Photometric redshifts for the Kilo-Degree Survey

Machine-learning analysis with artificial neural networks

M. Bilicki[1,2,3], H. Hoekstra[1], M. J. I. Brown[4], V. Amaro[5], C. Blake[6], S. Cavuoti[5,7,8], J. T. A. de Jong[9,1], C. Georgiou[1], H. Hildebrandt[10], C. Wolf[11], A. Amon[14], M. Brescia[7], S. Brough[12], M. V. Costa-Duarte[13,1], T. Erben[10], K. Glazebrook[6], A. Grado[7], C. Heymans[14], T. Jarrett[15], S. Joudaki[16], K. Kuijken[1], G. Longo[5], N. Napolitano[7], D. Parkinson[17,18], C. Vellucci[5], G. A. Verdoes Kleijn[9], and L. Wang[19,9]

[1] Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands
e-mail: bilicki@strw.leidenuniv.nl
[2] National Centre for Nuclear Research, Astrophysics Division, PO Box 447, 90-950 Łódź, Poland
[3] Janusz Gil Institute of Astronomy, University of Zielona Góra, ul. Szafrana 2, 65-516 Zielona Góra, Poland
[4] School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia
[5] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, 80126 Napoli, Italy
[6] Centre for Astrophysics & Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia
[7] INAF – Astronomical Observatory of Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[8] INFN – Section of Naples, Via Cinthia 6, 80126 Napoli, Italy
[9] Kapteyn Astronomical Institute, University of Groningen, Postbus 800, 9700 AV Groningen, The Netherlands
[10] Argelander-Institut für Astronomie, Auf dem Hügel 71, 53121 Bonn, Germany
[11] Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia
[12] School of Physics, University of New South Wales, NSW 2052, Australia
[13] Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, R. do Matão 1226, 05508-090 São Paulo, Brazil
[14] Scottish Universities Physics Alliance, Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[15] Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
[16] Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK
[17] School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia
[18] Korea Astronomy and Space Science Institute, Daejeon 34055, Korea
[19] SRON Netherlands Institute for Space Research, Landleven 12, 9747 AD, Groningen, The Netherlands

ABSTRACT

We present a machine-learning photometric redshift (ML photo-z) analysis of the Kilo-Degree Survey Data Release 3 (KiDS DR3), using two neural-network based techniques: ANNz2 and MLPQNA. Despite limited coverage of spectroscopic training sets, these ML codes provide photo-zs of quality comparable to, if not better than, those from the Bayesian Photometric Redshift (BPZ) code, at least up to $z_{\rm phot} \lesssim 0.9$ and $r \lesssim 23.5$. At the bright end of $r \lesssim 20$, where very complete spectroscopic data overlapping with KiDS are available, the performance of the ML photo-zs clearly surpasses that of BPZ, currently the primary photo-z method for KiDS. Using the Galaxy And Mass Assembly (GAMA) spectroscopic survey as calibration, we furthermore study how photo-zs improve for bright sources when photometric parameters additional to magnitudes are included in the photo-z derivation, as well as when VIKING and WISE infrared (IR) bands are added. While the fiducial four-band $ugri$ setup gives a photo-z bias $\langle \delta z/(1 + z) \rangle = -2 \times 10^{-4}$ and scatter $\sigma_{\delta z/(1+z)} < 0.022$ at mean $\langle z \rangle = 0.23$, combining magnitudes, colours, and galaxy sizes reduces the scatter by ~7% and the bias by an order of magnitude. Once the $ugri$ and IR magnitudes are joined into 12-band photometry spanning up to 12 μm, the scatter decreases by more than 10% over the fiducial case. Finally, using the 12 bands together with optical colours and linear sizes gives $\langle \delta z/(1 + z) \rangle < 4 \times 10^{-5}$ and $\sigma_{\delta z/(1+z)} < 0.019$. This paper also serves as a reference for two public photo-z catalogues accompanying KiDS DR3, both obtained using the ANNz2 code. The first one, of general purpose, includes all the 39 million KiDS sources with four-band $ugri$ measurements in DR3. The second dataset, optimised for low-redshift studies such as galaxy-galaxy lensing, is limited to $r \lesssim 20$, and provides photo-zs of much better quality than in the full-depth case thanks to incorporating optical magnitudes, colours, and sizes in the GAMA-calibrated photo-z derivation.

Key words. galaxies: distances and redshifts – catalogs – large-scale structure of Universe – methods: data analysis – methods: numerical – methods: statistical

1. Introduction

The distance to an astronomical object is arguably one of the most important quantities that we want to measure. In extragalactic studies, except for sparse and mostly local samples of redshift-independent "distance indicators", the best way of estimating source distance is via its redshift. Redshifts can be measured precisely only from spectroscopy, and massive dedicated spectroscopic surveys have been very successful in obtaining them for millions of galaxies. But even the most advanced

techniques, such as multi-fibre spectroscopy, have their limitations: obtaining spectroscopic redshifts (spec-zs) is expensive and time-consuming. Today's largest imaging surveys already include hundreds of millions galaxies, and this number is expected to grow by at least an order of magnitude in the coming years. It is already now infeasible to obtain spectra for even a significant fraction of catalogued galaxies.

Fortunately, many applications do not require the redshift precision available from spectroscopy. Various approaches can be employed instead to estimate redshifts, both on an individual basis, as well as for redshift distributions of particular samples. As far as the individual redshifts are concerned, broad-band photometry can be used to derive photometric redshifts (photo-zs; Baum 1957; Koo 1985; Loh et al. 1986), using two main approaches, sometimes in concert (Brodwin et al. 2006; Hildebrandt et al. 2010): (i) empirical, usually machine-learning (ML); and (ii) source energy distribution (SED), or template, fitting.

In the ML domain, techniques such as artificial neural networks (ANNs, Tagliaferri et al. 2003; Firth et al. 2003), boosted decision or regression trees (BDTs, Gerdes et al. 2010), Gaussian processes (Way et al. 2009), or genetic algorithms (Hogan et al. 2015), to list just a few, are calibrated (trained) on spec-z samples, which have the relevant set of passbands measured, to derive the mapping from photometry to spec-zs, and the best-fit solution is then propagated to the target data with photometry only. These methods are usually agnostic to any physics, and thus need well-controlled and representative training sets to work properly. If the latter are available, the ML photo-z approaches usually provide both very accurate (minimal bias) and precise (low scatter) estimates. In addition to magnitudes, they can also directly use other galaxy observed properties as inputs, such as colours, sizes, half-light radii, and so on (for example Collister & Lahav 2004; Wadadekar 2005; Wray & Gunn 2008). A recently proposed extension of ML photo-z estimation is by working directly on imaging data instead of using post-processed source catalogues; this is possible thanks to "deep learning" (for example Hoyle 2016; D'Isanto & Polsterer 2018).

Among the advantages of ML methods (MLMs) is their ability to automatically handle some systematics in the data, such as varying aperture bias as a function of wavelength, which can produce errors in SED fitting if not dealt with correctly. Last but not least, the empirical methods are able to "learn from data" – their performance gets increasingly better as the training data improve in quantity and quality. The major drawback of MLMs is their poor performance in extrapolation, that is, ML photo-zs are usually not reliable beyond the range of magnitudes, colours, etc., spanned by the training sets.

On the other hand, SED-fitting uses a more direct and physically motivated approach of matching the measured multi-band magnitudes, or fluxes, to the best-fit redshifted spectrum, the latter coming from libraries of either real galaxy spectra and/or artificial ones (for example Benítez 2000; Bolzonella et al. 2000; Brammer et al. 2008). The main advantage of these methods is that they are largely independent of spectroscopic calibration, although they might require priors to avoid assigning unrealistically high redshifts to galaxies of bright observed magnitudes (Kodama et al. 1999; Brammer et al. 2008). The two main drawbacks of SED-fitting photo-zs are: (i) template model dependence, which requires knowledge of realistic galaxy SEDs at various redshifts; (ii) their general inability to use parameters other than magnitudes or fluxes (such as galaxy sizes or shapes).

The empirical methods for deriving individual photo-zs always require spectroscopic calibration data, even if the requested properties of these data differ for various techniques.

Overlapping spec-zs are also needed to judge the performance of the methods, and this includes the SED-fitting ones as well. Generally, it can be stated that every approach for redshift estimation requires spec-z samples at some stage of its application or performance testing.

In this paper we present a machine-learning photo-z analysis for the Kilo-Degree Survey (KiDS, de Jong et al. 2013). KiDS is one of the major wide-angle photometric surveys currently undertaken, along with the Dark Energy Survey (The Dark Energy Survey Collaboration 2005) and the Hyper Suprime-Cam Subaru Strategic Programme (Aihara et al. 2018), and all three are precursors for even more ambitious efforts such as the Large Synoptic Survey Telescope (LSST Science Collaboration et al. 2009) and Euclid (Laureijs et al. 2011). These surveys face a common challenge of the necessity of using photo-zs for scientific analyses, as spec-zs are and will be available only for a very small fraction of detected sources.

The KiDS pipeline photo-z solution, used in most of the scientific analyses so far, comes from the Bayesian Photometric Redshift (BPZ, Benítez 2000) SED-fitting code. However, two ML approaches are also used for deriving alternative photo-zs in KiDS: MLPQNA (Cavuoti et al. 2012), and ANNz2 (Sadeh et al. 2016). This paper aims at quantifying the performance of these MLMs in the most recent Data Release 3 (DR3) of KiDS. This has already been briefly presented in the DR3 publication (de Jong et al. 2017) and here we provide a more detailed discussion. The paper is accompanied by two ANNz2-based KiDS photo-z catalogues and serves as a reference for their end-users.

The overall structure of this paper is the following. First, in Sect. 2 we present the photo-z codes used in this work: ANNz2 (Sect. 2.1), and MLPQNA (Sect. 2.2). Next, in Sect. 3 we describe the data employed in our studies: photometric from KiDS (Sect. 3.1), VIKING (Sect. 3.2), and WISE (Sect. 3.3), as well as spectroscopic coming from various samples overlapping with KiDS (Sect. 3.4). A summary of the joint photo-spectro sample is provided in Sect. 3.5.

We then explore ML photo-zs in two different regimes and setups of the KiDS data. First, in Sect. 4 we study the performance of the two ANN-based algorithms at almost the full depth of KiDS, using various overlapping spec-z datasets as training and test samples; we also compare the results with those from the fiducial KiDS photo-z solution from BPZ (Sects. 4.1–4.3). We conclude that Section by describing in Sect. 4.4 the publicly released KiDS DR3 full-depth photo-z catalogue obtained by applying the ANNz2 algorithm. An earlier version of that dataset was already made available with the DR3 release[1] (de Jong et al. 2017) and is now updated with this paper.

In the second set of experiments, described in Sect. 5, we use ANNz2 for the bright end of KiDS, for which there is very complete spectroscopic training data from the Galaxy And Mass Assembly (GAMA, Driver et al. 2011) survey. We study how the basic KiDS *ugri* parameter space can be extended to improve photo-zs at the GAMA depth, by adding further imaging information, such as galaxy morphology. We also examine what can be gained in terms of photo-z quality if the wavelength range is extended by adding VIKING near-infrared (IR) and WISE mid-IR information. This is of particular importance because dedicated reductions of the relevant data are either ongoing (KiDS-VIKING) or planned (KiDS-WISE). The results of these tests are detailed in Sects. 5.1–5.3. The GAMA-based analysis is also accompanied by a public catalogue release, in this case

---

[1] http://kids.strw.leidenuniv.nl/DR3/ml-photoz.php#annz2

limited to $r \lesssim 20$ mag, with much more accurate and precise photo-zs than in the global solution; see Sect. 5.4. Such a sample with precise and accurate photo-zs is of particular interest for studies such as galaxy-galaxy lensing, which require foreground data with well-constrained redshift estimates.

In Sect. 6 we conclude and mention future prospects regarding KiDS photo-zs.

2. Photometric redshift algorithms used

In this Section we provide details of the two approaches used to obtain KiDS ML photo-zs, ANNz2 and MLPQNA. The results from these two codes will be compared to the KiDS pipeline solution derived with the Bayesian Photometric Redshift algorithm (BPZ, Benítez 2000), and made publicly available together with the DR3 photometric data (de Jong et al. 2017). For the details of how BPZ was implemented in the KiDS pipeline, please see the relevant papers: Kuijken et al. (2015) and de Jong et al. (2017).

2.1. ANNz2

Most of the analysis of this paper, as well as the two accompanying photo-z catalogues, are based on the ANNz2 code (Sadeh et al. 2016). ANNz2 is a versatile ML package[2], originally designed as a successor of the ANNz software (Collister & Lahav 2004). However, unlike its predecessor, ANNz2 is not limited to using only artificial neural networks (ANNs) but it also incorporates other machine-learning methods (MLMs), such as boosted decision or regression trees (BDTs). ANNz2 is based on the Toolkit for Multivariate Data Analysis (TMVA) package[3] (Hoecker et al. 2007), which itself is part of the ROOT C++ software[4] (Brun & Rademakers 1997), and therefore allows the user to use various MLMs. In this study we have limited ourselves to exploring only the fiducial MLMs of ANNz2, namely ANNs and BDTs. ANNz2 provides also other important improvements over ANNz. The first one is a high level of work automatisation via Python scripts, thanks to which the user does not have to define the individual MLM properties, allowing the software to generate their architectures randomly (which we applied here). By training a large ($\gtrsim 100$) number of ANNs and/or BDTs with various architectures – in the Randomized Regression mode which we employed in our study – the photo-z derivation can be optimised both by using the "best" solution, as well as by folding all or part of all the solutions from each run. This allows for an overall improvement in the photo-z quality without much user involvement in the training procedure.

The Randomized Regression mode of ANNz2 allows for deriving the probability distribution functions (PDFs) of the computed photo-zs, by folding selected individual MLM results with their uncertainty estimates, the latter being derived using a k-nearest neighbours (kNN) estimator (Oyaizu et al. 2008). However, these PDFs should not be treated as actual error distributions with respect to the true redshift (which is unknown) but rather as quantification of the uncertainties of the photo-z derivation method. This will however apply to most photo-z techniques that derive PDFs, including the fiducial KiDS method, BPZ (see the accompanying analysis by Amaro et al. 2018). In general, we do not store these PDFs in the catalogues presented here, but they can be generated on request.

Last but not least, a major improvement in ANNz2 over ANNz (and several other ML photo-z codes) is the possibility to weight the training data to mimic the target set. These weights can then be propagated throughout the training and evaluation procedure, by assigning a correction factor to the training objects depending on the input parameters. The weighting is done via the kNN method in the parameter space chosen by the user (for instance magnitudes, colours) by comparing the density of input sources to that of the target ones (Lima et al. 2008). A similar approach was taken in the KiDS cosmic shear analysis by Hildebrandt et al. (2017) to estimate the true redshift distributions of KiDS sources from the matched spectroscopic catalogues (the "DIR" calibration method therein).

The general framework of ANNz2 is similar to most other photo-z MLMs. The code is fed with training and validation sets that have both the input (for example photometric) and output (for example redshift) parameters. If weighting of the training and validation data is requested, this is done at the beginning in the "`generate input trees`" stage of the procedure. A user-defined number and type of MLMs are first trained and then validated on the relevant data; the latter procedure is called optimisation in ANNz2. Thus trained and validated MLMs can then be applied to "blind" data – evaluation sets – either including spec-zs for performance checks, or photometric-only for generating the final catalogues.

We followed the recommendations of Sadeh et al. (2016) to use at least 100 MLMs for Randomized Regression. Training BDTs is much faster than training ANNs for the same number of MLMs; on the other hand, the former requires more storage space and more memory in the optimisation and evaluation process than the latter. The two types of MLMs also differ in performance: our experiments show that using BDTs generally gives worse results than ANNs, even if the number of the former is (much) larger than of the latter. In this paper we thus present results based on ANNs only; in most cases we used 250 ANNs for each experiment, with architectures always defined randomly within the code. We note that a different, perhaps more optimal, setup of ANNz2 is possible if the ANNs are not generated randomly by the code but rather defined by the user, adjusted to the properties of the data (for example to the number of input parameters). In such a case, using fewer ANNs could give similar results to the approach we adopted here (John Soo, priv. comm.). However, running ANNz2 would then require more user supervision; we thus opted for the fully randomised approach which allowed us to execute the computations in the background.

