magnitudes. Figure 4 presents two examples of the colour distribution of the training sources (black circles).

The accuracy of the photometric redshifts estimated by TPZ was quantified by two widely used statistical parameters, the normalized absolute median deviation, $\sigma_{\rm nmad}$, and the percentage of outliers, η. $\sigma_{\rm nmad}$ is defined as

$$\Delta(z_{\rm norm}) = \frac{z_{\rm spec} - z_{\rm phot}}{1 + z_{\rm spec}},$$

$$\mathrm{MAD}(\Delta(z_{\rm norm})) = \mathrm{Median}(|\Delta(z_{\rm norm})|),$$

$$\sigma_{\rm nmad} = 1.4826 \times \mathrm{MAD}(\Delta(z_{\rm norm})). \quad (1)$$

The percentage of outliers, η, is defined as

$$\eta = \frac{100}{N} \times (\text{Number of sources with } |\Delta(z_{\rm norm})| > 0.15). \quad (2)$$

Since the near-IR data come from different surveys, the training sample was used to calibrate any possible dependencies on the different filters used, that is, the differences between the K filter on UKIDSS and the $K_{\rm s}$ filter on VISTA and 2MASS. Our tests revealed that there are no differences, regardless of whether we ignored the different filters or scaled K magnitudes to $K_{\rm s}$. For example, using the SDSS+NIR sample for point-like and extended sources, the percentage of outliers differs by $<\pm 0.8\%$ and the difference in $\sigma_{\rm nmad}$ is negligible. Therefore, we ignored this difference in filters in our analysis.

Our initial tests revealed that the performance of the TPZ algorithm in estimating photometric redshifts improved when we split the sources based on their morphology (Salvato et al. 2011). Using the SDSS photometric bands and estimating photometric redshifts without dividing the sources into point like and extended, we obtained $\sigma_{\rm nmad} = 0.12$ and $\eta = 0.35\%$. These numbers are higher than those derived when splitting the sources based on their optical morphology (see Table 3). We also tried to use the morphology as one of the features used to train the algorithm. Our tests revealed that there is no improvement in the accuracy of the photo-z estimations. For example, using ten photometric bands, $\sigma_{\rm nmad} = 0.05$ and $\eta = 11.8\%$. These estimates are in between the values obtained when the sources are split based on their morphology (Table 3). We therefore split the training sources into point like and extended, using their SDSS classification. The number of sources in each subsample is shown in Table 2. Their redshift distribution is presented in Fig. 1. Based on the two distributions, we can reach redshifts of up to 3.5 and 2.5 for point-like and extended sources, respectively.

Table 3 presents the values for the various parameters of TPZ we used to estimate photometric redshifts for each subsample. Nrandom is the number of random realizations that TPZ performs, NTrees is the number of trees used, and Natt the number

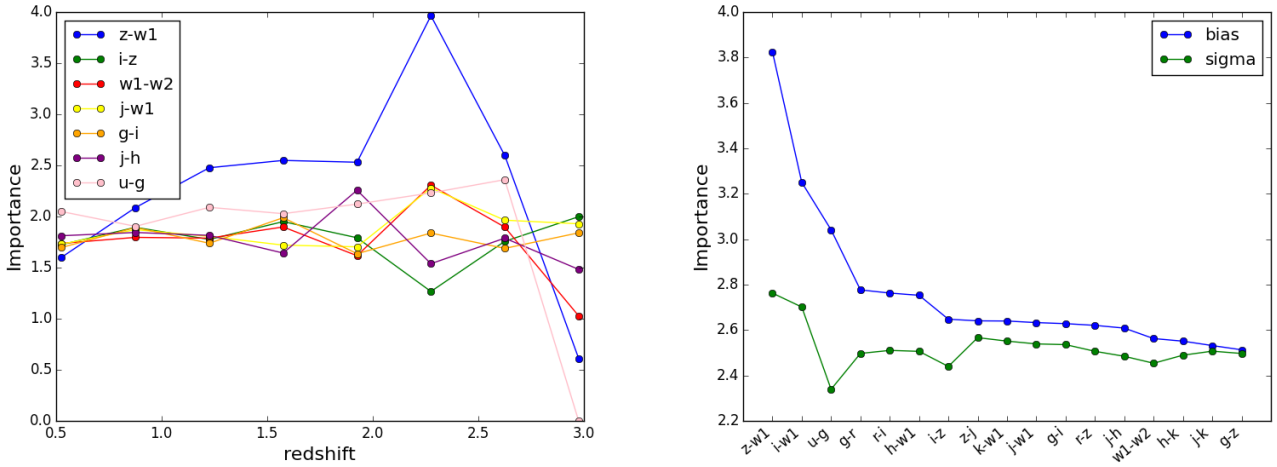

Fig. 2. Point-like sources. *Left*: importance of attributes as a function of redshift. *Right*: RMS importance factor as a function of the attributes computed using the bias and its scatter.

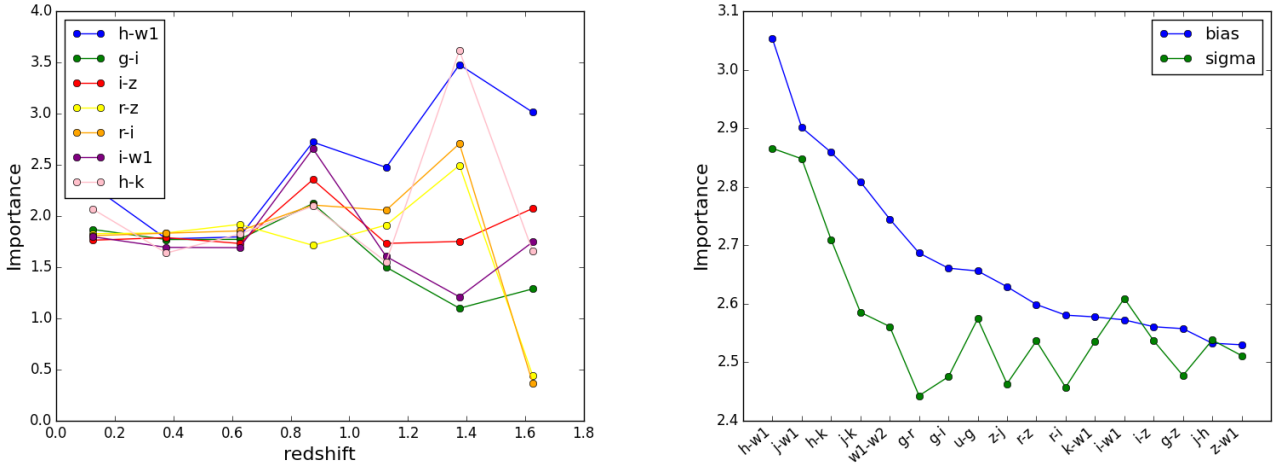

Fig. 3. Same measurements as presented in Fig. 2, but for extended sources.