ANNz2 provides various parameters to be set up by the user. We tested the influence of several of them on the final results and we eventually decided for the following configuration (see Sadeh et al. 2016 as well as the ANNz2 online documentation for details):

- `optimCondReg`: a metric used to rank the performance of individual MLMs, its options are the bias, the 68th percentile scatter, or the outlier fraction; in our experiments we found no significant difference between results for the "`sig68`" and "`bias`" options, and we used `optimCondReg = bias` everywhere;
- `optimWithScaledBias`: used as an optimisation criterion for the best MLM and the PDFs; we used `True`: the normalised bias $(z_{phot} - z_{spec})/(1 + z_{spec})$ was employed for optimisation;
- `optimWithMAD`: we used `True`: the best MLM and the PDFs were optimised using the MAD (median absolute deviation) rather than the 68th percentile of the bias distribution;

- splitting of the training + validation data into separate training and validation sets was done randomly into two halves using the ANNz2 option `glob.annz["splitType"] = "random"`
- by default, ANNz2 does not use the actual errors of the training parameters but derives an error model from the data using the kNN-error method; the user can, however, propagate the actual parameter errors directly; we have tested this latter option for our deep calibration data (zCOSMOS; Sect. 4.3), as well as for the case when low signal-to-noise WISE data were additionally used (Sect. 5.2) and found only slight improvements in the results, or none at all; therefore, we used the default setup;
- in some cases, as described in the text, we applied weighting of the training data (`useWgtkNN = True`) using a relevant reference sample; these weights were then used in the whole photo-z estimation procedure;
- ANNz2 outputs five types of point estimates of photo-zs; the first of them, `ANNZ_best`, comes from the single MLM which provides the best combination of performance metrics; the remaining ones are based on photo-z PDFs which are derived internally but do not have to be stored by the user (`glob.annz["doStorePdfBins"] = False`); the PDFs come in two options (one based on the true target as known from the training data, the other based on the results of the best MLM) and two pairs of related photo-z point estimates are derived: `ANNZ_PDF_avg_0` and `ANNZ_PDF_avg_1` – averages of the first (second) PDF types (using the full weighted set of MLMs, convolved with uncertainty estimators), as well as `ANNZ_MLM_avg_0` and `ANNZ_MLM_avg_1` – unweighted averages of all the MLMs which have non-zero PDF weights, that is, of those MLMs that have good performance metrics; our experiments show that the best performance is usually achieved by `ANNZ_MLM_avg_1` and we will be reporting statistics based on this point estimate;
- we do not use full PDFs in any other way than by employing point estimates based on them as described above; the PDFs for the published datasets can however be derived on request.

All the input features used in training as well as in kNN-weighting were normalised to the range $[-1; 1]$ via linear rescaling; this is the default ANNz2 setup (`doWidthRescale = True`).

2.2. MLPQNA

In the KiDS DR3 experiments of Sect. 4 we compare the ANNz2 results with those from another machine-learning approach used in the survey, namely MLPQNA (Cavuoti et al. 2012), which stands for the Multi Layer Perceptron feed-forward neural network (MLP; Rosenblatt 1962), trained by the Quasi Newton Algorithm (QNA; Byrd et al. 1994) learning rule. This ML model is among the most efficient optimisation methods searching for the minimum of the MLP training error function, since it makes use of a statistical approximation of the Hessian of this error, obtained by an iterative MLP network error gradient calculation. MLPQNA makes use of the L-BFGS algorithm (Limited-memory Broyden-Fletcher-Goldfarb-Shanno; Byrd et al. 1994), originally designed for problems with a wide parameter space.

The analytical details of the MLPQNA model, as well as its performance for photo-z estimation, have been extensively discussed elsewhere (Cavuoti et al. 2012, 2015a; Brescia et al. 2013), and the method has been to an earlier KiDS data release, DR2 (Cavuoti et al. 2015b). Within KiDS DR3, it is embedded as a photo-z prediction kernel into the METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts) pipeline (Cavuoti et al. 2017), able to extend the photo-z estimation by providing also their error PDFs. The details of its application to the DR3 data are discussed in de Jong et al. (2017) and the resulting catalogue was released together with the overall DR3 data[5].

MLPQNA is publicly available through the DAMEWARE (DAta Mining & Exploration Web Application REsource; Brescia et al. 2014) web-based infrastructure[6].

3. Input data

In this Section we present the data used in our studies. Most of the results described here are based on public photometric data from the KiDS DR3 (de Jong et al. 2017), supplemented with some additional photometry outside of the nominal KiDS footprint, as well as with public and proprietary spectroscopic datasets. Part of the analysis also uses infrared photometry derived from VIKING and WISE surveys. Below we provide the details of the samples used in this paper.

3.1. KiDS photometric data

The Kilo-Degree Survey (KiDS, de Jong et al. 2013) is a wide-angle imaging campaign being conducted with the Omega-CAM camera (Kuijken 2011) at the VLT Survey Telescope (Capaccioli et al. 2012), using four broad-band optical filters (*ugri*). The target area of the survey is ~1500 deg^2 in two patches, one on the celestial Equator, and the other in the South Galactic Cap. The main science goal of KiDS is to map the large-scale distribution of matter, and extract related cosmological information, using weak lensing techniques (Hildebrandt et al. 2017; Joudaki et al. 2017, 2018; Köhlinger et al. 2017; van Uitert et al. 2018), it is however also perfectly suitable for studying galaxy evolution (Tortora et al. 2016), structure of the Milky Way (Pila Díez 2015), detecting galaxy clusters (Radovich et al. 2017) and high-redshift quasars (Venemans et al. 2015), as well as looking for strong lenses (Petrillo et al. 2017), or even Solar System objects (Mahlke et al. 2018), to name just a few applications.

KiDS has had three data releases so far (de Jong et al. 2015, 2017) and DR3 includes about 450 deg^2 of photometric data, with typical 5σ depth of 24.3, 25.1, 24.9, 23.8 mag in 2″ apertures in *ugri*, respectively. Accurate colours and absolute photometric calibration down to ~2% in *gri* and ~3% in *u* are ensured via a specific photometric homogenisation scheme. In the *r* band, which is used for galaxy shape measurements, the typical PSF size is below 0.7″; sub-arcsecond seeing is also used for the *g* and *i* band observations, while in *u* the mean PSF is 1″. All this guarantees excellent-quality deep imaging, perfectly suitable for astrophysical studies where precise photometry is crucial.

The details of KiDS data reduction are provided in the relevant papers (de Jong et al. 2015, 2017); of importance for this work is that the basic catalogues are produced using the SExtractor (Bertin & Arnouts 1996) software in dual-image mode, which provides several magnitude types for each band, measured directly on astrometrically and photometrically calibrated, stacked images ("coadds"). Among them are Kron-like automatic aperture magnitudes `MAG_AUTO`, as well as isophotal ones, `MAG_ISO`. Two types of catalogues are produced: single-band, with source extraction and photometry done independently

---

in each band, and multi-band, which we use here, where source detection is based on the *r* band, and aperture-matched photometry is derived for the other filters.

KiDS data reduction also involves a post-processing stage in which Gaussian Aperture and Photometry (GAaP, Kuijken 2008) magnitudes are derived (Kuijken et al. 2015). For this, the coadds are first "Gaussianised", meaning that the point spread function (PSF) is homogenised across each individual coadd. The photometry is then measured using a Gaussian-weighted aperture (the size and shape of which are set by the *r*-band major and minor axis lengths and orientation) that compensates for the seeing differences between the filters because each part of the source gets the same weight across all filters. We will call this procedure "PSF homogenisation" from now on.

Additional "photometric homogenisation" is achieved by adjusting the zeropoints across the full survey area. This is done using the coadd overlaps in the *r* and *u* bands, homogenising the photometry in these two filters, and then *g* and *i* bands are tied to the *r* band using stellar locus regression, which homogenises the *g* − *r* and *r* − *i* colours, and therefore the *g* and *i* band zeropoints. The photometric homogenisation is done using the GAaP photometry, and in the final catalogues the resulting zeropoint offsets ("ZPT_OFFSET_band" for each filter) are reported in separate columns, together with Galactic extinction corrections which are based on the Schlegel et al. (1998) maps. The zeropoint-calibrated and extinction-corrected magnitudes will be denoted as "calib" from now on:

$$\texttt{MAG\_type\_band\_calib} =$$
$$\texttt{MAG\_type\_band} + \texttt{ZPT\_OFFSET\_band} - \texttt{EXT\_SFD\_band} \,, \quad (1)$$

where the uncalibrated measurements were taken directly from the KiDS multi-band catalogue. However, since the zeropoint offsets were derived from GAaP measurements, they work better for the GAaP photometry than for other types.

The GAaP magnitudes are the default ones for KiDS, and are used in most of the scientific analyses. They are also applied in the pipeline-photo-*z* derivation with BPZ (Kuijken et al. 2015), as they provide very good galaxy colours. Our studies presented here will also use GAaP magnitudes as defaults. In Sect. 5 we show quantitatively that indeed this type of photometry is the most optimal for photo-*z* estimation among the 3 tested types available from KiDS multi-band data (the other being ISO and AUTO), even for bright sources. One should bear in mind, though, that the GAaP magnitudes cannot be generally used as proxies for total fluxes of galaxies, especially at the bright end where they severely underestimate the total flux (by ~1 mag or more).

Unless indicated otherwise, the KiDS data we use have undergone appropriate cleaning of bad photometry. First of all, in all the analysis we used only those sources which have GAaP magnitudes measured for each band, to guarantee that photo-*z*s are estimated using the full *ugri* information. These cuts apply mostly to the *u* and *i* bands, in which respectively 13% and 7% of KiDS sources do not have magnitude measurements in the multi-band catalogue because of a combination of intrinsically lower source brightnesses in *u* and decreased depth in both *u* and *i* bands, as compared to *g* and *r* (cf. Table 3 in de Jong et al. 2017). Once this filtering is applied in all the bands, the DR3 sample is reduced to 39.2 million objects.

Such a four-band requirement is obviously a limitation for the current analysis, especially compared to the BPZ approach where the photo-*z*s are derived for all the KiDS sources, and upper limits, non-detections, and lacking measurements are handled appropriately. However, the photo-*z*s using fewer bands will

be obviously of worse overall quality than the *ugri*-based ones, which would lead to inhomogeneities in the eventual ML photo-*z* catalogue. We postpone a detailed analysis of the influence of missing bands on KiDS photo-*z*s to the forthcoming KiDS-VIKING nine-band data release, where this situation will be much more common.

Furthermore, we defined a "CLEAN" sample by additionally requiring that magnitude errors are provided in each band, as well as by removing artefacts with any of the following masking flags set: readout spike, saturation core, diffraction spike, secondary halo, or bad pixels[7], following Radovich et al. (2017). The resulting CLEAN dataset includes 36.9 million KiDS-DR3 objects out of 48.7 million in the full multi-band catalogue.

For the purpose of photo-*z* derivation in DR3 we also define a "FIDUCIAL" dataset, which is based on the CLEAN sample additionally purified of stars (by applying the SG2DPHOT = 0 flag[8]) and trimmed at the faint end to encompass the magnitude ranges of the spectro-photo training set described in Sect. 3.4. More precisely, we removed from the KiDS DR3 those sources for which any of the *ugri* magnitudes were beyond the 99.9th percentile of the spectroscopic catalogue distribution. These cuts are MAG_GAAP_u_calib < 25.4, MAG_GAAP_g_calib < 25.6, MAG_GAAP_r_calib < 24.7 & MAG_GAAP_i_calib < 24.5. Applying these cuts on the artefact-purified DR3 dataset gives 20.5 million sources in the FIDUCIAL sample. This sample will be used as the reference set for weighting the spectroscopic catalogue, used for training of the global DR3 photo-*z* solution, as discussed in Sect. 4.4.

We emphasise that in the released full-depth catalogue, the photo-*z*s are derived for all the sources that have the 4 *ugri* GAaP magnitudes measured, although they will be most likely unreliable outside the FIDUCIAL dataset, and of course do not have any meaning for stars. In order not to propagate residual bad photometry to photo-*z* calibration, in the training and validation (optimisation) phase we additionally applied MAGERR_GAAP_band < 1 for each band, but not in the tests nor the final evaluation in the target catalogue. Such an additional cut affects mostly the *u* filter, and removes an extra ~3% from the training data.

We also used KiDS-like observations outside of the nominal KiDS footprint, namely from VST imaging of deep spectroscopic fields described in Sect. 3.4: CDFS (from the VOICE survey, Vaccari et al. 2016) and two DEEP2 fields (2 h and 23 h). Details of observing conditions of these observations are provided in Hildebrandt et al. (2017) and Appendix C. Here it is sufficient to note that they were of comparable quality as the full KiDS.

3.2. VIKING photometry

We also tested how going beyond KiDS photometry can improve the photo-*z*s. The planned KiDS footprint is practically fully covered by the VISTA Kilo-degree Infrared Galaxy survey (VIKING, Edge et al. 2013) providing five near-IR bands $zYJHK_s$ at a similar depth to KiDS, and a joint KiDS-VIKING data reduction is ongoing. At the time of performing the experiments described in this paper, we did not yet have access to these joint data, and thus limit our tests to GAMA-LAMBDAR

---

[7] This was done by applying the bitwise operator IMAFLAGS_ISO_band & 01010111 = 0 for each band. See Appendix A.2 of de Jong et al. (2017) for more details of these flags.

[8] SG2DPHOT is a KiDS star-galaxy classification flag derived from the *r*-band source morphology (de Jong et al. 2015, 2017). Extended sources are assigned a value of 0.

(Wright et al. 2016) forced VIKING photometry on the GAMA sources. These tests are therefore currently limited to KiDS-GAMA objects in the equatorial fields, and apply only to GAMA depth in KiDS ($r \lesssim 20$ mag). The input photometry, and in particular the apertures used for these forced-photometry VIKING measurements, came from SDSS DR7. They are therefore of worse quality than what can be expected from a similar approach using KiDS sources instead. They also had no homogenisation of a similar form as in KiDS applied.

The LAMBDAR measurements come in the form of fluxes, and we also used those that were negative or zero[9]. We discarded only those sources where at least one of the VIKING bands had no measurement at all (`band_flux = −999`); at GAMA depth this is however a small number, ∼3%, of all the objects. No extinction corrections nor zero-point offsets were applied in this test phase. In the near future, once joint optical – near-IR photometry becomes available for KiDS sources, also outside the GAMA regions, these experiments will be extended. In particular, we expect the photo-zs derived from KiDS + VIKING to improve over what is presented in Sect. 5 thanks to incorporating VIKING GAaP magnitudes, zero-point calibrated and extinction-corrected in the same manner as the KiDS *ugri* measurements.

3.3. WISE

In the GAMA-depth experiments, we also used date from the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010), which cover the full sky in four mid-IR bands ($W1$–$W4$) ranging from 3.4 μm to 23 μm. WISE is the most sensitive in its two shorter-wavelength channels, $W1$ (3.4 μm) and $W2$ (4.6 μm), reaching respectively 54 μJy and 71 μJy (5σ), which in $W1$ is equivalent to ∼21 mag in the AB system. The public WISE catalogue[10] is however limited to sources with a 5σ detection in at least one band. Therefore, rather than using that dataset, which is very incomplete even at GAMA depth (Cluver et al. 2014; Jarrett et al. 2017), we employed the GAMA-LAMBDAR catalogue which includes forced-photometry WISE flux measurements for all the GAMA sources in the equatorial fields.

Because of the much lower sensitivity of the $W4$ (23 μm) channel than the three others, it has a very high number of non-detections (`W4_flux = 0`) even in the LAMBDAR catalogue and will not be used. Also the $W3$ band (12 μm) has a considerable number of measurements lacking (17%), so part of our experiments employing WISE use either the $W1 + W2$ bands or $W1 + W2 + W3$. At present such WISE forced photometry for KiDS sources is not available, so these tests were limited only to the GAMA depth (Sect. 5) and cannot currently be extended beyond that. We are planning to obtain WISE measurements for a subsample of KiDS sources, but this will be limited to the bright end of the latter survey because of its much larger depth (cf. Lang et al. 2016b).

3.4. Spectroscopic: compilation of various datasets

As any other ML photo-z tool, ANNz2 and MLPQNA used in this study require training sets of sources from the target

photometric sample which have also spectroscopic redshifts measured. Empirical photo-z methods perform optimally if the training set is representative of the target data. Ideally, the former should be a random subset of the latter to provide the same distributions in magnitudes, colours, and redshift. However, even if this ideal setup cannot be met, ML will perform well as long as the important parameters such as magnitudes span the same range in training and target data, especially if some weighting is applied on the training data to mimic the target set.