Table 3. Performance of the TPZ algorithm, estimated by splitting our spectroscopic sample (see Sect. 3.2) into train and test files.

Sample	Point like		Extended		TPZ parameters		
	$\sigma_{\rm nmad}/\eta$ (%)	⟨error⟩	$\sigma_{\rm nmad}/\eta$ (%)	⟨error⟩	Nrandom	NTrees	Natt
SDSS	0.08/27.0	0.33	0.06/18.0	0.21	6	8	7
SDSS+WISE	0.06/17.4	0.25	0.06/13.0	0.20	8	10	8
SDSS+WISE+NIR	0.05/13.7	0.23	0.04/9.0	0.18	6	8	12
SDSS+NIR	0.06/20.0	0.27	0.05/11.5	0.19	8	10	10

Notes. The accuracy of the photometric redshifts is quantified by estimating the normalized absolute median deviation, $\sigma_{\rm nmad}$ and the percentage of outliers, η. The median error of the photometric redshift for each subsample is shown. The values of the TPZ parameters we used for each subsample are also presented.

of attributes for TPZ. The number of the bins used was 50 in the case of extended sources and 70 for the point-like sources. To estimate the PDFs and the confidence level of the estimated photometric redshifts (see Carrasco Kind & Brunner 2013), the rms factor was set to 0.06. The same values for each parameter were used to estimate the photo-z for the 1031 X-ray sources in the ATLAS field (next section).

Figure 2 presents the importance of some of the attributes we used in the training process of the TPZ algorithm. The left panel presents the importance of the attribute as a function of redshift for the point-like sources when ten photometric bands are available. A factor of one in the importance implies that the attribute acts as a random variable (for more details see Sect. 4.1.1. in Carrasco Kind & Brunner 2013). The right panel presents the RMS importance factor as a function of the attributes computed using the bias, defined as $\Delta z = z_{\rm spec} - z_{\rm phot}$, and its scatter. Figure 3 shows the same measurements for the extended sources.

The left panels of Figs. 2 and 3 show that the importance of each attribute is different at different redshifts. In the case of point-like sources, the $z - W1$ colour is the most important attribute up to redshift 2.5, but its importance significantly drops at $z = 3$. Similarly, the importance of the $h - k$ colour in the

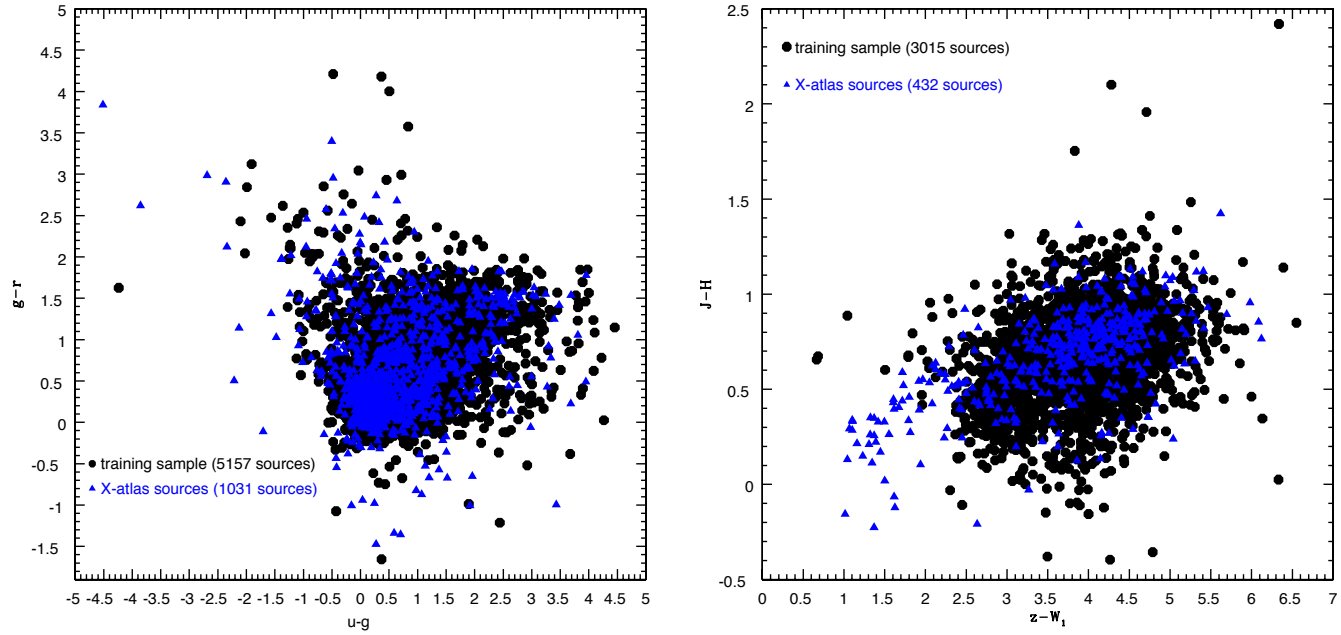

Fig. 4. *Left*: $u - g$ vs. $g - r$ colour distribution of the training sample (black circles) and the X-ATLAS sources (blue triangles). *Right*: $z - W1$ vs. $J - H$ colour distribution of the training sample (black circles) and the X-ATLAS sources (blue triangles). The fraction of the X-ATLAS sources that is well covered by the training set is different for different colour combinations. This is quantified in Table 4.