On the other hand, MLMs usually do badly in extrapolating; for instance, training on a bright subset of much deeper target data is likely to give very biased results at the faint end. In addition, it must be remembered that ML photo-zs usually perform best at the median redshift (where they should provide practically zero bias), and by construction they tend to overestimate the redshifts at low z and underestimate them at high z (for example Bilicki et al. 2014). On the other hand, if applied properly, MLMs should give unbiased redshift as a function of $z_{\rm phot}$ in a sense that $\langle z_{\rm spec}|z_{\rm phot}\rangle = z_{\rm phot}$, which is not necessarily the case for template-fitting approaches.

In modern deep photometric surveys we hardly ever have spectroscopic subsets that are sufficiently representative for photo-z training at the full depth (for example Sánchez et al. 2014; Masters et al. 2015; Beck et al. 2016) and the situation will get worse with planned campaigns such as LSST or Euclid (cf. Newman et al. 2015), especially when one takes into account the requirements that photo-zs must meet in order not to heavily degrade cosmological constraints (Ma et al. 2006).

In the case of KiDS, the original footprint was optimised to first cover four GAMA fields as well as the COSMOS area. Of these, only the latter offers spectroscopy at a depth comparable to KiDS photometric data. On the other hand, the whole KiDS footprint is covered by either SDSS or 2dFLenS spectroscopic observations (see below), and these two samples have very similar properties in terms of their target selection for spectroscopy. Although very useful as a part of the overall training set, neither of these reach the full KiDS depth, and both offer only sparse sampling of colour-preselected objects (mostly luminous red galaxies, LRGs) beyond the local volume of $z < 0.1$. There are however several deep spectroscopic fields in the southern sky, and for the purpose of extending our spectroscopic calibration data, we have either included external measurements or asked for dedicated observations of some of them, as discussed in Hildebrandt et al. (2017).

Below we provide details of the spectroscopic data integrated into the training and calibration set used in this study. Their basic properties are summarised in Table 1 and their redshift distributions are shown in Fig. 1. All the spec-z samples had appropriate redshift quality cuts applied to preserve only science-grade measurements. Cross-matches between KiDS photometric sources and the spectroscopic objects were done using a 1″ matching radius.

3.4.1. GAMA

Galaxy And Mass Assembly (GAMA, Driver et al. 2011) is a spectroscopic survey of five fields, which employed the AAOmega spectrograph on the Anglo-Australian Telescope, with targets selected mostly from the Sloan Digital Sky Survey (SDSS), as well as from other surveys, including KiDS. It spans 3 equatorial fields (G09, G12 and G15) and two southern ones (G02 and G23) of which only G02 is outside the KiDS footprint. GAMA is 98.5% complete spectroscopically for SDSS galaxies with $r_{\rm Petro} < 19.8$ mag in the equatorial fields, and 94.2%

---

[9] We did not have to convert the fluxes to any magnitude system, because ML photo-z methods are agnostic to physical units. What matters is that each particular photometric parameter is measured self-consistently. This is a useful advantage of these methods over the SED-fitting ones.
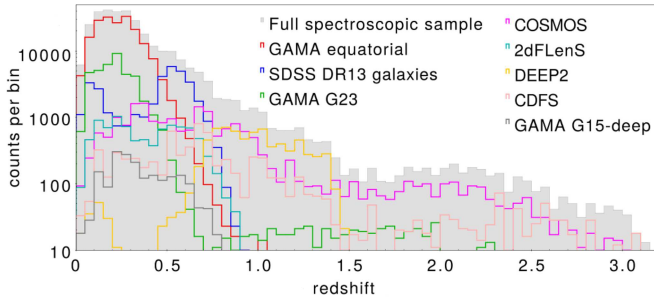
[10] Available from http://irsa.ipac.caltech.edu/Missions/wise.html

Table 1. Spectroscopic samples constituting the KiDS DR3 photo-z training set.

Sample	Number of sources[a]	Mean z[a]	Mean r mag[a]	Reference(s)
GAMA-II equatorial	190 741	0.234	19.5	Liske et al. (2015)
SDSS DR13 galaxies	56 911	0.349	19.6	Albareti et al. (2017)
GAMA G23	38 854	0.238	19.3	*proprietary*
zCOSMOS[b]	25 888	0.813	22.2	*private comm.* & Davies et al. (2015)
2dFLenS	11 873	0.362	19.6	Blake et al. (2016)
DEEP2 DR4 (two fields)	8924	0.962	23.5	Newman et al. (2013)
CDFS[c]	7044	0.846	22.9	*online* & Cooper et al. (2012)
GAMA G15-deep	2286	0.340	21.1	*proprietary*
Total[d]	312 501	0.335	19.9	
Total cleaned[e]	278 946	0.332	19.9	

Notes. [a] After cross-match with KiDS DR3, without masking nor quality cuts in KiDS. [b] Data from zCOSMOS public and non-public catalogues, as well as from the GAMA-G10 catalogue. [c] Data from GOODS/CDF-S compilation and from ACES. [d] Duplicate entries removed. [e] After cleaning of bad photometry as described in Sect. 3.5.

Fig. 1. Redshift distribution of the full KiDS DR3 spectroscopic training sample and of particular datasets included. The histograms show sources with 4-band *ugri* photometry in KiDS or in auxiliary datasets outside the nominal footprint.

complete for KiDS galaxies to $i < 19.2$ mag in G23 (Liske et al. 2015). Some of the measured sources are however fainter, and there additionally exists an unpublished catalogue of deeper observations in the G15 field (2300 sources of good redshift quality, with $\langle z \rangle = 0.34$) which we also use here.

These four fields give us in total almost 230 000 KiDS sources with GAMA spectroscopic redshift measurements, and their $\langle z \rangle = 0.23$. This, together with the excellent spectroscopic completeness of GAMA and no colour preselection therein other than star and quasar removal, makes GAMA the photometric redshift calibration set at the bright end of KiDS. Indeed, we will devote Sect. 5 to a GAMA-depth analysis, where GAMA spec-zs were used to calibrate KiDS ML photo-zs with excellent accuracy and precision.

3.4.2. SDSS

The Sloan Digital Sky Survey (SDSS, York et al. 2000) is a photometric and spectroscopic survey of $\sim\pi$ steradians of the northern sky, performed from the Apache Point Observatory in New Mexico, USA. SDSS is currently in Stage IV of its operations (Blanton et al. 2017) and we use its spectroscopic sources from the Data Release 13 (DR13, Albareti et al. 2017) which encompasses and supersedes all the earlier releases.

SDSS overlaps with KiDS in the equatorial fields above $\delta = -3°$. From the SDSS spectroscopic dataset, we only use sources with class "GALAXY", and do not include those which are "QSO", as training on the latter might bias the photo-zs. We

verified that it is indeed the case: training with SDSS QSOs included gives slightly worse overall results than if they are not used (but see Soo et al. 2018). There are almost 57 000 SDSS DR13 spectroscopic galaxies with KiDS DR3 photometric measurements, however those with $r < 19.8$ are mostly included in GAMA, and eliminating them gives about 43 000 unique KiDS × SDSS galaxies. While the full SDSS-matched sample has a mean redshift of only $\langle z \rangle \sim 0.35$, those that remain after removal of GAMA are at much higher redshifts, $\langle z \rangle \sim 0.71$. This is mostly thanks to the completed Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al. 2013) and first data from the extended BOSS (eBOSS, Dawson et al. 2016), both targeting preselected higher-z galaxies. A caveat is that these are mostly LRGs, which are not representative of the whole population and could bias the photo-zs if used as the sole calibration sample (Rozo et al. 2016). In our analysis we employ them as part of the overall training set, and the spec-z sample weighting applied in the photo-z derivation procedure should mitigate the related effects of an unevenly populated colour space.

3.4.3. 2dFLenS

The 2-degree Field Lensing Survey (2dFLenS, Blake et al. 2016) is a spectroscopic survey conducted at the Australian Astronomical Observatory between September 2014 and January 2016, covering an area of 731 deg^2 principally located in the KiDS regions. By expanding the overlap area between galaxy redshift samples and gravitational lensing imaging surveys, 2dFLenS aims to facilitate the joint analysis of lensing and clustering observables including all cross-correlation statistics (for example Joudaki et al. 2018), and to assist with photo-z calibration by direct training methods (Wolf et al. 2017) and by cross-correlation (Johnson et al. 2017). The 2dFLenS spectroscopic dataset contains two main target classes: $\sim$40 000 LRGs across a range of redshifts $z < 0.9$, selected by SDSS-inspired cuts, and a magnitude-limited sample of $\sim$30 000 objects in the range $17 < r < 19.5$.

In KiDS DR3 we have almost 12 000 2dFLenS galaxies, of which 9000 are unique (after excluding sources in common with SDSS and GAMA). The mean redshift of 2dFLenS, after eliminating the SDSS and GAMA overlap, is $\langle z \rangle \sim 0.39$. As in the case of SDSS, a caveat of using the 2dFLenS sources for photo-z training is that outside the local volume they are mostly LRGs.

3.4.4. zCOSMOS

The COSMOS field, centred roughly at $\alpha = 150°$, $\delta = 2.2°$, is currently one of the most comprehensively sampled areas in terms of deep spectroscopy. The original KiDS footprint was designed to cover 1 deg^2 of this field, so the photometric data here come from the main KiDS pipeline. For the photo-z experiments, we joined two main spectroscopic datasets in this field. The first one is a non-public dataset from the zCOSMOS team, that is deeper than the public release (Lilly et al. 2009), kindly shared by the zCOSMOS team. It incorporates spectroscopic data from various other observational campaigns in this field. After cleaning of bad-quality redshifts, this catalogue includes almost 28 000 sources, of which over 19 000 have a counterpart in KiDS-DR3 with $\langle z \rangle = 0.87$.

We supplement this catalogue with a GAMA-team reanalysis of public COSMOS data, dubbed G10 (Davies et al. 2015), which includes almost 24 000 spectroscopic measurements of appropriate quality. As there is large overlap between the zCOSMOS and G10 samples, we removed the duplicates, and eventually were left with about 6700 unique sources from a G10 cross-match with KiDS, of $\langle z \rangle = 0.61$, which were added to zCOSMOS.

The two samples together give about 25 900 sources with KiDS measurements, of which 21 100 have all four $ugri$ bands measured. These data have $\langle z \rangle = 0.71$ but span up to $z = 3$ (Fig. 1) which makes them crucial for photo-z calibration at the high-z tail of KiDS.

3.4.5. CDFS

The Chandra Deep Field South (CDFS), centred at $\alpha \simeq 53.1°$, $\delta \simeq -27.8°$, is another area surveyed by VST that has deep spectroscopy available. Unlike COSMOS, however, it is located outside the KiDS footprint and the photometric data we use here come from a KiDS-like reduction of VST imaging from the VOICE project (Vaccari et al. 2016). As in the zCOSMOS case, here the spectroscopic data were also composed of two datasets: an ESO-released compilation of GOODS/CDFS spectroscopy[11], including about 5600 sources with "secure" or "likely" redshifts (of which 3500 with KiDS measurements, $\langle z \rangle = 1.04$), supplemented with data from the Arizona CDFS Environment Survey (ACES, Cooper et al. 2012; 6400 with quality flag ≥ 3, of which 4440 in KiDS, $\langle z \rangle = 0.59$).

After removing duplicates we have 7000 spec-zs in the CDFS area, of which 5600 with all four bands available. This sample is slightly deeper on average ($\langle z \rangle = 0.74$) than KiDS-zCOSMOS, but has much fewer spectroscopic sources; it however also spans to large redshifts of $z \sim 3$ (Fig. 1), which makes it equally important for photo-z calibration and helps mitigate sample variance related to using very small areas for this purpose.

3.4.6. DEEP2

The DEEP2 Galaxy Redshift Survey (Newman et al. 2013) covers 2.8 deg^2 in four patches and is colour-selected in a way to target high redshift ($z \sim 1$) galaxies. Although not appropriate for photo-z calibration on its own, it is very useful when joined with the other samples, adding data in the $0.5 < z < 1.5$ range.

Two of the DEEP2 fields are within reach of VST and we have KiDS-like observations for them: the 2h field, centred at $\alpha \simeq 37.2°, \delta \simeq 0.5°$, and the 23h field at $\alpha \simeq 352.0°, \delta \simeq 0.0°$.

---

[11] http://www.eso.org/sci/activities/garching/projects/goods/MasterSpectroscopy.html

There are over 16 000 DEEP2 sources with ZQUALITY ≥ 3 in there, of which some 9000 have KiDS-like measurements. Among these, 7100 have measurements in all the four $ugri$ filters, with $\langle z \rangle = 0.97$, but limited almost entirely to $0.6 \lesssim z \lesssim 1.4$.

3.5. Properties of the photo-spectro compilation

In total we have over 310 000 sources with good-quality spectroscopic redshift measurements available for KiDS DR3. However, for these to be applicable as a photo-z training set, the data had to be cleaned of bad photometry as discussed in Sect. 3.1. We also required $z > 0.001$ to avoid residual stellar contamination and local volume galaxies with a possibly significant contribution of peculiar velocity to measured redshift. After these cuts, the full DR3 spectroscopic set used in this paper includes almost 280 000 objects. We reiterate though that this sample, having $\langle z \rangle = 0.33$, is dominated by GAMA with $z < 0.6$, and at higher redshifts it is very limited – see Fig. 1 and Table 1 for details. We would also like to emphasise that, what is a general problem for photo-z estimation and calibration, deep spectroscopic surveys preferentially measure redder galaxies. This is also the case for our training compilation beyond the GAMA depth, where it includes mostly red objects, unlike the target data, dominated by blue galaxies at the faint end (the "faint blue galaxy problem", Ellis 1997).
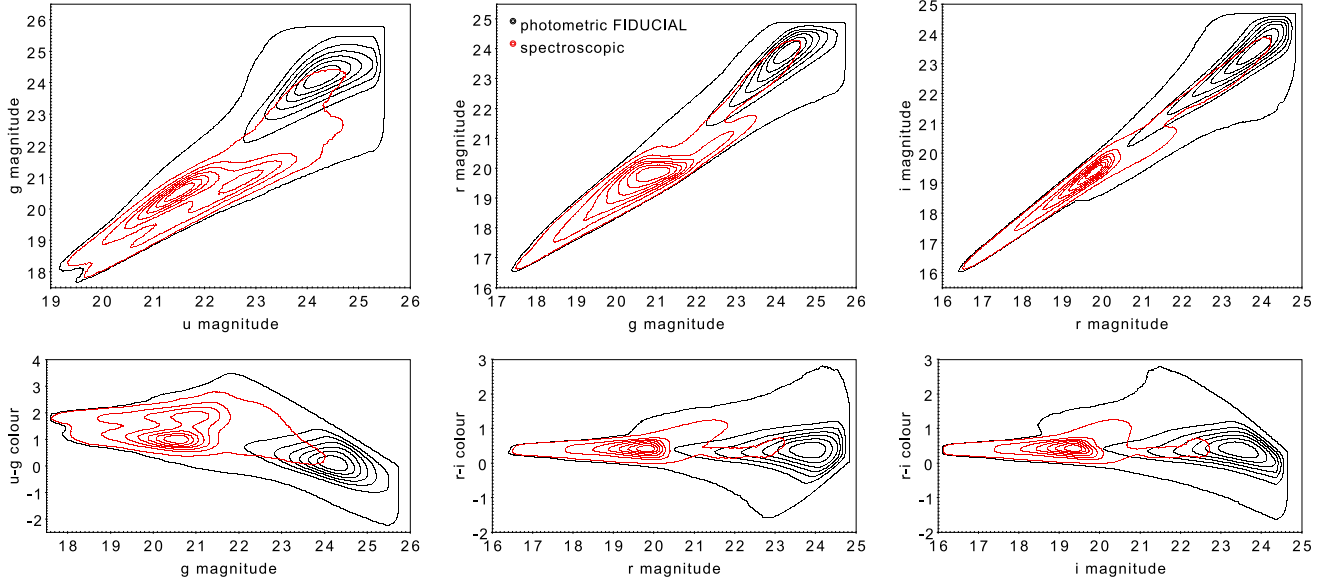
We illustrate the non-representativeness of our spectroscopic data in Fig. 2, which compares selected magnitude-magnitude (top) and magnitude-colour (bottom) distributions for the spectroscopic (red) and photometric (black) data. For the latter we show the FIDUCIAL sample (as defined in Sect. 3.1), which is the one used as reference for weighting the training set employed for the full-depth DR3 photo-z catalogue (Sect. 4.4). Clearly, both in magnitude and colour space of KiDS DR3 there are regions not well sampled by the current spec-z data. This issue cannot be fully overcome without adding further deep and appropriately preselected spectroscopic data to the calibration sample (Masters et al. 2017), although we mitigate its importance by the aforementioned weighting using the kNN procedure (Lima et al. 2008) implemented in ANNz2. On the other hand, as far as weak lensing analyses using KiDS data are concerned, the objects that are missing in the overlapping spec-z samples are mostly faint galaxies at high redshift, which are unresolved by KiDS and are thus either not included or are heavily downweighted when measuring lensing shear.

4. KiDS DR3 experiments and associated photometric redshift catalogue

In this Section we quantify the performance of ML photo-zs in KiDS DR3, and compare them to the pipeline solution from BPZ. This is done by running several photo-z experiments in which we applied ANNz2 and MLPQNA to different training and test subsets of the KiDS DR3 spectro-photo compilation presented above. We also describe the publicly released photo-z catalogue derived with ANNz2, which includes all the DR3 sources that have the four $ugri$ bands measured (39.2 million objects). An earlier version of this catalogue was made available with the DR3 publication (de Jong et al. 2017). Here we update that dataset and provide more details on its properties.

The tests below will be obviously limited to the spectroscopic data, so the conclusions based on them may not be easily extrapolated to the full photometric set. This is however a general truth in photo-z performance checks if incomplete spec-z

Fig. 2. *Top row*: comparison of magnitude distributions for the KiDS-DR3 photometric FIDUCIAL sample (black) and the spectroscopic redshift calibration dataset (red). *Bottom row*: similar comparison but for selected magnitude-colour planes. The contours are linearly spaced. The FIDUCIAL sources are used as the reference for weighting the spec-z training set in the derivation of photo-zs for the full catalogue (Sect. 4.4). See also Fig. 3 for colour-colour plots where weighting of the training is additionally illustrated.

samples are used as calibrators, as is the case for most of the modern photometric surveys (Hildebrandt et al. 2010). Due to the nature of spectroscopic campaigns, which either explicitly target or are more efficient at measuring spectra of red and intrinsically luminous galaxies, the colour space of spec-z samples is undersampled in some areas (Masters et al. 2015) which may lead to biases in direct comparisons of spectroscopic and photometric redshifts.

In what follows, by a test sample we will always mean data not used in the training and validation phase. We note that if both are selected randomly, the training and test samples will be statistically equivalent, so such tests will mostly tell how well the MLMs did for representative training data but not necessarily how well they do for the target photometric sample. We thus performed two types of experiments: (i) where the training and test data were statistically equivalent (Sects. 4.1–4.2), as well as (ii) those where the training and test samples were very different (Sect. 4.3); in the latter case, in some of the tests weighting was applied to the training data. Such comparisons with available spectroscopic redshifts do not however provide the full picture on photo-z performance due to biases in the calibration data such as their preference for red galaxies over blue ones, and limited depth. Therefore, in Sect. 4.4 and Appendix B we also analyse output photo-z redshift distributions of the target photometric sample.

The performance of the photo-zs will be measured using the following statistics:
- bias, $\langle \delta z \rangle = \langle z_{phot} - z_{spec} \rangle$, unclipped;
- normalised bias, $\langle \delta z/(1 + z_{spec}) \rangle$, unclipped;
- standard deviation of normalised error, $\sigma_{\delta z/(1+z_{spec})}$, unclipped;
- scaled median absolute deviation of normalised error, $\mathrm{SMAD}\left(\delta z/(1 + z_{spec})\right)$, where $\mathrm{SMAD}(x) = 1.4826$ median $(|x - \mathrm{median}(x)|)$;
- percentage of catastrophic outliers for which $|\delta z/(1 + z_{spec})| > 0.15$; we use this particular definition of outliers to be consistent with other KiDS photo-z analyses (Kuijken et al. 2015; de Jong et al. 2015, 2017).

For non-Gaussian distributions which usually characterise photo-z errors, the unclipped scatter is not always informative, and SMAD, converging to the standard deviation (SD) for Gaussians, is preferred as the measure of the actual scatter. We also provide the SD as its comparison with SMAD helps judge how non-Gaussian the distribution is.

The statistics for MLM results will be computed for the test sets unseen by the algorithm in the training phase. They will also be compared to the results from the fiducial KiDS photo-z solution, BPZ, which is independent of any training; for consistency, in such comparisons, we will use exactly the same test sets for the MLM and BPZ cases. The BPZ statistics will be based on the central Z_B values only. In the case of ANNz2, we use the unweighted MLM-average (`ANNZ_MLM_avg_1`) generally found to perform best among the five types of point estimates from this software (Sect. 2.1). For MLPQNA, we use the output of the regression network without any further manipulation.

4.1. Random subsample of the spectroscopic data

In the first experiment we chose a random subsample (1/3) of the full spectroscopic data for training and validation and used the remaining 2/3 as a blind test set. We have checked that the exact proportions of this split do not have a large importance for the results, provided that there are enough sources both in training and test samples to guarantee good statistics. The results for this test, compared with BPZ, are provided in the top rows of Table 2. Except for the normalised bias, both ANNz2 and MLPQNA clearly outperform BPZ for this low-z dominated sample, the two ML approaches having statistics very comparable between each other. We have to note that in this case, the test data had the same properties as the training set, which means that this particular experiment shows only the performance of the MLMs in an ideal setup of the training being fully representative for the target data, which is not the case in KiDS. This experiment is thus mostly useful to judge the performance of the methods for the bright end of the sample. See also

Table 2. Statistics of photometric redshift performance obtained for KiDS DR3 experiments with ANNz2 and MLPQNA vs. BPZ.

Sample	Method	Mean of $\delta z = z_{\rm ph} - z_{\rm sp}$	Mean of $\delta z/(1 + z_{\rm sp})$	St.dev. of $\delta z/(1 + z_{\rm sp})$	SMAD[a] of $\delta z/(1 + z_{\rm sp})$	% of outliers $|\delta z|/(1 + z_{\rm sp}) > 0.15$
Random subsample	ANNz2	-3.3×10^{-3}	3.3×10^{-3}	0.073	0.026	3.5%
$\langle z \rangle = 0.332$	MLPQNA	-2.0×10^{-3}	3.9×10^{-3}	0.079	0.026	3.4%
	BPZ[b]	-1.9×10^{-2}	-1.5×10^{-3}	0.089	0.035	4.1%
Random 10% of $r < 20$,	ANNz2	-2.4×10^{-3}	8.3×10^{-3}	0.102	0.034	7.1%
all from $r \geq 20$,	MLPQNA	-3.2×10^{-3}	7.2×10^{-3}	0.116	0.034	7.4%
$\langle z \rangle = 0.489$	BPZ[b]	-5.8×10^{-2}	-1.9×10^{-2}	0.120	0.042	8.4%
Trained w/o COSMOS,	ANNz2	-4.4×10^{-2}	1.2×10^{-2}	0.183	0.091	25.0%
tested on COSMOS	ANNz2w[c]	-6.7×10^{-2}	-4.6×10^{-4}	0.184	0.086	22.7%
$\langle z \rangle = 0.784$	MLPQNA	-8.0×10^{-2}	-2.7×10^{-3}	0.204	0.086	23.6%
	BPZ[b]	-2.4×10^{-1}	-8.5×10^{-2}	0.195	0.085	24.5%
Trained w/o CDFS,	ANNz2	3.0×10^{-2}	5.2×10^{-2}	0.232	0.108	25.7%
tested on CDFS	ANNz2w[c]	3.9×10^{-2}	5.4×10^{-2}	0.206	0.101	26.0%
$\langle z \rangle = 0.742$	MLPQNA	1.0×10^{-2}	3.8×10^{-2}	0.222	0.100	25.8%
	BPZ[b]	-1.9×10^{-2}	-7.2×10^{-2}	0.183	0.083	23.7%

Notes. Results for the particular tests are provided in blocks of rows. See text for details. [a] SMAD is the scaled median absolute deviation, converging to standard deviation for Gaussian distributions. [b] BPZ is independent of the training sets – the numbers are given for comparison (for the same test samples). These statistics are based on the KiDS pipeline solution. [c] Training data weighted with the kNN method, weights propagated throughout the training and evaluation procedure.

Cavuoti et al. (2015b) and Amaro et al. (2018) for a more detailed discussion of how MLPQNA performs in this regime, as well as Sect. 5 of this paper for a dedicated study of ANNz2 performance at the bright end of KiDS.

4.2. Downweighting the bright end

As the training set is dominated by bright galaxies (cf. Fig. 2), in the second step we constructed a sample in which we artificially down-weighted the bright end. This was done by randomly selecting 10% of the bright-end ($r < 20$) sources from the full KiDS spectro-photo compilation, while keeping all the objects with $r \geq 20$. The subsampling percentage was chosen to obtain the mean redshift of the joint sample in between that of the fully random one from Sect. 4.1 and those of the COSMOS and CDFS datasets analysed in Sect. 4.3. This procedure gave us a joint sample of 118 000 galaxies with $\langle z \rangle = 0.49$ and $\langle r \rangle = 21$ mag. This dataset was again divided into training and test sets in proportions 1:2. Photo-z statistics are provided in Table 2, second set of rows. In this case all the computed statistics for ANNz2 and MPLQNA are better than for BPZ, and the two empirical methods gave results very comparable to each other.
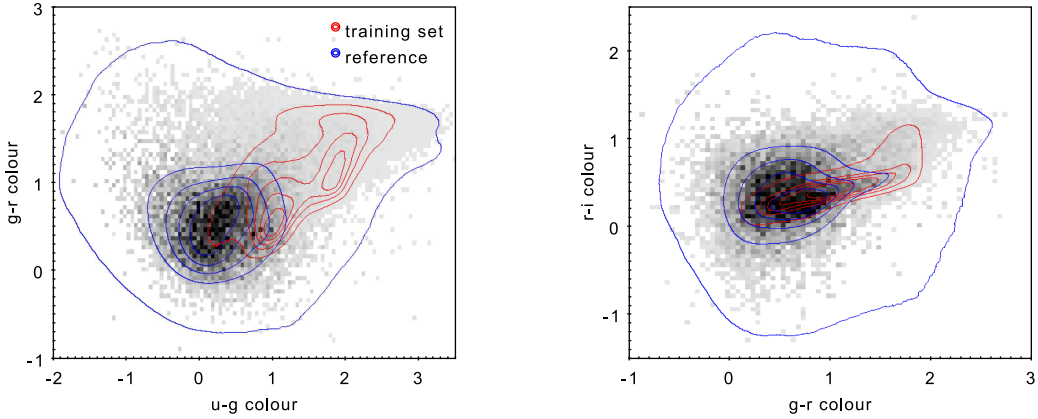
4.3. COSMOS and CDFS as independent test samples

The most informative approach to judge the performance of KiDS ML photo-zs is to use separate deep training end test data. Therefore, as a next step, we trained ANNz2 and MLPQNA on KiDS spectroscopic sources from outside the COSMOS field and tested the results on KiDS-COSMOS spec-z data; then we repeated the exercise this time with CDFS (train excluding CDFS, test on KiDS-CDFS). This way the test sets were fully independent from the training ones, and had very different characteristics. On the other hand, these two target samples have similar mean redshifts of $z \sim 0.75$, closer to what we expect from the full KiDS than the mean redshift of the current spectroscopic calibration data would suggest. Therefore, these experiments provide the most insight into the true performance of the photo-z

methods at the full depth available from spec-z samples overlapping with KiDS.

In the case of the ANNz2 experiments, two approaches were taken: in the first one we trained on a random subsample of non-COSMOS/non-CDFS data (respectively 10% and 3%) without any weighting; in the second one we trained on all the non-COSMOS/non-CDFS data but this time weighting the training sample in GAaP *ugri* magnitude space with the kNN method (as implemented in the ANNz2 code) to mimic the properties of the target COSMOS/CDFS data, respectively. These weights were then used in the whole photo-z procedure. The reason for taking just a small random subsample for the no-weighting experiments was that otherwise there would be a huge, unrealistic imbalance between the size of the training and test sets; the subsampling percentages used made the training and test sets comparable in size. On the other hand, in the weighting case, the weights for most of the training objects were much smaller than unity, so the effective weighted number of training sources was also comparable to the target set sizes. For MLPQNA, the experiments had the same setup as ANNz2 without any weighting.

The results of these experiments are compared in the two bottom set of rows of Table 2. If no weighting is applied, then both MLPQNA and especially ANNz2 perform worse than BPZ in terms of scatter, but better in terms of bias. Weighting does improve the ANNz2 results, although not significantly; in the COSMOS case, they provide similar scatter to the BPZ case while still have much smaller bias. For CDFS, MLPQNA performed generally better than both the unweighted and weighted ANNz2 experiments, but the scatter from both ML approaches remains visibly worse than from BPZ. The large fraction of outliers for these two deep comparison datasets is partly due to how these outliers were defined, namely with respect to a fixed normalised error value of 0.15. For BPZ, these results are consistent with what was shown in Kuijken et al. (2015) where test samples of similar depth as in here were used (CDFS and non-public zCOSMOS). On the other hand, de Jong et al. (2017) used a shallower public zCOSMOS sample and consequently found a smaller outlier fraction both for BPZ and ANNz2.

Fig. 3. Illustration of the kNN weighting procedure applied to the training data for the KiDS DR3 photo-zs, as projected to colour-colour planes. Red contours show the unweighted spectroscopic training data, while the blue ones are for the reference photometric sample (FIDUCIAL). The greyscale pixels show the distribution of the weighted spectroscopic sample. The contours are linearly spaced.
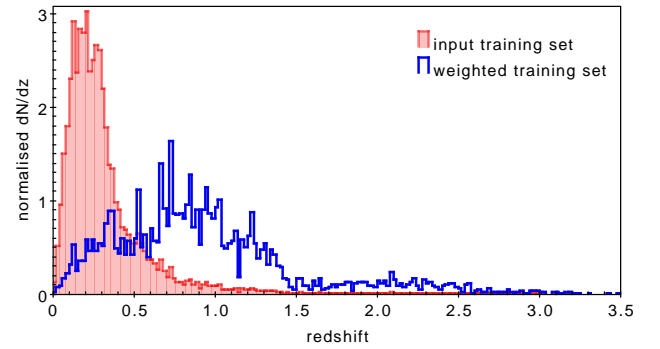
4.4. KiDS DR3 ANNz2 photo-z catalogue release

Having performed the above tests, we used the full KiDS-matched spectroscopic sample as the training+validation set to train ANNz2, and produced the full-depth DR3 photo-z catalogue, originally released with the DR3 paper (de Jong et al. 2017), and now updated[12]. This catalogue includes all the 39.2 million DR3 sources that have the full set of *ugri* bands measured, but only part of them will have photo-zs of sufficient quality to be considered reliable. Below we quantify the performance of these ML photo-zs.

In the whole photo-z procedure we used the kNN weighting of the training data, as implemented in ANNz2 (Sect. 2.1), applied in the *ugri* magnitude space. The reference dataset was the FIDUCIAL sample described in Sect. 3.1, constructed in such a way to include only likely galaxies and encompass magnitude ranges of the training data. Fig. 3 compares the 2D contours of the training sample (red) in colour space to those of the reference FIDUCIAL dataset (blue), and to the weighted distribution of the spec-z sources illustrated as background greyscale pixels. Fig. 4 shows the unweighted (red) and weighted (blue) input spectroscopic redshift distributions of the training set. The latter, of weighted $\langle z^{(\mathrm{w})} \rangle = 0.93$, can be regarded as a proxy for what should be expected from the true redshift distribution of the target sample (Hildebrandt et al. 2017; Soo et al. 2018), although at high redshifts $z > 1.5$ this might be just a crude approximation due to sample variance in the very limited calibration data (Fig. 1).

In de Jong et al. (2017) we described an earlier version of the KiDS DR3 ANNz2 photo-z catalogue, for which 100 ANNs were used in the training phase. Here we update that catalogue, having found a small issue in selecting the FIDUCIAL sources for weighting the spectroscopic sets. The changes are very small and all the conclusions from de Jong et al. (2017) regarding the performance of ANNz2 photo-zs remain valid; the catalogue is updated for consistency.