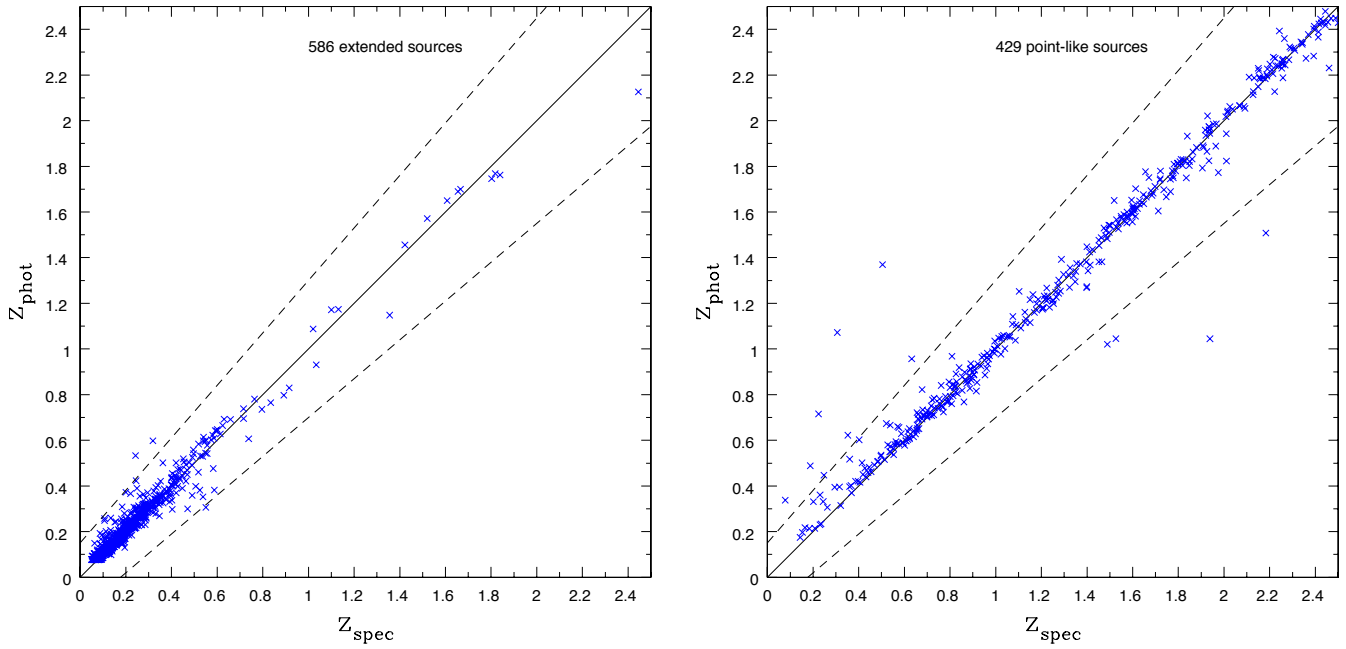

Fig. 5. Performance of TPZ using the ten available photometric bands (SDSS+WISE+near-IR). The training sample has been split into train and test files to compare the estimated photometric redshifts with the spectroscopic redshifts of the sources. The dashed lines correspond to $\Delta z_{\rm norm} = \pm 0.15$. Based on our analysis, the number of outliers is $\eta = 9\%$ and $\eta = 13\%$, for the extended and point-like sources, respectively. The normalized absolute median deviation is $\sigma_{\rm nmad} \approx 0.04$–$0.05$.

case of extended sources significantly drops at $z > 1.4$. Moreover, same colours have a different importance for point-like and extended sources, as can be more clearly seen in the right panels of the two figures. For instance, the $z - W1$ colour is the most important attribute for the point-like sources, but is least important in the case of extended sources. Therefore, the importance of the colours used to estimate photometric redshifts for X-ray sources strongly depends on the morphology of the source and the redshift range of interest.

The results of our measurements are presented in Table 3. When we use optical photometry alone (SDSS), the number of outliers is high, especially in the case of point-like sources. When we add mid-IR colours (WISE), the results improve significantly, while TPZ performs best when we also include near-IR magnitudes in the training process of the algorithm. Figure 5 compares the estimated photometric redshifts with the available spectroscopic redshifts of the sources. Figure 6 presents examples of photometric redshift PDFs produced by TPZ.