Figure 5 summarises the properties of the KiDS DR3 ANNz2 photo-zs as compared to spec-zs from the datasets overlapping with the DR3 footprint (that is, a subset of the full training sample, excluding CDFS and DEEP2). They show that the photo-zs are stable for $z_{\mathrm{spec}} \lesssim 0.9$ and $z_{\mathrm{phot}} \lesssim 0.9$, as well as for $r \lesssim 23.5$, above which their quality quickly deteriorates. These could be

---

[12] Data available from http://kids.strw.leidenuniv.nl/DR3/ml-photoz.php#annz2
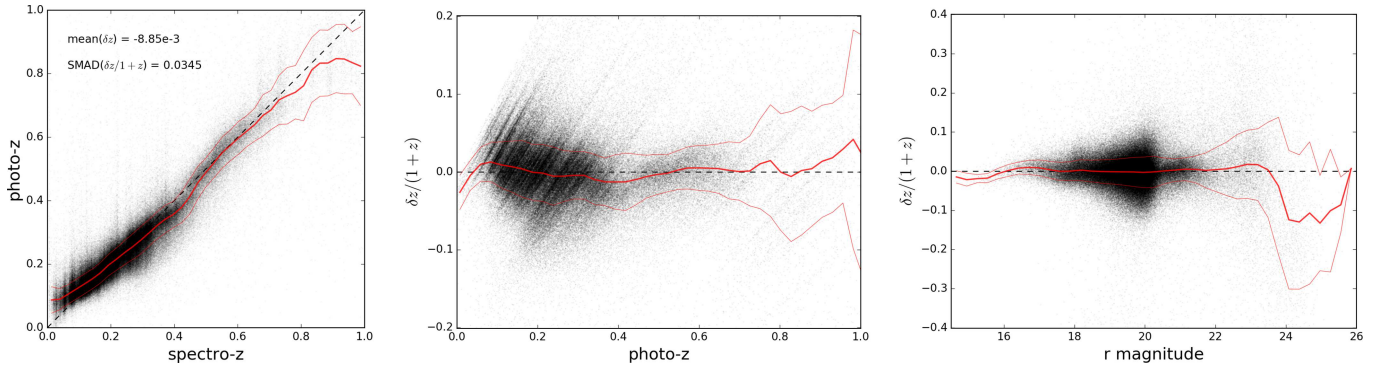

Fig. 4. Comparison of spectroscopic redshift distributions of the unweighted training set (red) and after applying the kNN weights to it (blue). The weights were derived with reference to the KiDS DR3 FIDUCIAL dataset, and subsequently used in ANNz2 training and evaluation for the public KiDS DR3 photo-z data release. Histograms are normalised to unit area.

then considered the limits up to which the presented here ANNz2 photo-zs are relatively reliable. In de Jong et al. (2017) the performance of photo-zs was illustrated as a function of spec-z and the r-band magnitude, but for shallower calibration samples than here (public GAMA DR2 and public zCOSMOS). Here we prefer to focus on the error behaviour as a function of magnitude and photo-z, as these latter are the quantities available to the end user of the catalogue. Table 3 quantifies this performance in bins of photo-z, for both ANNz2 and BPZ (binning is done in the respective photo-z type). The statistics were derived using the same overlapping spec-z samples as employed for Fig. 5, which become very incomplete at $z \sim 1$. At present there are no sufficiently deep and complete spectroscopic datasets available in the KiDS footprint that would allow for a reliable quantification of photo-z performance at the full depth of the survey.

We have also verified that both ANNz2 and BPZ perform better for red galaxies than for blue ones. For instance, if we split the overlapping spectroscopic sample according to the colour-colour line $g - r = 0.8 - 0.8(r - i)$, then sources redwards of this division have $\delta z_{\mathrm{BPZ}} = 0.015$, $\mathrm{SMAD}(\delta z_{\mathrm{BPZ}}/(1 + z)) = 0.030$ for BPZ and $\delta z_{\mathrm{ANNz2}} = -8.0 \times 10^{-3}$, $\mathrm{SMAD}(\delta z_{\mathrm{ANNz2}}/(1 + z)) = 0.032$ for ANNz2, while those on the blue side have $\delta z_{\mathrm{BPZ}} = -0.098$, $\mathrm{SMAD}(\delta z_{\mathrm{BPZ}}/(1 + z)) = 0.049$ and $\delta z_{\mathrm{ANNz2}} = -0.014$, $\mathrm{SMAD}(\delta z_{\mathrm{ANNz2}}/(1 + z)) = 0.050$. Similar worsening when going from red to blue is also observed for other statistics as well

Fig. 5. Performance of the KiDS DR3 ANNz2 photo-*z*s from the released catalogue as compared to the overlapping spectroscopic samples. *Left-hand panel*: direct spec-*z*–photo-*z* comparison; central panel: photo-*z* error as a function of photo-*z*; *right-hand panel*: photo-*z* error as a function of the *r*-band magnitude. The thick solid line shows the running median while the thin lines encompass the scatter (SMAD). Note different scalings of the $\delta z/(1 + z)$ axes. Based on these comparisons, we judge the published photo-*z*s to be reliable within $z_{\rm phot} < 0.9$ and $r < 23.5$.

Table 3. Statistics of photometric redshift performance for the released KiDS DR3 catalogue, as obtained from a comparison with overlapping spectroscopic redshifts.

photo-*z* bin	Method	Mean of $\delta z = z_{\rm ph} - z_{\rm sp}$	Mean of $\delta z/(1 + z_{\rm sp})$	St.dev. of $\delta z/(1 + z_{\rm sp})$	SMAD of $\delta z/(1 + z_{\rm sp})$
$0.0 < z_{\rm phot} \leq 0.1$	ANNz2	3.0×10^{-3}	6.2×10^{-3}	0.045	0.025
	BPZ	-7.8×10^{-2}	-4.4×10^{-2}	0.109	0.040
$0.1 < z_{\rm phot} \leq 0.2$	ANNz2	1.7×10^{-3}	3.9×10^{-3}	0.044	0.034
	BPZ	-2.9×10^{-3}	6.7×10^{-3}	0.069	0.032
$0.2 < z_{\rm phot} \leq 0.3$	ANNz2	-1.0×10^{-2}	-5.1×10^{-3}	0.048	0.035
	BPZ	3.5×10^{-3}	5.6×10^{-3}	0.044	0.028
$0.3 < z_{\rm phot} \leq 0.4$	ANNz2	-1.7×10^{-2}	-8.3×10^{-3}	0.055	0.032
	BPZ	1.0×10^{-2}	1.0×10^{-2}	0.043	0.028
$0.4 < z_{\rm phot} \leq 0.5$	ANNz2	-2.0×10^{-2}	-5.8×10^{-3}	0.079	0.032
	BPZ	4.5×10^{-2}	3.4×10^{-2}	0.044	0.035
$0.5 < z_{\rm phot} \leq 0.6$	ANNz2	-5.0×10^{-3}	3.0×10^{-3}	0.074	0.026
	BPZ	2.6×10^{-2}	2.0×10^{-2}	0.054	0.028
$0.6 < z_{\rm phot} \leq 0.7$	ANNz2	2.2×10^{-3}	9.3×10^{-3}	0.088	0.032
	BPZ	6.4×10^{-3}	1.1×10^{-2}	0.081	0.040
$0.7 < z_{\rm phot} \leq 0.8$	ANNz2	-9.2×10^{-3}	9.0×10^{-3}	0.123	0.057
	BPZ	-2.9×10^{-2}	-1.1×10^{-3}	0.118	0.057
$0.8 < z_{\rm phot} \leq 0.9$	ANNz2	-1.1×10^{-2}	1.4×10^{-2}	0.146	0.079
	BPZ	-1.5×10^{-1}	-3.4×10^{-2}	0.183	0.084
$0.9 < z_{\rm phot} \leq 1.0$	ANNz2	2.7×10^{-3}	3.8×10^{-2}	0.197	0.082
	BPZ	-4.8×10^{-1}	-1.4×10^{-1}	0.239	0.256

Notes. Results are presented per bin of ANNz2 and BPZ photo-*z*s, respectively.

as for differently defined red-blue separations. This general behaviour should not be surprising: the observed optical colours of red galaxies are a strong function of redshift while those of the blue ones much less depend on *z*. Regardless of the approach used, this means that photo-*z*s for blue galaxies are expected to be worse than for red ones. Let us however reiterate that red galaxies dominate at the faint end of our spectroscopic calibration sample. This means that our possibility to reliably quantify photo-*z* performance for faint blue galaxies is limited.

The limitations of the spectroscopic calibration data mentioned above mean that classic spec-*z*–photo-*z* comparisons do not give the complete picture on the performance of the latter. Therefore, a useful test, even if rather qualitative, is provided by the verification of photo-*z* distributions of target photometric samples. In Fig. 6 we first compare $\mathrm{d}N/\mathrm{d}z_{\rm phot}$ of the

FIDUCIAL sample, for BPZ (grey bars) and ANNz2 (blue line). We see that they are very different with the ANNz2 one being smooth and extending to high redshifts, while the BPZ $\mathrm{d}N/\mathrm{d}z$ shows several significant peaks, likely resulting from aliasing in colour-redshift space (that is, emission lines moving between the filters). The SED-fitting solution has here practically no redshifts beyond $z_{\rm BPZ} > 1.5$; many sources are instead assigned low $z_{\rm BPZ}$. This latter behaviour of BPZ is related to the prior which has been used in its implementation for KiDS purposes, directly propagated from an earlier CFHTLenS analysis (Hildebrandt et al. 2012), where the original prior from Benítez (2000) was modified to behave better for that survey. It was not optimised further for KiDS as the default redshift calibration in KiDS cosmological analyses is not based on individual redshift estimates but on external spectroscopic samples (the "DIR" method of
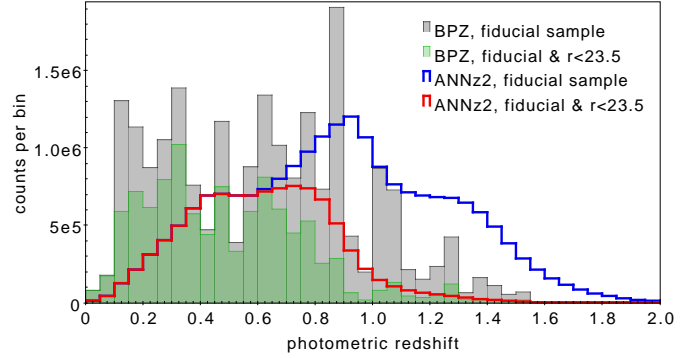
Hildebrandt et al. 2017). However, this prior is now being revised for new KiDS releases to provide also more reliable individual photo-zs from BPZ. As far as the abundance of low-z_{BPZ} sources is concerned, these are mostly galaxies with observed blue colours. More discussion on this can be found in Appendix B.

The ANNz2 redshift distribution in Fig. 6 is much more regular than that of BPZ, although probably not trustworthy beyond $z > 1$ or so (as discussed a couple paragraphs above), where practically all the sources have $r > 23$. We also observe a flattening of dN/dz_{ANNz2} at $z_{phot} \sim 0.5$. which may reflect worse performance of ANNz2 in this regime (cf. Fig. 5) and is probably related to the properties of the training set. Namely, at low redshifts the calibration data are dominated by the complete flux-limited ($r < 19.8$) GAMA sample. Its dN/dz quickly drops off at $z \sim 0.5$, beyond which the training is composed of various deeper but not as complete datasets (Sect. 3.4). Despite the weighting applied to the training data, this imbalance is apparently propagated into the photo-z solution. Part of the reason might be also under-performance of the weighting procedure, which in 4-dimensional parameter space could be prone to biases from the large-scale structure and noise, as evidenced by various peaks and dips in Fig. 4. We will be testing if these issues could be mitigated in future KiDS releases, which will include more spectroscopic calibration data and will be extended with VIKING near-IR measurements, providing thus nine-dimensional magnitude space.

In the same Fig. 6 we also show dN/dz_{phot} for the FIDU-CIAL sample but trimmed at $r < 23.5$ (green bars for BPZ and red line for ANNz2), which we have judged above to be the limit up to which the published ANNz2 redshifts are reliable. Indeed, we see that the main peak observed in the FIDUCIAL sample at $z_{ANNz2} \sim 0.95$ as well as most of the tail at $z > 1$ come from sources fainter than this magnitude cut, which is probably a sign of extrapolation. Interestingly, such a flux limit removes also several high peaks in the distribution of BPZ photo-zs, although the dN/dzs for the $r < 23.5$ sources remain very different between the BPZ and ANNz2 solutions. Noting here that we do not expect this particular flux-limited selection to provide improved z_{BPZ} over other cuts, as we show in Appendix B the main source of this persisting discrepancy seems to be very different treatment of blue galaxies by the two photo-z approaches. Namely, ANNz2 assigns them a flat and extended dN/dz while BPZ limits the output redshifts to a couple of rather narrow ranges. In particular, a significant fraction of blue galaxies are allocated to relatively low photo-zs ($z_{BPZ} < 0.4$) by the KiDS DR3 BPZ implementation. On the other hand, applying a colour cut on the sample to separate out redder galaxies allows us to largely mitigate the photo-z differences.

We note that until now, galaxies with photo-zs beyond $z_{BPZ} > 0.9$ have not been used for KiDS scientific analyses mostly due to the inability of their proper calibration at this high-z end. Forthcoming developments from using additional VIKING data as well as extending the spectroscopic training samples should provide the possibility of deriving better photo-zs at this range both using BPZ and MLMs, which is certainly of great interest for lensing studies.

As we showed in this Section, the currently derived ANNz2 photo-zs are, at least to $z_{phot} < 0.9$ and $r < 23.5$, of quality comparable to the default KiDS BPZ ones in terms of the overall statistics, and fare considerably better in terms of bias in most of the regimes, and also in terms of scatter at the bright, low-redshift end. In the near future we expect both the ML and template-fitting KiDS solutions to improve. For the ML case, extending the training sample is important, and will be made possible thanks to currently processed or ongoing KiDS-like

Fig. 6. Comparison of photometric redshift distributions for two KiDS DR3 samples: FIDUCIAL (see Sect. 3.1 for details), and FIDUCIAL with an additional cut of $r < 23.5$, for BPZ (bars) and ANNz2 (lines). The comparison suggests that for $r > 23.5$ the ANNz2 photo-zs in DR3 may not be trustworthy as they are based on extrapolation. The shape of the BPZ dN/dz is driven by the prior that was adopted (see text for details).

observations of some of the VVDS (Le Fevre et al. 2003) and VIPERS (Guzzo et al. 2014) fields. These will give additional calibration samples spanning redshifts $0 < z < 1.6$, which will help mitigate sample variance plaguing the derivation of high-z photo-zs.

Both the ML and SED-fitting methods will benefit from the major extension of photometry, namely the addition of five VIKING NIR bands. In the following Section we present the improvement possible thanks to adding the VIKING data to ANNz2 derivation at low redshifts, but our early tests show that similar gains should be also expected at larger depths. In future KiDS releases, starting from DR4, we plan to derive the ANNz2 photo-zs from nine-band KiDS+VIKING photometry employing the extended training data.

To summarise this comparison of the two photo-z solutions, it is clear that both have their limitations which the user should be aware of. Photo-z accuracy is expected to be a function of apparent magnitude, colour, and true redshift. This is inevitable given errors in photometry and the SEDs of galaxies – the optical colours of blue galaxies have a relatively weak dependence on redshift. We can mitigate these differences to some extent, for instance by weighting the training data in ML photo-z derivation, but not entirely remove their impact. As far as the ANNz2 photo-zs are concerned, they should be preferred in the range where sufficient spectroscopic training data are available, which is brighter and redder sources. Outside of this range, where MLMs suffer from biases incurred by extrapolation beyond the training coverage, the recommended solution is an SED-fitting one such as BPZ. However, the lack of sufficiently deep and complete, especially in terms of the blue population, calibration data overlapping with KiDS DR3 does not allow us to reliably quantify the performance of both these approaches at $z > 1$.

5. GAMA-depth experiments and associated photometric redshift catalogue

In this Section we analyse ANNz2 photo-z performance at the bright end of KiDS, and describe the associated catalogue release which includes 800 000 sources with $r < 20.3$. As already mentioned, ML photo-zs usually work best at the median redshift of the training set and tend to over-(under-)estimate redshifts at the low-(high-)z regime, that is, at the bright (faint) end

of the sample. This is also the case for KiDS DR3 where, partly due to the training set weighting to mimic the target data, the ANNz2 photo-z solution is not optimised for the bright end of the sample. Also the BPZ photo-zs calculated in the KiDS DR3 pipeline do not perform very well at the bright end, especially in terms of bias.

However, there is considerable interest in obtaining a KiDS dataset with well-constrained photo-zs in the relatively nearby Universe, as such a sample could be then used for such measurements as galaxy-galaxy lensing (for example Velander et al. 2014), general galaxy evolution studies (for example Tortora et al. 2016; Costa-Duarte et al. 2018) or for studying the effects of the cosmic web (for example Gruen et al. 2016). Indeed, this type of study have already been undertaken by the KiDS team, using spectroscopic data for the foreground sample, thanks to the (by-design) full overlap of KiDS with GAMA equatorial fields (Viola et al. 2015; Sifón et al. 2015; van Uitert et al. 2016, 2017; Brouwer et al. 2016; Dvornik et al. 2017).

The GAMA survey has however already finished, its overlap with KiDS ($\sim$200 deg^2) will thus not increase. Having a GAMA-like catalogue within the full planned KiDS coverage of $\sim$1500 deg^2 would make it possible to reduce the statistical errors of the aforementioned KiDS analyses by a factor of $\sim$2.5. We therefore examined what accuracy and precision can be obtained by training KiDS photo-zs on GAMA spec-zs, and how these could be improved by extending the parameter space with redshift-dependent measurements other than magnitudes, as well as by adding IR photometry. This is also of interest for ongoing and future photometric surveys such as the HSC or LSST, which will overlap with GAMA, but will provide even deeper photometry. An alternative route towards a KiDS foreground sample with well-constrained photo-zs could be via identifying LRGs, for instance using the "redMaGiC" algorithm (Rozo et al. 2016). This type of analysis is currently ongoing (Vakili et al., in prep.).

All the tests described hereafter will be using the ANNz2 software only, and will be restricted to the GAMA equatorial fields to guarantee very high completeness of the training data (except for the catalogue release of Sect. 5.4). Unlike in Sect. 4, here we kept the same training and test sets for all the experiments; what was varied were the parameters used in the photo-z derivation. We tested practically all the KiDS multi-band measurements from the DR3 public release that correlate with redshift, such as (observed) magnitudes, colours, angular sizes, and other related photometric parameters. In addition, when analysing various extensions to the basic KiDS $ugri$ magnitudes, we also took advantage of the availability of GAMA LAMBDAR catalogues (Wright et al. 2016), which include VIKING and WISE forced photometry measurements on GAMA targets. These extra features were first added "individually" to the basic $ugri$ setup (in relevant groups, for example magnitudes in a single fixed aperture but from all the bands) and once those providing the most amelioration had been determined, they were combined into multiparameter setup used at a further stage. Below we present the main results of these tests, focusing on those photometric measurements which brought the most improvement to the photo-zs over using only the default (GAaP) $ugri$ magnitudes.

We note that a possibly more optimal way of extending the parameter space would be to first apply a dimensionality reduction algorithm such as Principal Component Analysis (PCA) or related (for example Singal et al. 2011). This would remove redundancy from the feature space and speed up the training process. At this stage of data exploration, we preferred however to work directly with the parameters provided in the catalogues, to verify which among them are the most useful for photo-z

estimation; PCA might however blur such information. In future releases of KiDS data, where also the 5-band VIKING information will be added by default, and perhaps also WISE for the bright sources, the parameter space will grow considerably, so such pre-processing may indeed become necessary.

The usefulness of parameters other than magnitudes, especially the morphological ones, on photo-z estimation, has been studied by several authors (for example Wray & Gunn 2008; Way 2011; Singal et al. 2011; Hoyle et al. 2015; Jones & Singal 2017; Gomes et al. 2018), with mixed conclusions. We refer the reader to the recent "Morpho-z" analysis by Soo et al. (2018) and references therein for an overview of these earlier efforts. As far as we are aware, perhaps with the exception of the more recent analysis by Gomes et al. (2018), none of these previous studies of that kind operated at the same regime as ours here, namely for relatively bright galaxies with excellent photometry (very high signal-to-noise) and using a complete spectroscopic sample for photo-z calibration.

As far as the the recent results from Soo et al. (2018) are concerned, they found improvement in photo-zs only if morphology was used with a very limited set of passbands (as small as one in some cases), while for a more complete setup such as their fiducial $ugriz$, even some deterioration of photo-zs was obtained after adding non-magnitude parameters. Our experiments discussed here apply to a very different regime of magnitudes and redshifts than those by Soo et al. (2018), and unlike in that paper, we also incorporate colours in addition to purely morphological parameters such as sizes. These two independent analyses, and their conclusions, should then be regarded as largely complementary.

There are over 190 000 GAMA galaxies with good-quality redshifts (NQ $\geq$ 3) that have a KiDS counterpart in the equatorial fields. We split this sample randomly in proportion 1:2 into a training and test set; after cleanup of bad photometry and masking we are left with $\sim$56 000 galaxies for training and validation, and $\sim$112 000 for most of the tests. Due to occasionally missing IR fluxes, these numbers decrease by typically $\sim$1000 sources for some of the experiments where for instance VIKING or WISE measurements were additionally required; this has no influence on our conclusions.

5.1. Choice of the magnitude type

We start by looking at the properties of GAMA-depth photo-zs based on $ugri$ magnitudes only. As the benchmark we will use the default (pipeline) KiDS results from BPZ, as well as those from the ANNz2 training of the full KiDS-DR3 described in Sect. 4, but evaluated on the GAMA test set only. The detailed statistics for these two cases are provided in the top rows of Table 4. At this bright end, BPZ performs $\sim$10% better than the overall ANNz2 DR3 solution in terms of scatter, but over one order of magnitude worse in terms of bias. The third row of Table 4 shows that training the ANNs on GAMA sources only allows us to significantly reduce the scatter (by over 35%) while the bias remains minimal. These photo-z results for KiDS-GAMA sources are considerably better than those of WISE × SuperCOSMOS (Bilicki et al. 2016) where practically the same spectroscopic training data were used, also employing 4 bands at similar depth as here, although of worse photometric quality (in the optical based on digitised photographic plates). This highlights the importance of photometry for photo-zs even if the training sets are the same and the target data have similar depths.

As discussed in Sect. 3.1, the default KiDS GAaP magnitudes (Kuijken 2008), used for instance in the BPZ pipeline and

Table 4. Statistics of photometric redshift performance obtained for the KiDS-GAMA spectroscopic sample (extract).

Parameter setup	Mean of $\delta z = z_{ph} - z_{sp}$	Mean of $\delta z/(1 + z_{sp})$	St.dev. of $\delta z/(1 + z_{sp})$	SMAD[a] of $\delta z/(1 + z_{sp})$
ugri GAaP magnitudes, BPZ[b]	1.34×10^{-2}	1.17×10^{-2}	0.0481	0.0317
ugri GAaP magnitudes, DR3 ANNz2 overall solution[c]	-2.03×10^{-3}	8.61×10^{-4}	0.0508	0.0345
ugri GAaP magnitudes, GAMA-depth ANNz2 solution	-1.94×10^{-3}	-1.91×10^{-4}	0.0328	0.0219
ugri ISO magnitudes	-2.45×10^{-3}	-3.26×10^{-4}	0.0360	0.0259
ugri GAaP magnitudes +				
KiDS FLUX_APER_10_*ugri*	-1.98×10^{-3}	-2.71×10^{-4}	0.0321	0.0214
KiDS A, B linear semi-axes	-1.96×10^{-3}	-2.54×10^{-4}	0.0318	0.0213
KiDS six GAaP *ugri* colours	-2.13×10^{-3}	-3.99×10^{-4}	0.0321	0.0210
WISE $W1, W2, W3$ fluxes	-1.61×10^{-3}	-1.47×10^{-5}	0.0312	0.0206
VIKING $zYJHK_s$ fluxes	-1.58×10^{-3}	-2.22×10^{-5}	0.0312	0.0202
KiDS A, B + FLUX_APER_10_*ugri*	-1.71×10^{-3}	-1.16×10^{-4}	0.0311	0.0209
KiDS FLUX_APER_10_*ugri* + GAaP colours	-1.71×10^{-3}	-1.23×10^{-4}	0.0313	0.0205
KiDS A, B + GAaP colours	-1.59×10^{-3}	-7.61×10^{-5}	0.0305	0.0204
KiDS GAaP colours + VIKING fluxes	-1.72×10^{-3}	-1.53×10^{-4}	0.0311	0.0199
KiDS GAaP colours + WISE $W1, W2, W3$ fluxes	-1.50×10^{-3}	-9.24×10^{-6}	0.0303	0.0198
VIKING fluxes + WISE $W1, W2, W3$ fluxes	-1.18×10^{-3}	2.47×10^{-4}	0.0302	0.0197
KiDS A, B + FLUX_APER_10_*ugri* + GAaP colours	-1.69×10^{-3}	-1.67×10^{-4}	0.0304	0.0200
KiDS GAaP colours + VIKING fluxes + WISE $W1, W2, W3$ fluxes	-1.32×10^{-3}	7.55×10^{-5}	0.0296	0.0191
KiDS A, B + VIKING fluxes + WISE $W1, W2, W3$ fluxes	-1.29×10^{-3}	7.42×10^{-5}	0.0290	0.0188
KiDS A, B + GAaP colours + VIKING fluxes + WISE $W1, W2, W3$ fluxes	-1.33×10^{-3}	3.57×10^{-5}	0.0288	0.0186

Notes. See Appendix C for the full table. All the experiments used the same training and test sets of $\langle z \rangle = 0.23$; what was varied were the photometric parameters used for the photo-*z* derivation, as given in the first column. See text for details of the photometric parameters. [a] SMAD is the scaled median absolute deviation, converging to standard deviation for Gaussian distributions. [b] BPZ is independent of the training sets – the numbers are given for comparison (for the same GAMA test sample). These statistics are based on the KiDS pipeline solution. [c] ANNz2 overall KiDS-DR3 solution from Sect. 4.4 calculated for the same GAMA test set as in all the experiments of this table.

for the ANNz2 solution of Sect. 4.4, underestimate the fluxes of bright galaxies. We have therefore tested if this influences the photometric redshifts and whether other choices of magnitudes could be used to improve the results. This was done by training and testing on the same data but with *ugri* GAaP magnitudes replaced by first ISO and then AUTO measurements provided in the DR3 catalogue, zero-point calibrated and extinction corrected ("calib"). The results for both these options are shown in the second set of rows of Table C.1 in Appendix C (the shortened Table 4 only lists the ISO case). It is obvious that both cases gave significantly worse results in terms of scatter (respectively by almost 20% and over 25%) than the GAaP solution. We would like to reiterate however that these ISO and AUTO magnitudes have not undergone the PSF homogenisation discussed in Sect. 3.1, unlike the GAaP ones. As was shown by Hildebrandt et al. (2012), performing PSF homogenisation on the image level guarantees very accurate colours also for the ISO magnitudes, and therefore should improve the photo-*z*s based on such measurements. However, as such processing has not been done in KiDS except for the GAaP magnitudes, we will be thus using only the latter as the basic set of parameters, and other quantities will be supplemented to them to look for photo-*z* improvement. We did test the option of using the GAaP and ISO magnitudes together (that is, eight training parameters), but the improvement was minimal, and several other combinations worked much better, as described hereafter.

5.2. GAaP magnitudes with one additional set of parameters

Having determined that the GAaP magnitudes are optimal for photo-*z* estimation, we checked how adding other parameters,

correlated with redshift, can improve the results. We tested the following measurements from the KiDS multi-band catalogue (see Table A.2 in de Jong et al. 2017 for details on these quantities):

- A & B, which are linear semi major and minor axes of the galaxy, derived from *r*-band imaging; at the GAMA redshift range, these are expected to trace the monotonically increasing angular diameter distance;
- FLUX_APER_*size*_*band*, which is flux in *size* aperture, where $size \in \{4, 6, 10, 14, 25, 40, 100\}$ pixels. The tests were performed for each of the *ugri* bands in two configurations:
  - fixed *size*, all possible *bands*, that is, four parameters used together with the GAaP *ugri* magnitudes, for instance: FLUX_APER_4_u, FLUX_APER_4_g, FLUX_APER_4_r & FLUX_APER_4_i, and so on for each *size*;
  - fixed *band*, all possible *sizes*, that is, seven parameters used together with the GAaP *ugri* magnitudes, for instance: FLUX_APER_4_u, FLUX_APER_6_u, ..., FLUX_APER_40_u & FLUX_APER_100_u, and so on for each *band*; this configuration can be regarded as a proxy for surface brightness profile, expected to provide very good constraints on photo-*z*s (Kurtz et al. 2007);
- ISOAREA_IMAGE_*band*, that is, isophotal aperture in pixel^2, for each *band* (four parameters used together with the *ugri* GAaP magnitudes);
- colour_GAAPHOM_*band1*_*band2*, that is, homogenised and extinction corrected GAaP colours, for all combinations of *band1* and *band2* (six parameters used together with the *ugri* GAaP magnitudes).

Except for the GAaP-based magnitudes and colours, the remaining parameters used in these tests were derived by SExtractor. This software may not be optimal for source extraction in crowded fields and new, presumably better, tools are being developed for such a purpose, for example The Tractor (Lang et al. 2016a) or ProFound (Robotham et al. 2018). Testing their performance would be however beyond the scope of this work as source extraction is embedded deeply in KiDS processing pipeline. In other words, we cannot verify at present which of the SExtractor-based measurements are unreliable. Moreover, these parameters were not corrected for PSF variations so the photo-z improvement they provide should be considered as lower limits.

We note however that the "GAMA-like" galaxies we are concerned with in this Section are typically much larger than the PSF, so the variations of the latter are not expected to significantly bias the measurements that we use here. This is especially true in the most stable r band, in which galaxy sizes, used in the photo-z derivation for the released catalogue (Sect. 5.4), are measured. The PSF in this band is well-behaved both within individual coadds (Kuijken et al. 2015), as well as between the tiles (de Jong et al. 2017). A possibly more important PSF-related effect could be for those features which use multi-band information, such as FLUX_APER_size_band. Here indeed differences in the PSF between the bands could affect the photo-z measurements. Being unable to quantify this effect at present, we note that overall, even if small, improvement in photo-zs after adding these various parameters suggests that the related noise does not dominate.

The third set of rows of Table 4 presents selected results from the combinations of the $ugri$ magnitudes with one set of those of the above additional parameters which performed best in terms of scatter; a more extended set of results is given in Table C.1 in Appendix C. Some of the parameters do not appear in the Tables because they brought less improvement than those listed, but all of them were tested. The best results were obtained by adding GAaP colours, and a bit worse by using the sizes (semi-axes), as well as for several FLUX_APER combinations. The improvement is not huge (no more than by 5% in scatter with respect to the basic $ugri$ setup) but clearly visible.

The fact that using colours together with magnitudes improves the magnitude-only results may seem puzzling at first, because the two sets of parameters are redundant. We verified that the same effect exists not only for ANNs but also for BDTs, so it seems to be a general property of this type of MLMs. Moreover, similar improvement was observed by Hoyle et al. (2015) where SDSS magnitudes were also combined with colours, and ANNs used for photo-z experiments. We interpret this improvement as due to the fact that the parameter space of magnitudes only is not constraining enough for the MLMs to converge on a solution sufficiently close to the truth. The MLMs do not "know" a priori that colours are simple combinations of magnitudes. Using both together adds physical information on how galaxy observed properties are related to redshift and forces the MLMs to work in a much better constrained region of the parameter space, which improves the mapping between photometry and redshift.

We also tested how adding photometry external to KiDS (namely from VIKING and WISE) can improve the photo-zs, and the results are very promising in view of the forthcoming KiDS-VIKING data, as well as of the forced photometry of WISE on KiDS sources that we are planning to obtain. As is clear from Table 4, already adding WISE $W1$ and $W2$ to KiDS $ugri$ gives results better than the best KiDS-only solution discussed in this Subsection. Extending the parameter space by adding $W3$ brings in further improvement, with a caveat that

part of the objects (~16%) have no flux measurements in this band (that is, W3_flux = 0), which means that for them the $W3$ information was effectively ignored. As was already discussed in Sect. 2.1, the default setup of our photo-z experiments does not use the parameter errors provided in the catalogues but relies on an internal error model built by ANNz2. In case of WISE we tested however the option of directly using the provided errors, which are here considerable especially in the $W3$ channel. We found negligible (sub-percent) differences in the photo-z statistics between the two setups, as is shown in Table C.1 in Appendix C.

The KiDS+WISE statistics can be compared for instance with the 2dFLenS study by Wolf et al. (2017) where it was found that adding the $W1$, $W2$ bands to the $ugriz$ optical setup (based on VST-ATLAS photometry) improved the results by 5–10%, which was however a smaller difference than between some of the photo-z methods tested there. Last but not least, using the five additional bands from VIKING (without employing WISE) gives the best results of all considered so far, although the improvement is not dramatic (~−8% in scatter over the fiducial $ugri$ setup). This is somewhat similar to some of the results from a DES + VHS analysis by Banerji et al. (2015), although those experiments used a very different setup from that presented here, so the results are not directly comparable.