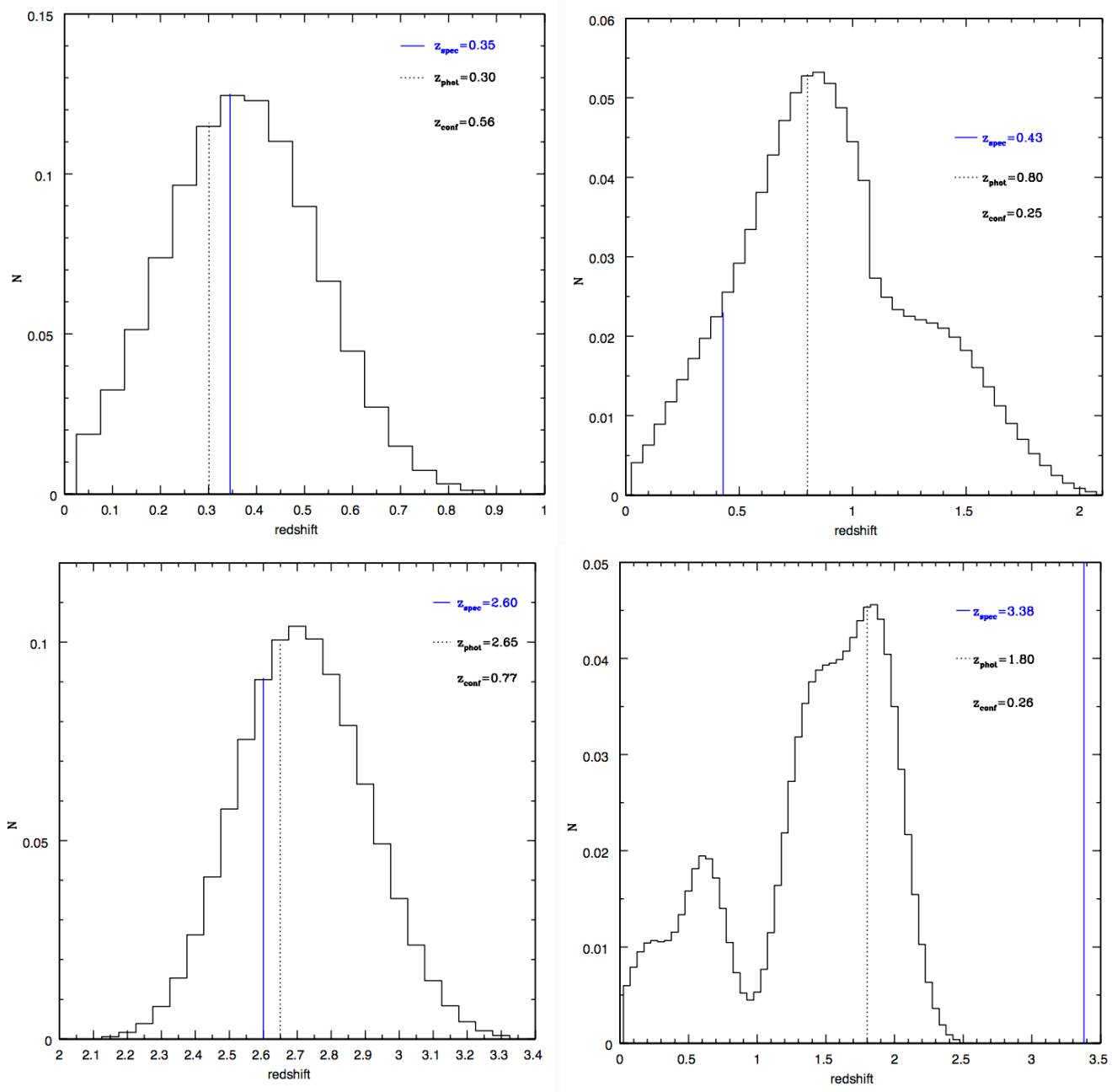

Fig. 6. Examples of PDFs produced by TPZ during the validation process. The *top panels* present results for extended sources and the *bottom panels* for point-like sources. In the *left panels*, the estimated photo-z (dotted line) is in agreement with the spectroscopic redshift (solid line) of the source. In the *right* panels, the estimated photo-z differs significantly from the spectroscopic redshift. These measurements are also characterized by a low confidence level of the photometric redshift.

The number of outliers drops to 9–14% when ten bands are used for the photo-z estimation (Table 3). Although this number is significantly lower than the outlier percentage that we obtain when we used fewer photometric bands, there is a non-negligible number of outliers even among our best photo-z estimates. Figure 7 presents the colour space occupied by the training sample (black circles) for different colour combinations. Outliers (blue triangles) lie within the boundaries of the training set. Therefore, their existence cannot be attributed to the extrapolation in colour space that TPZ may be required to perform. Although the cause of these outliers is uncertain, their percentage can be significantly reduced by applying a cut in the confidence level, $z_{\rm conf}$ (Carrasco Kind & Brunner 2013), of the photo-z. For example, for $z_{\rm conf} > 0.6$, $\eta = 4.5\%$ in the case of point-like sources. The percentage further decreases ($\eta = 2.4\%$) when we consider only photo-z estimated using ten photometric bands. When we apply a $z_{\rm conf} > 0.5$ cut for the extended sources, the corresponding numbers are $\eta = 4.0\%$ and $\eta = 1.2\%$. Figure 8 presents the distribution of $z_{\rm conf}$ for point-like and extended sources.

Variability of AGN can affect the accuracy of the estimated photometric redshifts (Simm et al. 2015). This is not a problem for the optical bands of SDSS we used, since all bands have been observed simultaneously. Variability is also minimum in the mid-IR photometric bands. No estimate of the variable sources in our sample can be made for the near-IR bands, however. We

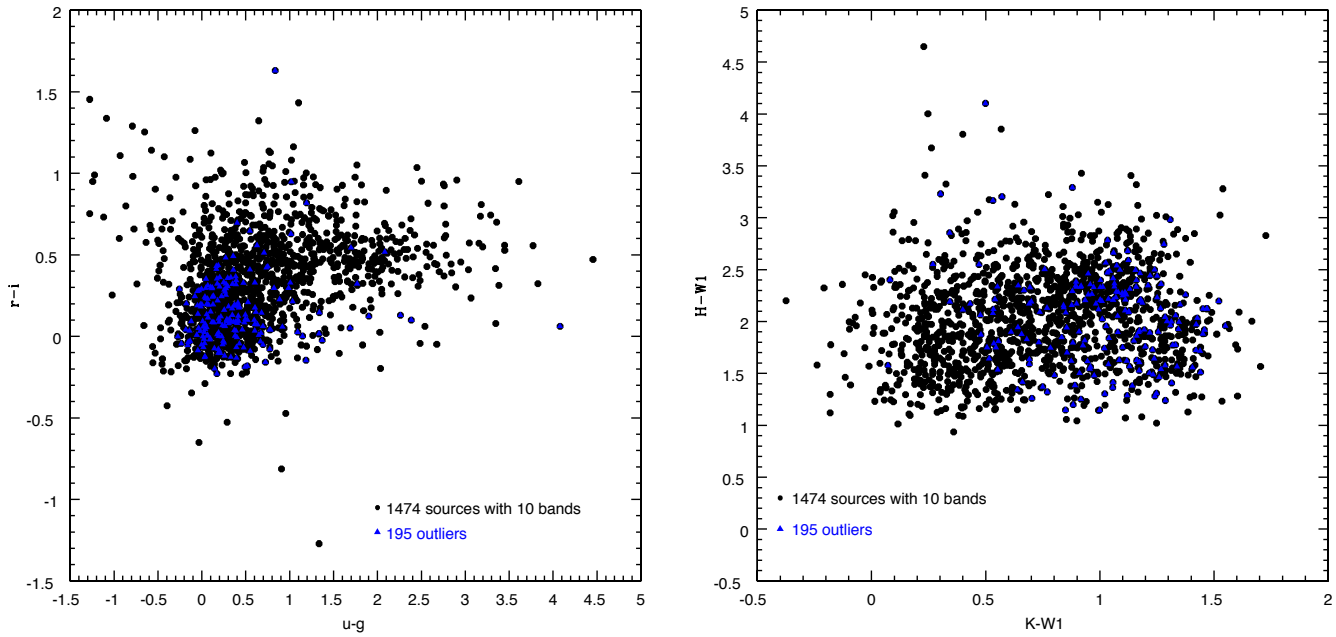

Fig. 7. *Left*: $r-i$ vs. $u-g$ colour distribution of the training sample (black circles) and the outliers (blue triangles). *Right*: $H-W1$ vs. $K-W1$ colour distribution of the training sample (black circles) and the outliers (blue triangles).

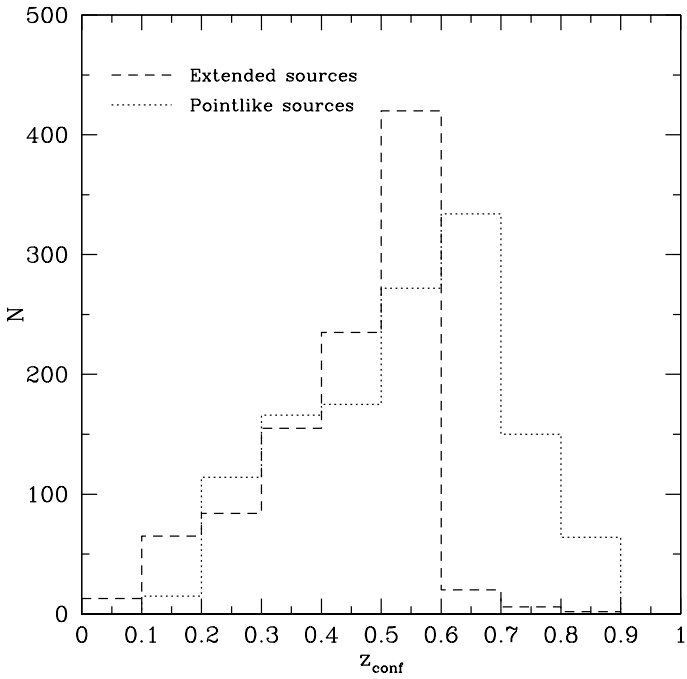

Fig. 8. Distribution of z_{conf} for the extended (dashed line) and the point-like (dotted line) sources in our training sample.

Fig. 9. $i-z$ vs. $g-i$ colour space diagram. Black dots present the sources in our training sample. The black solid line defines the region of the colour space that contains 90% of the training sources as estimated by the KDE test. Green dots are the sources from the X-ATLAS sample inside the 90% region, and red crosses present the remaining X-ATLAS sources.

would expect most of these sources to be excluded when a z_{conf} cut were applied, as discussed above, but a flag cannot be assigned to indicate these sources in the full catalogue.

4. Results

Following the results of the tests during the validation process (see previous section), we split the 1031 X-ATLAS X-ray AGN into point-like and extended sources using their SDSS classification. The number of sources divided based on their optical morphology as well as the available photometry is presented in Table 1.

Machine-learning methods, such as TPZ, are known to perform poorly when no training set coverage is available and extrapolation must be performed (Beck et al. 2017). Figure 4 compares the colour distribution of the X-ATLAS AGN (blue triangles) with that of the training sample (black circles). In both examples, the coverage of the training set seems sufficient to properly train TPZ to estimate the photometric redshift of the X-ATLAS sources. To quantify the differences among the colours between the training and the X-ATLAS samples, we performed a kernel-density estimation (KDE) test. Using KDE, we defined the region in colour space that contained 90% of the training sample. Then we estimated the fraction of the X-ATLAS

Table 4. Fraction of the X-ATLAS sample that is well covered in all possible combinations of colours as well as in at least one colour-colour combination.

Available photometry	Fraction of sources that is well covered in all colour combinations	Fraction of sources that is well covered in at least one colour-colour combination
	Extended/Point like	Extended/Point like
SDSS	51%/56%	98%/91%
SDSS+WISE	40%/44%	99%/94%
SDSS+WISE+NIR	25%/35%	100%/100%
SDSS+NIR	37%/46%	99%/99%

Notes. An X-ATLAS source is considered well covered by the training set in a colour-colour combination when it lies in a region of the colour space that contains 90% of the training sources.

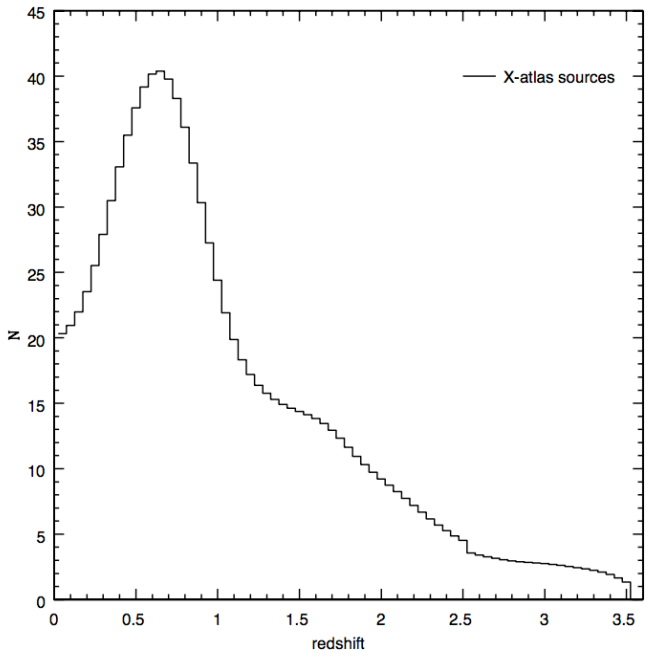

Fig. 10. Redshift distribution of the 933 X-ATLAS sources taking into account the full PDF of each source. Photo-z are estimated using the TPZ algorithm.

sources that were contained in that region, that is, the sources that are well covered by the training sample. This is illustrated in Fig. 9 for the $g - i$ vs. $r - z$ colours. Table 4 presents the fraction of X-ATLAS sample that is well covered in all possible combinations of colours as well as in at least one colour-colour combination.