It is also worth noting that the normalised bias of photo-zs is improved by adding IR, by an order of magnitude over most of the so-far discussed experiments using only optical data. Furthermore, these results should be treated as lower limits to the improvement in photo-zs possible by adding the IR data, for several reasons. Firstly, the LAMBDAR forced photometry was based on SDSS DR7 apertures as input, which are more noisy than KiDS measurements. Secondly, we may expect – similarly as found in Sect. 5.1 – that using GAaP magnitudes for the IR bands (being currently derived by the KiDS team for VIKING) will improve the derived photo-zs. Last but not least, of some importance should be proper zero-point calibration and extinction corrections, not applied to these bands in the present tests.

We would like to emphasise however that the NIR data are expected to help with photo-z estimation mostly when the Balmer break is redshifted into the appropriate filters, leaving the KiDS $ugri$ coverage, which happens for $z > 1.35$. Therefore, data much deeper than used in this Section are needed for the NIR to bring the most benefit for photo-zs.

5.3. GAaP magnitudes with multiple parameter set combinations

Having determined the single set of additional parameters bringing the most improvement to photo-zs when used with the basic $ugri$ setup, we proceeded to joining them. Here we explored only the combinations of these sets from the previous tests that gave the best results. From the third block of rows of Tables 4 and C.1 it is clear that the best option of two additional parameter sets should be by adding VIKING and WISE; however, due to the current unavailability of proper photometry for these surveys outside of the GAMA equatorial patches, it is of interest to study also KiDS-only parameter combinations. In particular, we examined the unions of the A, B sizes, GAaP colours, FLUX_APER_10_band and FLUX_APER_size_r measurements, first pairwise (that is, by adding two sets of parameters to GAaP magnitudes), and then in multiple combinations based on the results of the GAaP+pairwise experiments.

In the fourth set of rows of Tables 4 and C.1 we provide results for the best cases of the GAaP+pairwise options, first

for KiDS-only and then for KiDS+VIKING+WISE ones. In the latter case we were able to break the barrier of 0.02 in scatter, which is more than 10% improvement over using only *ugri* magnitudes. But even without IR photometry, combining GAaP magnitudes, colours, and linear sizes gives very well constrained photo-*z*s with bias $\sim 10^{-3}$ and scatter ~ 0.02. This will be taken advantage of in the publicly released photo-*z* catalogue described in Sect. 5.4. Interestingly, whenever *ugri* magnitudes are supplemented with optical colours and/or morphological parameters, adding WISE $W1$–$W3$ seems more beneficial than adding VIKING z–K_s bands.

Using more extended parameter setups further improves the results, although it is at the expense of the computation time, which for ANNs scales non-linearly with the number of training parameters. As far as KiDS-only quantities are concerned, we stopped our experiments on four-set combinations, the best results of which came from using GAaP magnitudes, colours, linear sizes, and 10-pixel aperture magnitudes in the four bands. The improvement over using *ugri* magnitudes+colours+sizes is not large. On the other hand, combining the optical magnitudes, colours, and sizes with VIKING and WISE measurements allowed us to obtain SMAD ($\delta z/(1 + z)$) < 0.019 with normalised bias $\sim 10^{-5}$. In the case of further adding the 10-pixel aperture *ugri* magnitudes, some further improvement is seen (Table C.1), but we have reasons to believe that this setup of even 24-dimensional parameter space might be too large for the ANNs to work efficiently (see for example Soumagnac et al. 2015 for a related discussion).

5.4. KiDS-GAMA photo-*z* catalogue release

Based on the results of the above analysis, we computed very accurate and precise photo-*z*s for a sample of 800 000 bright ($r \lesssim 20$) KiDS galaxies over the whole DR3 area of 450 deg^2. This time the training set included both the GAMA equatorial data, as well as those from the G23 field. Although the latter are less complete than the former, the G23-GAMA galaxies were also selected for spectroscopic measurements based on their apparent magnitudes and not colours. This guarantees that we do not introduce biases into the training that could be related to using for instance LRGs or other sources for which photo-*z*s are usually better constrained.

The target GAMA-like KiDS catalogue will have somewhat different selections than the spectroscopic GAMA sample. The main reason are differences in photometry: GAMA input sources were selected from SDSS DR7 using the *r*-band Petrosian magnitude measurements as reference (except for G02 and G23 where additional photometry from CFHTLenS, KiDS and VIKING was used; Liske et al. 2015). In KiDS, we do not have Petrosian magnitudes, and the closest to them among the 3 possibilities discussed in Sect. 3.1 are the AUTO ones. Using the GAMA equatorial fields with an $r_{\mathrm{Petro}} \leq 19.8$ cut as the reference, we obtained $\langle r_{\mathrm{AUTO}}^{\mathrm{KiDS}} - r_{\mathrm{Petro}}^{\mathrm{SDSS}} \rangle = 0.018 \pm 0.201$ mag (mean with standard deviation). These differences are related not only to different ways of measuring the magnitudes, but also partly to the lower photometric quality of GAMA-input SDSS photometry than that of the KiDS measurements.

Nevertheless, in order to maximise the completeness of KiDS "GAMA-like" galaxies with respect to the actual GAMA sources, we need to take a fainter cut in KiDS than $r < 19.8$. For instance, $r_{\mathrm{AUTO}} \leq 20.3$ retains 99.5% of GAMA galaxies, while using $r_{\mathrm{AUTO}} \leq 19.8$ would decrease the completeness to only 95.1%. On the other hand, going on average 0.5 mag fainter than the fiducial GAMA limits will introduce many sources into the

KiDS GAMA-like sample that are not well covered by the spectroscopic training sample (although it is partly alleviated due to the fact that GAMA does include some fainter objects with redshift measurements). We emphasise however that all these considerations do not influence the ANNz2 photo-*z* training process itself, which uses only confirmed KiDS × GAMA galaxies. One should only bear in mind that the evaluated photo-*z*s of sources fainter than the GAMA limits might be not reliable, and cuts on the presented here GAMA-like KiDS catalogue might be necessary to mitigate this.

To summarise, the released data product[13] includes KiDS DR3 sources cut at $r_{\mathrm{AUTO}} \leq 20.3$, extinction-corrected and zero-point calibrated, that is, using "MAG_AUTO_r_calib" for selection. Together with this, star removal was also applied by using KiDS star-galaxy separator SG2DPHOT = 0. At the bright end of KiDS, this parameter is reliable enough to guarantee practically 100% purity of the galaxy sample, while it only minimally influences the completeness with respect to GAMA ($\sim 99.2\%$ once combined with the $r_{\mathrm{AUTO}} \leq 20.3$ magnitude cut). Such a selection from the full KiDS-DR3, together with the requirements of |MAG_GAAP_band|<99 and MAGERR_GAAP_band > 0 for each band, gives 801 000 sources over the KiDS DR3 footprint. Applying additionally the masking flag of IMAFLAGS_ISO_band & 01010111 = 0 for each band leaves 695 000 galaxies in the GAMA-like KiDS-clean sample. We do not apply this latter flagging to the published data, leaving this to the end-users.
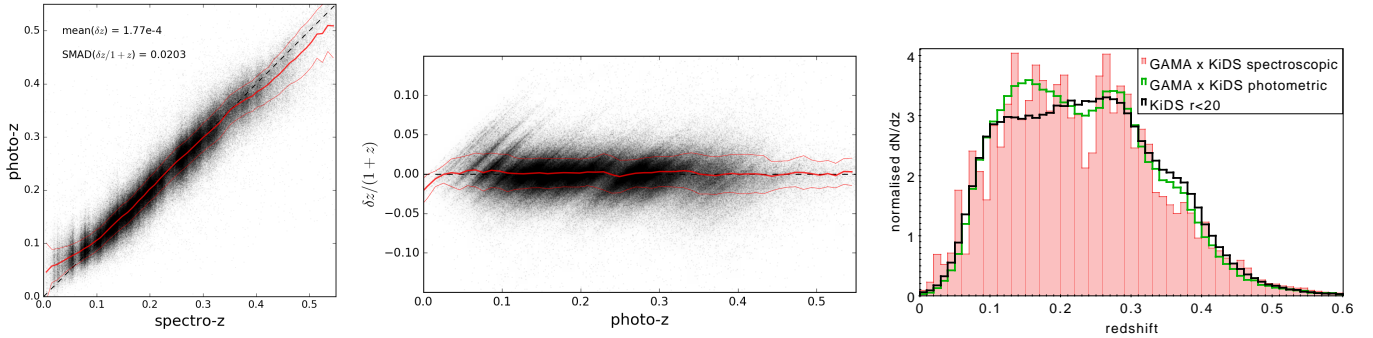
As mentioned, the GAMA-depth photo-*z*s released with this paper are based on the training set composed of KiDS×GAMA galaxies from the equatorial (G09, G12, G15) and southern (G23) fields. This sample includes almost 227 500 galaxies with $\langle z \rangle = 0.23$ and $\langle r_{\mathrm{GAaP}} \rangle = 19.4$. The parameters supplied to the ANNs were GAaP *ugri* magnitudes, related colours, and the A and B linear sizes; this is the setup with the best performance among the KiDS-only combinations of the magnitudes with two additional sets of parameters (fourth set of rows in Table 4). As in all other experiments, this training sample was split randomly in two halves, one for the actual training, and the other for validation (optimisation). A total of 250 ANNs were trained, of architectures generated randomly each time.

Figure 7 illustrates the performance of the GAMA-depth photo-*z*s in the released catalogue, as judged from a comparison with the spectroscopic GAMA data in the equatorial fields. We see that except for the very local volume of $z \sim 0$, the photo-*z*s are extremely stable and well-constrained up to the limits of GAMA at $z \sim 0.6$, and their overall performance for this sample is $\langle \delta z \rangle = 1.77 \times 10^{-4}$ and SMAD($\delta z/(1 + z)$) = 0.0203. For comparison, the KiDS pipeline photo-*z*s from BPZ give for the same data $\langle \delta z \rangle = 0.0153$ and SMAD($\delta z/(1 + z)$) = 0.0317. The redshift distributions in the right-hand panel of Fig. 7 show that the ML photo-*z*s are so good to trace even a "dip" in dN/dz of the GAMA equatorial catalogue which is caused by the large-scale structure crossing the GAMA fields (Eardley et al. 2015). This dip is of course not observed in the full KiDS-GAMA photo-*z* dataset, as shown with the black line illustrating all the KiDS $r < 20$ sources for which photo-*z*s were derived as described in the present Section.

This "GAMA-like" KiDS catalogue has been already used in scientific analyses, and its first published application is presented in Brouwer et al. (2018), where it is employed as the foreground for a weak lensing analysis of galaxy troughs and ridges.

---

[13] Data available from http://kids.strw.leidenuniv.nl/DR3/ml-photoz.php#annz2

Fig. 7. Performance of the KiDS-GAMA ANNz2 photo-zs as compared to the GAMA spectroscopic redshifts in the equatorial fields. *Left-hand panel*: direct spec-z–photo-z comparison; *central panel*: photo-z error as a function of photo-z; *right-hand panel*: comparison of redshift distributions for the same set of KiDS × GAMA sources (red bars for spec-zs, green line for photo-zs), with also dN/dz_{phot} of the full bright-end KiDS sample ($r < 20$) overplotted (black line), all normalised to unit area under the histograms.

6. Conclusions and future prospects

In this paper we presented an analysis of machine learning photometric redshifts in the Kilo-Degree Survey Data Release 3, and quantified the properties of two accompanying photo-z catalogue releases, one at the full depth of the survey and the other limited to its bright end. In the latter case, we additionally studied possible extensions of the fiducial *ugri* parameter space, both by adding extra imaging information (galaxy colours, sizes, fluxes in fixed apertures), as well as by using infrared photometry from VIKING and WISE.

At the full depth available from overlapping spectroscopy we made a comparison of two MLMs used in KiDS – ANNz2 and MLPQNA – between each other as well as against the KiDS pipeline photo-z solution from BPZ. This was done for various samples extracted from the current overlap between KiDS DR3 (plus some auxiliary photometric data) and external spectroscopic catalogues. We showed that at the bright, low-redshift end ($z < 0.5$) of KiDS, the two ML photo-z methods perform better than BPZ in most statistics, which is expected, as this is where the spectroscopic calibration data (mostly from GAMA) is the most abundant. But also for dimmer and higher-redshift sources (up to $z \sim 1$) the MLMs provide well-constrained photo-zs of comparable quality to the BPZ solution, despite much worse training data coverage there.

These general conclusions apply also to the publicly-released KiDS DR3 photo-z catalogue derived using ANNz2. This dataset includes all the KiDS DR3 sources having 4-band *ugri* measurements (over 39 million objects), although for part of these the photo-zs are based on extrapolation over the limits of the training set and must be used with caution. For scientific applications we therefore defined a FIDUCIAL subsample of 20.5 million extended sources, which is limited to the photometric coverage of the training sets used by ANNz2, and provides more secure photo-zs (additionally improved thanks to weighting the training set). We judge that the photo-zs in this catalogue are trustworthy to at least $z_{phot} \lesssim 0.9$ and $r \lesssim 23.5$.

In the second part of the paper we focused on the bright end of the KiDS catalogue ($r < 20$, $\langle z \rangle = 0.23$) and made a comprehensive analysis of ML photo-zs in this regime, taking advantage of the excellent KiDS photometry and of the high spectroscopic completeness of the largely overlapping GAMA survey. Having obtained very accurate, $\langle \delta z/(1 + z) \rangle \sim 10^{-4}$, and precise, $\sigma_{\delta z} < 0.022(1 + z)$, photo-$z$s for these KiDS sources when training the ANNz2 algorithm on 4-band *ugri* magnitudes, we further studied how extending this basic parameter space can improve

the redshifts. We looked at both adding KiDS-only quantities, such as colours, galaxy sizes, and fixed-aperture magnitudes, but also – in view of the forthcoming or planned forced-photometry reduction of VIKING and WISE data in KiDS fields – at additionally using near- and mid-infrared photometry.

The general conclusion from the bright-end study is that extending the parameter space used for photo-z derivation with galaxy colours as well as morphological quantities does improve the results, both in terms of bias as well as scatter. These improvements, although noticeable, are not huge, and in the best case of using *ugri* magnitudes + colours + linear sizes + 10-pixel-aperture magnitudes, the scatter in $\delta z/(1 + z)$ is reduced by ~9% over the fiducial case of magnitudes only. Adding IR measurements, on the other hand, is more promising, especially if they are also combined with the optical colours and morphological parameters. Photo-zs derived from 12-band photometry (from u up to WISE 12 μm) have scatter smaller by $> 10\%$ than in the optical-only case. Combined further with colours and sizes, photo-zs of $\sigma_{\delta z} < 0.019(1 + z)$ will be possible.

It is worth emphasising that these improvements should be considered as lower limits for KiDS-based estimates, as the IR photometry we used was based on SDSS apertures, and we lacked the GAaP magnitudes, shown in this paper to provide much more robust photo-zs than other magnitude types available in KiDS.

Future prospects for KiDS photo-zs, in particular the ML ones, look bright. Both the photometry coverage as well as the training sets are being significantly extended. The availability of nine KiDS+VIKING bands at practically the full depth of KiDS will allow us to improve the accuracy and precision of photo-zs both at the bright end of the sample (as shown in this paper), but also – and perhaps more importantly – at the faint end where the NIR bands start to play a significant role in constraining photo-zs (Banerji et al. 2015). For this amelioration to be achieved, more high-redshift spectroscopic training data is however needed to mitigate the sample variance. And indeed, observations of relevant fields both in the optical with VST and in the NIR with VISTA have been already made (VVDS fields) or are ongoing, or solicited for (VIPERS areas). We also plan to add the overlapping WiggleZ (Drinkwater et al. 2010) to the spectroscopic sample, not used until now in KiDS analyses.

It is worth emphasising that the forthcoming nine-band KiDS+VIKING photo-zs will place these surveys in a unique position as far as datasets of such angular extent and depth are concerned. Until now, photo-zs with (at least) that many bands have been available either for wide-angle but very shallow

samples (for example SDSS-based, Way et al. 2009; 2MPZ, Bilicki et al. 2014) or for deep but small-area ones (for example EGS, Barro et al. 2011; COSMOS, Laigle et al. 2016). The joint KiDS + VIKING photo-*z* analysis and related data products will fill this important gap.