TPZ estimated photo-z for 933 out of the 1031 sources. Most of the remaining 98 sources have missing photometry, that is, only SDSS bands are available, and therefore the algorithm cannot be properly trained to give a photometric redshift estimate. The distribution of the photometric redshifts for the 933 X-ATLAS X-ray sources, estimated by TPZ and taking into account the full PDF of each sources, is shown in Fig. 10. Of the 933 AGN, 174 have available spectroscopic redshifts from the SDSS and GAMA surveys. In Fig. 11 we compare our photometric redshifts, estimated using TPZ, with the available spectroscopic redshifts. Table 5 presents the median error and the median confidence level, $z_{\rm conf}$, of the photometric redshifts, calculated by TPZ as a function of the available photometric bands.

Table 5. Median error of the photometric redshifts and their median confidence level, estimated by TPZ, for each subsample of the X-ATLAS dataset based on the available photometry.

Available photometry	$\langle z_{\rm conf} \rangle$	$\langle$error$\rangle$
	Extended/Point like	Extended/Point like
SDSS	0.44/0.36	0.21/0.26
SDSS+WISE	0.44/0.46	0.20/0.25
SDSS+WISE+NIR	0.49/0.48	0.19/0.24
SDSS+NIR	0.48/0.47	0.19/0.26

The full catalogue with the estimated photometric redshifts is available at the CDS[1].

To verify how many of the X-ATLAS sources are AGN ($\log L_{\rm X} > 42\,{\rm erg\,s^{-1}}$), we used the X-ray fluxes provided by the XMM-ATLAS catalogue (Ranalli et al. 2015) and the estimated photometric redshifts to calculate the X-ray luminosities. This information is available for 894 sources. Our calculations show that 883 of the sources have $\log L_{\rm X} > 42\,{\rm erg\,s^{-1}}$.

5. Summary and discussion

We presented a catalogue with photometric redshift estimates for 933 X-ray AGN in the ATLAS field. For the first time, we used the largest available X-ray sample to train a machine-learning technique (TPZ) and estimate photo-z for X-ray sources. Our analysis shows that our redshift estimates are accurate when optical photometry is combined with mid-IR photometry in the training process of the algorithm. When additional photometric bands (near-IR) are used, the precision of the photometric redshifts is further improved. Our photo-z estimates have a normalized absolute median deviation, $\sigma_{\rm nmad} \approx 0.06$ and the percentage of outliers is $\eta = 10$–14%, depending on whether the sources are extended or point like. These numbers significantly improve when a cut in the confidence level of the photometric redshift is applied ($z_{\rm conf} > 0.5$–0.6.).

Valiante et al. (2016) and Bourne et al. (2016) presented a catalogue of 120 230 sources with identifications of optical counterparts to submm sources in Data Release 1 (DR1) of the H-ATLAS sample. The sources are located in three fields on the celestial equator, covering a total area of 161.6 deg^2, which was previously observed in the GAMA spectroscopic survey. The catalogue contains photometric redshifts (Smith et al. 2011)

[1] And at http://xraygroup.astro.noa.gr/atlas/atlas-photoz-online.dat

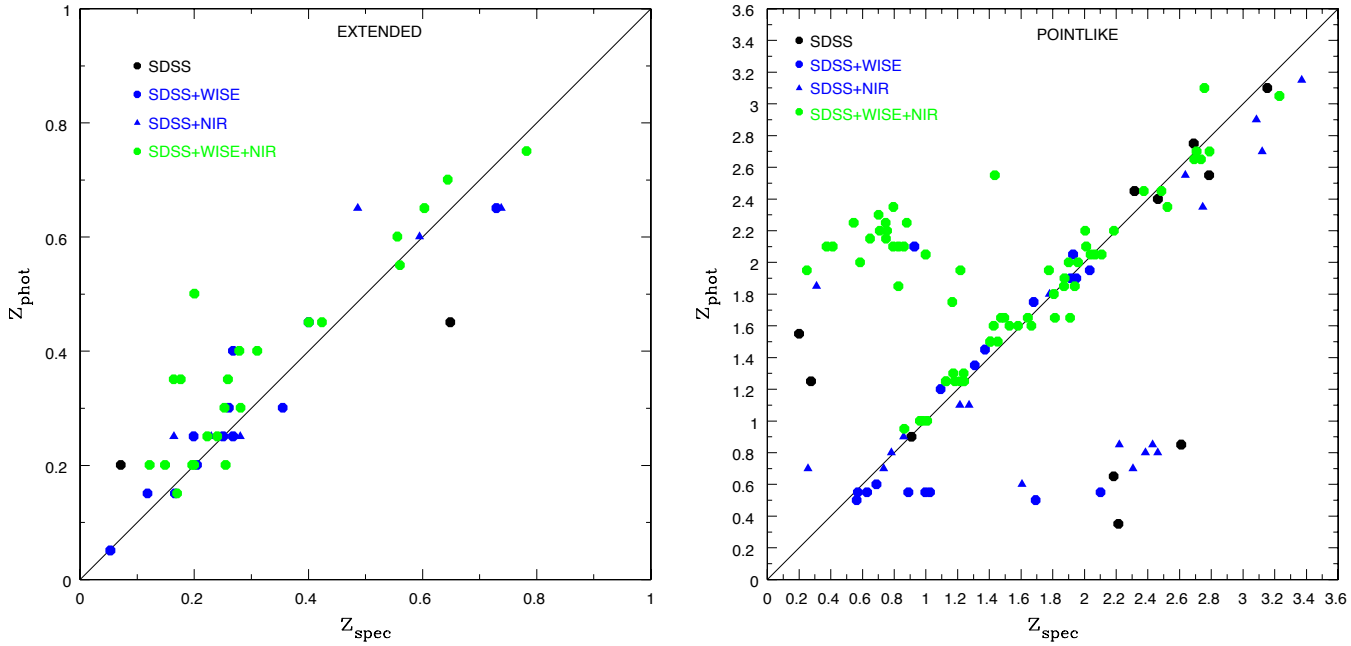

Fig. 11. Comparison of the photometric redshifts estimated using TPZ with the spectroscopic redshifts from the SDSS and GAMA surveys for the 174 of the 933 sources in the ATLAS field. The *left panel* shows the comparison for 55 extended sources and the *right panel* for 119 point-like sources. The median error of the photo-z varies from 0.19 to 0.26 and the median confidence level from 0.36 to 0.49, depending on the morphology of the source and the available photometric bands (Table 5). A significant fraction of outliers exists in the case of the point-like sources, even when seven or even ten photometric bands are used. This number can be greatly reduced when a cut is applied on the confidence level of the photometric redshift, as discussed in the text ($z_{\rm conf} > 0.6$).

measured from the SDSS *ugriz* and UKIDSS *YJHK* photometry using the neural network technique of ANNz (Collister & Lahav 2004). Photometric redshifts have been estimated using a training sample constructed from spectroscopic redshifts from GAMA I, SDSS DR7, 2SLAQ (Cannon et al. 2006), AEGIS (Davis et al. 2007), and zCOSMOS (Lilly et al. 2009), covering redshifts $z < 1$. Of these sources, 5500 lie in the X-ATLAS region, and 3515 have a photometric redshift estimate using ANNz. Sixty-five of these sources are common between the two samples. Figure 12 presents the redshift distribution of the 3515 sources (solid line) and that of the 65 common sources, based on our TPZ photo-z estimations (dashed line). The vast majority of the ANNz photo-z estimates are at $z < 1$ because of the galaxy training sample used for ANNz. In Fig. 13 we compare our photometric redshift estimates using TPZ with those using the ANNz method. Most of the discrepancy between the two photo-z estimates is located in the upper left part of the plot, that is, ANNz computes lower redshift values than we find from our TPZ measurements. Most of this difference is likely due to the different training sets used in the two methods. The training sample of ANNz was constructed to better suit their test sample, the vast majority of which consists of galaxies. Our training sample (Sect. 3.2) consists of X-ray AGN and extends to higher redshifts (up to $z \sim 3.5$; see Fig. 1). Our analysis has shown (Figs. 2, 3, and 9 and Table 4) that the coverage of our training set in feature space, that is to say, in colours, is also sufficient at high redshifts ($z > 1$). The results of this comparison is not an indication that ANNz generally performs poorer than TPZ, but that for the specific X-ray sources our X-ray training set is probably better suited.

Large-scale structure studies (e.g. weak lensing, gravitational waves, clustering) require accurate redshifts in their analysis. Georgakakis et al. (2014) examined the effect of the

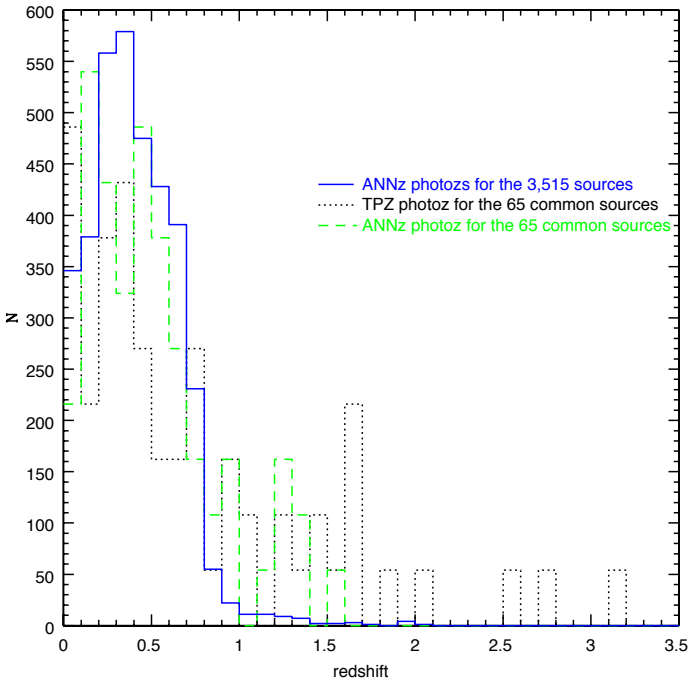

Fig. 12. Redshift distribution of the 3515 sources with ANNz photo-z estimates in the X-ATLAS field (solid line) and the N(z) using TPZ (normalized to the number of sources with ANNz estimates) of the 65 sources that also belong to our X-ray AGN sample. The redshift distribution of the photo-z estimated by ANNz peaks at low redshifts ($z \sim 0.3$), and very few sources have $z > 1$ (solid line). This is expected since ANNz has been trained to estimate photometric redshifts for galaxies. The N(z) estimated using TPZ has been specifically trained to estimate photo-z for X-ray sources and presents a second peak at $z \sim 1.5$ (dotted line).

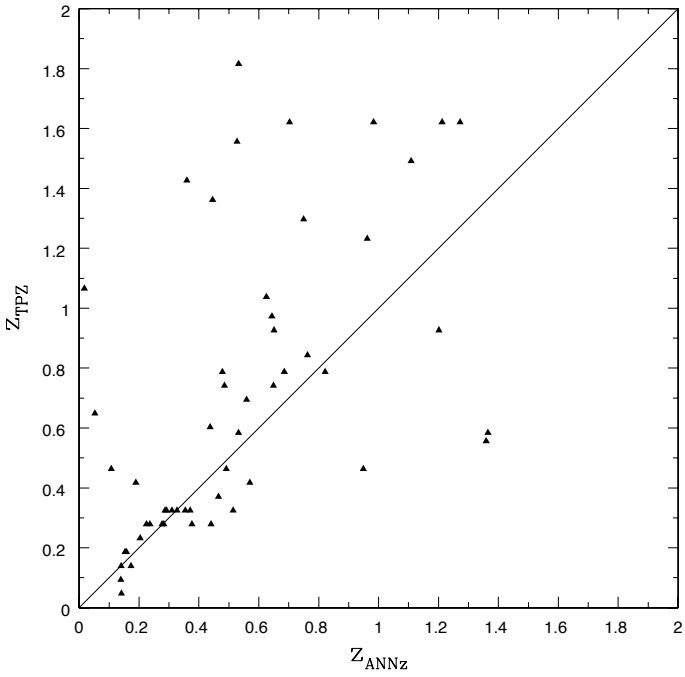

Fig. 13. Comparison of our photometric redshifts estimated using TPZ and those estimated using ANNz (Smith et al. 2011) for the 65 common sources with our X-ATLAS X-ray AGN catalogue and the submm catalogue described in Valiante et al. (2016) and Bourne et al. (2016). Most of the discrepancy between the two photo-z estimates is located in the upper left part of the plot, i.e., ANNz computes lower redshift values than are obtained with our TPZ measurements. Most of this difference is likely due to the different training sets used in the two methods. The training sample of ANNz is constructed to better suit their test sample, the vast majority of which consists of galaxies. Our training sample (Sect. 3.2) consists of X-ray AGN (see text for more details).

accuracy of photometric redshifts on the estimation of the correlation function in clustering measurements. They concluded that a $\sigma \sim 0.04$ (standard deviation of the photo-z) is required in photo-z estimations that are to be used to calculate the AGN correlation function in clustering studies. This accuracy is challenging to obtain, although Georgakakis et al. argued that the clustering signal can be recovered even if the normalized absolute median deviation is $\sigma = 0.08$, when the AGN/galaxy cross-correlation function is measured and the galaxy sample has very accurate photometric redshifts ($\sigma \approx 0.01$). Their analysis takes the error of the photometric redshifts into consideration, but does not account for outliers. Even our best photometric redshift measurements (extended sources with ten available photometric bands) have a considerable percentage of outliers (∼9–10%). Our preliminary results (Mountrichas et al., in prep.) indicate that the clustering signal can be recovered using photometric redshifts derived by TPZ when a cut is applied on the confidence level of the photometric redshift.

The 3XMM catalogue is the largest available X-ray catalogue, containing about 470 000 unique sources covering a total area of 1000 deg^2 on the sky. XMMFITCAT-Z[2] (Corral et al. 2015) is a spectral fit database for 124 000 sources with good photon statistics in the 3XMM. The potential of these catalogues will increase significantly with the addition of the distance information for their sources. We will apply the analysis presented in this work to the 3XMM catalogue to estimate photometric redshifts for all the X-ray sources with at least optical photometry. In the 3XMM-DR5 catalogue, 42 697 sources have available SDSS photometry and 22 619 also have WISE counterparts. 3XMM-DR6 and usage of PanSTARRS in the southern sky will increase the numbers of available X-ray sources. The resulting X-ray catalogue will exceed any other current X-ray catalogue with available redshift information by an order of magnitude.

Acknowledgements. The authors thank the anonymous referee for their careful reading of the paper and their constructive comments. The research leading to these results has received funding from the European Union's Horizon 2020 Programme under the AHEAD project (grant agreement No. 654215). G.M. acknowledges financial support from the AHEAD project, which is funded by the European Union as Research and Innovation Action under Grant No: 654215. F.J.C. and A.C.R. acknowledge financial support through grant AYA2015-64346-C2-1-P (MINECO/FEDER). A.C.R. also acknowledges financial support by the European Space Agency (ESA) under the PRODEX program.

References

Albareti, F. D., Comparat, J., Gutiérrez, C. M., et al. 2015, MNRAS, 452, 4153
Baldry, I. K., Robotham, A. S. G., Hill, D. T., et al. 2010, MNRAS, 404, 86
Barcons, X., Carrera, F. J., Ceballos, M. T., et al. 2007, A&A, 476, 1191
Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371
Beck, R., Lin, C.-A., Ishida, E. E. O., et al. 2017, MNRAS, 468, 4323
Benitez, N. 2000, ApJ, 536, 571
Bourne, N., Dunne, L., Maddox, S. J., et al. 2016, MNRAS, 462, 1714
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brescia, M., Cavuoti, S., & Longo, G. 2015, MNRAS, 450, 3893
Brusa, M., Civano, F., Comastri, A., et al. 2010, ApJ, 716, 348
Cannon, R., Drinkwater, M., Edge, A., et al. 2006, MNRAS, 372, 425
Carrasco Kind, M., & Brunner, R. J. 2013, MNRAS, 432, 1483
Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Corral, A., Georgantopoulos, I., Watson, M. G., et al. 2015, A&A, 576, A61
Dalton, G. B., Caldwell, M., Ward, A. K., et al. 2006, SPIE, 6269, 62690X
Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, ApJ, 660, L1
Della Ceca, R., Maccacaro, T., Caccianiga, A., et al. 2004, A&A, 428, 383
Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971
Eales, S., Dunne, L., Clements, D., et al. 2010, PASP, 122, 499
Emerson, J., McPherson, A., & Sutherland, W. 2006, The Messenger, 126, 41
Esquej, P., Page, M., Carrera, F. J., et al. 2013, A&A, 557, 11
Georgakakis, A., Mountrichas, G., Salvato, M., et al. 2014, MNRAS, 443, 3327
Georgakakis, A., Salvato, M., Liu, Z., et al. 2017, MNRAS, 469, 3232
Hambly, N. C., Collins, R. S., Cross, N. J. G., et al. 2008, MNRAS, 384, 637
Hsu, L.-T., Salvato, M., Nandra, K., et al. 2014, ApJ, 796, 22
Irwin, M. J. 2008, in Processing Wide Field Imaging Data (Berlin Heidelberg: Springer-Verlag), 541
Lahav, O., & Collister, A. A. 2012, Astrophysics Source Code Library [record ascl:1209.009]
Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, ApJS, 184, 218
Liu, Z., Merloni, A., Georgakakis, A., et al. 2016, MNRAS, 459, 1602
Menzel, M.-L., Merloni, A., Georgakakis, A., et al. 2016, MNRAS, 457, 110
Pineau, D. C. 2016, ArXiv e-prints [arXiv:1609.03457]
Ranalli, P., Georgantopoulos, I., Corral, A., et al. 2015, A&A, 577, 10
Rigby, E. E., Maddox, S. J., Dunne, L., et al. 2011, MNRAS, 415, 2336
Salvato, M., Hasinger, G., Ilbert, O., et al. 2009, ApJ, 690, 1250
Salvato, M., Ilbert, O., Hasinger, G., et al. 2011, ApJ, 742, 61
Simm, T., Saglia, R., Sabato, M., et al. 2015, A&A, 584, 22
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Smith, D. J. B., Dunne, L., Maddox, S. J., et al. 2011, MNRAS, 416, 857
Valiante, E., Smith, M. W. L., Eales, S., et al. 2016, MNRAS, 462, 3146
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

[2] http://xraygroup.astro.noa.gr/Webpage-prodec/xmmfitcatz.html