References

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4
Albareti, F. D., Allende Prieto, C., Almeida, A., et al. 2017, ApJS, 233, 25
Amaro, V., Cavuoti, S., Brescia, M., et al. 2018, MNRAS, submitted
Banerji, M., Jouvel, S., Lin, H., et al. 2015, MNRAS, 446, 2523
Barro, G., Pérez-González, P. G., Gallego, J., et al. 2011, ApJS, 193, 30
Baum, W. A. 1957, AJ, 62, 6
Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371
Benítez, N. 2000, ApJ, 536, 571
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., & Steward, L. 2014, ApJS, 210, 9
Bilicki, M., Peacock, J. A., Jarrett, T. H., et al. 2016, ApJS, 225, 5
Blake, C., Amon, A., Childress, M., et al. 2016, MNRAS, 462, 4240
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio A. 2013, ApJ, 772, 140
Brescia, M., Cavuoti, S., Longo, G., et al. 2014, PASP, 126, 783
Brodwin, M., Brown, M. J. I., Ashby, M. L. N., et al. 2006, ApJ, 651, 791
Brouwer, M. M., Cacciato, M., Dvornik, A., et al. 2016, MNRAS, 462, 4451
Brouwer, M. M., Demchenko, V., Harnois-Déraps, J., et al. 2018, MNRAS, submitted [arXiv:1805.00562]
Brun, R., & Rademakers, F. 1997, Nucl. Instrum. Methods Phys. Res. A, 389, 81
Byrd, R. H., Nocedal, J., & Schnabel, R. B. 1994, Math Program, 63, 129
Capaccioli, M., Schipani, P., de Paris, G., et al. 2012, in Science from the Next Generation Imaging and Spectroscopic Surveys, 1
Cavuoti, S., Brescia, M., Longo, G., & Mercurio, A. 2012, A&A, 546, A13
Cavuoti, S., Brescia, M., De Stefano, V., & Longo, G. 2015a, Exp. Astron., 39, 45
Cavuoti, S., Brescia, M., Tortora, C., et al. 2015b, MNRAS, 452, 3100
Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959
Cluver, M. E., Jarrett, T. H., Hopkins, A. M., et al. 2014, ApJ, 782, 90
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Cooper, M. C., Yan, R., Dickinson, M., et al. 2012, MNRAS, 425, 2116
Costa-Duarte, M. V., Viola, M., Molino, A., et al. 2018, MNRAS, 478, 1968
D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111
Davies, L. J. M., Driver, S. P., Robotham, A. S. G., et al. 2015, MNRAS, 447, 1014
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44
de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., & Valentijn, E. A. 2013, Exp. Astron., 35, 25
de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., et al. 2015, A&A, 582, A62
de Jong, J. T. A., Kleijn, G. A. V., Erben, T., et al. 2017, A&A, 604, A134
Drinkwater, M. J., Jurek, R. J., Blake, C., et al. 2010, MNRAS, 401, 1429
Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971
Dvornik, A., Cacciato, M., Kuijken, K., et al. 2017, MNRAS, 468, 3251
Eardley, E., Peacock, J. A., McNaught-Roberts, T., et al. 2015, MNRAS, 448, 3665
Edge, A., Sutherland, W., Kuijken, K., et al. 2013, The Messenger, 154, 32
Ellis, R. S. 1997, ARA&A, 35, 389
Firth, A. E., Lahav, O., & Somerville, R. S. 2003, MNRAS, 339, 1195
Fukugita M., Shimasaku, K., & Ichikawa, T. 1995, PASP, 107, 945
Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al. 2010, ApJ, 715, 823
Gomes, Z., Jarvis, M. J., Almosallam, I. A., & Roberts, S. J. 2018, MNRAS, 475, 331
Gruen, D., Friedrich, O., Amara, A., et al. 2016, MNRAS, 455, 3367
Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, A&A, 566, A108
Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31
Hildebrandt, H., Erben, T., Kuijken, K., et al. 2012, MNRAS, 421, 2355
Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454
Hoecker, A., Speckmayer, P., Stelzer, J., et al. 2007, ArXiv e-prints [arXiv:physics/0703039]
Hogan, R., Fairbairn, M., & Seeburn, N. 2015, MNRAS, 449, 2040
Hoyle, B. 2016, Astron. Comput., 16, 34
Hoyle, B., Rau, M. M., Zitlau, R., Seitz, S., & Weller, J. 2015, MNRAS, 449, 1275
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Jarrett, T. H., Cluver, M. E., Magoulas, C., et al. 2017, ApJ, 836, 182
Johnson, A., Blake, C., Amon, A., et al. 2017, MNRAS, 465, 4118
Jones, E., & Singal, J. 2017, A&A, 600, A113
Jones, E., Oliphant, T., Peterson, P., et al. 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/
Joudaki, S., Mead, A., Blake, C., et al. 2017, MNRAS, 471, 1259
Joudaki, S., Blake, C., Johnson, A., et al. 2018, MNRAS, 474, 4894
Kodama, T., Bell, E. F., & Bower, R. G. 1999, MNRAS, 302, 152
Köhlinger, F., Viola, M., Joachimi, B., et al. 2017, MNRAS, 471, 4412

Koo, D. C. 1985, AJ, 90, 418

Kuijken, K. 2008, A&A, 482, 1053

Kuijken, K. 2011, The Messenger, 146, 8

Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, MNRAS, 454, 3500

Kurtz M. J., Geller, M. J., Fabricant, D. G., Wyatt, W. F., & Dell'Antonio, I. P. 2007, AJ, 134, 1360

Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24

Lang, D., Hogg, D. W., & Mykytyn, D. 2016a, Astrophysics Source Code Library [record ascl:1604.008]

Lang, D., Hogg, D. W., & Schlegel, D. J. 2016b, AJ, 151, 36

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]

Laureijs, R., Amiaux, J., Arduini, S., et al. 2003, Proc. SPIE, 4834, 173

Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, ApJS, 184, 218

Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, MNRAS, 390, 118

Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, MNRAS, 452, 2087

Loh, E. D., & Spillar, E. J. 1986, ApJ, 303, 154

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, ArXiv e-print [arXiv:0912.0201]

Ma, Z., Hu, W., & Huterer, D. 2006, ApJ, 636, 21

Mahlke, M., Bouy, H., Altieri, B., et al. 2018, A&A, 610, A21

Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53

Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2017, ApJ, 841, 111

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5

Newman, J. A., Abate, A., Abdalla, F. B., et al. 2015, Astropart. Phys., 63, 81

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

Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129

Pila Díez B. 2015, Ph.D. Thesis, Leiden University

Radovich, M., Puddu, E., Bellagamba, F., et al. 2017, A&A, 598, A107

Robotham, A. S. G., Davies, L. J. M., Driver, S. P., et al. 2018, MNRAS, 476, 3137

Rosenblatt, F. 1962, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms (Spartan Books)

Rozo, E., Rykoff, E. S., Abate, A., et al. 2016, MNRAS, 461, 1431

Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502

Sánchez, C., Carrasco Kind, M., Lin, H., et al. 2014, MNRAS, 445, 1482

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

Sifón, C., Cacciato, M., Hoekstra, H., et al. 2015, MNRAS, 454, 3938

Singal, J., Shmakova, M., Gerke, B., Griffith, R. L., & Lotz, J. 2011, PASP, 123, 615

Soo, J. Y. H., Moraes, B., Joachimi, B., et al. 2018, MNRAS, 475, 3613

Soumagnac, M. T., Abdalla, F. B., Lahav, O., et al. 2015, MNRAS, 450, 666

Tagliaferri, R., Longo, G., Andreon, S., et al. 2003, in Lecture Notes in Comp. Sci., 2859, 226

Taylor, M. B. 2005, ASP Conf. Ser., 347, 29

Taylor, M. B. 2006, ASP Conf. Ser., 351, 666

The Dark Energy Survey Collaboration 2005, Dark Energy Task Force, submitted [arXiv:astro-ph/0510346]

Tortora, C., La Barbera, F., Napolitano, N. R., et al. 2016, MNRAS, 457, 2845

Vaccari, M., Covone, G., Radovich, M., et al. 2016, in Proc. of the 4th Annual Conference on HEASA [arXiv:1704.01495]

van der, Walt, S., Colbert, S. C., & Varoquaux G. 2011, Comput. Sci. Eng., 13, 22

van Uitert, E., Cacciato, M., Hoekstra, H., et al. 2016, MNRAS, 459, 3251

van Uitert, E., Hoekstra, H., Joachimi, B., et al. 2017, MNRAS, 467, 4131

van Uitert, E., Joachimi, B., Joudaki, S., et al. 2018, MNRAS, 476, 4662

Velander, M., van Uitert, E., Hoekstra, H., et al. 2014, MNRAS, 437, 2111

Venemans, B. P., Verdoes Kleijn, G. A., Mwebaze, J., et al. 2015, MNRAS, 453, 2259

Viola, M., Cacciato, M., Brouwer, M., et al. 2015, MNRAS, 452, 3529

Wadadekar, Y. 2005, PASP, 117, 79

Way, M. J. 2011, ApJ, 734, L9

Way, M. J., Foster, L. V., Gazis, P. R., & Srivastava, A. N. 2009, ApJ, 706, 623

Wolf, C., Johnson, A. S., Bilicki, M., et al. 2017, MNRAS, 466, 1582

Wray, J. J., & Gunn, J. E. 2008, ApJ, 678, 144

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

Wright, A. H., Robotham, A. S. G., Bourne, N., et al. 2016, MNRAS, 460, 765

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Appendix A: Data release details

Here we provide details of the two photometric redshift catalogues released with this paper, available via http://kids.strw.leidenuniv.nl/DR3/ml-photoz.php#annz2. We include only basic parameters in them, in order to enable end users to apply selections as described in the paper. For more sophisticated filtering, these datasets can be cross-matched with the overall DR3 data using the unique source identifier ID.

A.1. Full-depth photometric redshift catalogue

Table A.1 lists the columns included in the publicly-released full-depth KiDS DR3 photometric redshift catalogue, in which photo-zs were derived using ANNz2 trained on the full spectro-photo compilation (Sect. 3.5), as detailed in Sect. 4.4. The dataset includes 39.2 million sources extracted from DR3 by requiring that all the four *ugri* GAaP magnitudes are measured. The catalogue is meant for general-purpose uses, but additional filtering as described in Sects. 3.1 and 4.4 is needed to remove artefacts and to guarantee reliable photo-zs; we provide the "fiducial" flag to be applied as the minimum requirement. For details of the listed columns, please see Appendix A.2 of de Jong et al. (2017).

A.2. Bright-end photometric redshift catalogue

Table A.2 lists the columns included in the publicly-released bright-end KiDS DR3 photometric redshift catalogue, in which photo-zs were derived using ANNz2 trained on the GAMA spectroscopic sources (Sect. 3.4.1), as detailed in Sect. 5.4. The dataset includes 800 830 sources extracted from DR3 by applying the magnitude cut of `MAG_AUTO_R_calib`≤ 20.3, star-galaxy separation parameter `SG2DPHOT = 0`, and by requiring that all the four *ugri* GAaP magnitudes, and their errors, are measured. Additional filtering as described in Sect. 5.4 is needed to remove artefacts. For details of the listed columns, please see Appendix A.2 of de Jong et al. (2017).

Appendix B: Photometric redshift distributions for red and blue sources

In Sect. 4.4 we compared photo-z distributions for the BPZ and ANNz2 solutions of the FIDUCIAL DR3 photometric sample and found visible discrepancies at various redshift ranges. To probe further the source of these disagreements, we examined the differences between individual photo-zs from the two available full-depth solutions, namely $z_{ANNz2} - z_{BPZ}$, as a function of observed colours. By examining various colour-space projections we found that the two solutions are more consistent for redder sources. In particular by projecting on the $r - i$ vs. $g - r$ colour-colour plane, we determined that a good distinction between more consistent and less consistent photo-zs (in terms of $|z_{ANNz2} - z_{BPZ}|$) is given by $g - r = 0.8 - 0.8(r - i)$ and $r - i = 0.5$ lines. Sources lying above these two lines (that is, redder) have overall much smaller photo-z differences than the bluer ones. We note that this is just a general division irrespectively of morphology, and some of the "red" galaxies can in fact be spirals, or even irregulars of Magellanic type (Fukugita et al. 1995). In Fig. B.1 we compare the resulting dN/dz_{phot} distributions for thus selected "red" and "blue" galaxies, for the two photo-z solutions. This comparison is limited to the FIDUCIAL photometric sample, additionally cut at $r < 23.5$ to avoid ANNz2 extrapolation discussed in Sect. 4.4. This confirms that redder galaxies have much more consistent photo-z distributions

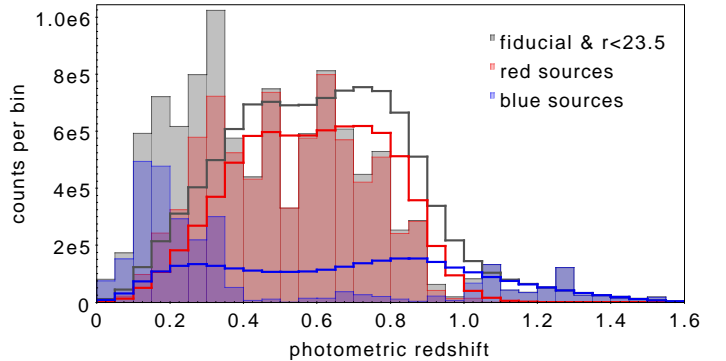**Table A.1.** Columns provided in the full-depth ANNz2 photometric redshift catalogue.

Label	Format	Unit	Description
ID	23A		Source identifier
RAJ2000	D	deg	Right ascension (J2000)
DECJ2000	D	deg	Declination (J2000)
SG2DPHOT	K		Source classification
IMAFLAGS_ISO_band	J		Mask flag
MAG_GAAP_band_calib	E	mag	Calibrated GAaP magnitude[a]
MAGERR_GAAP_band	E	mag	Error in GAaP magnitude
zphot_ANNz2	D		Photometric redshift derived with ANNz2
fiducial	I		Flag defining fiducial selection

Notes. [a]Extinction-corrected and zero-point calibrated GAaP magnitudes: `MAG_GAAP_band_calib = MAG_GAAP_band + ZPT_OFFSET_band − EXT_SFD_band`.

Table A.2. Columns provided in the bright-end ANNz2 photometric redshift catalogue.

Label	Format	Unit	Description
ID	23A		Source identifier
RAJ2000	D	deg	Right ascension (J2000)
DECJ2000	D	deg	Declination (J2000)
IMAFLAGS_ISO_band	J		Mask flag
MAG_AUTO_band_calib	E	mag	Calibrated Kron-like elliptical aperture magnitude[a]
MAGERR_AUTO_band	E	mag	RMS error for MAG_AUTO
MAG_ISO_band_calib	E	mag	Calibrated isophotal magnitude[a]
MAGERR_ISO_band	E	mag	RMS error for MAG_ISO
MAG_GAAP_band_calib	E	mag	GAaP "calibrated" magnitude[a]
MAGERR_GAAP_band	E	mag	Error in GAaP magnitude
zphot_ANNz2	D		Photometric redshift derived with ANNz2

Notes. [a]Extinction-corrected and zero-point calibrated magnitudes: `MAG_type_band_calib = MAG_type_band + ZPT_OFFSET_band − EXT_SFD_band`.

Fig. B.1. Comparison of photometric redshift distributions for KiDS DR3 galaxies preselected as red and blue, the division line being a join of the $g - r = 0.8 - 0.8(r - i)$ and $r - i = 0.5$ conditions. The sample is limited to the FIDUCIAL selection (Sect. 3.1) with an additional $r < 23.5$ cut. Filled histograms are for the BPZ solution and thick lines are for ANNz2. Grey is for the full sample, while red and blue are for respective colour selections.

than the full dataset, while the blue ones are assigned very different dN/dzs by the two methods. In particular, for BPZ a large fraction of blue sources have $z_{phot} < 0.4$ and these constitute most of the low-z_{BPZ} peak observed in the full sample without the colour division. On the other hand, ANNz2 produces a very flat dN/dz_{phot} distribution for blue galaxies and there are almost no red sources allocated to $z_{ANNz2} > 1$. The latter is equally true for $z_{BPZ} > 1$.

Appendix C: Extended table for GAMA-depth experiments

Table C.1 is an extended version of Table 4.

Table C.1. Statistics of photometric redshift performance obtained for the KiDS-GAMA spectroscopic sample.

Parameter setup	Mean of $\delta z = z_{\rm ph} - z_{\rm sp}$	Mean of $\delta z/(1 + z_{\rm sp})$	St.dev. of $\delta z/(1 + z_{\rm sp})$	SMAD[a] of $\delta z/(1 + z_{\rm sp})$
ugri GAaP mags, BPZ[b]	1.34×10^{-2}	1.17×10^{-2}	0.0481	0.0317
ugri GAaP mags, DR3 ANNz2 overall solution[c]	-2.03×10^{-3}	8.61×10^{-4}	0.0508	0.0345
ugri GAaP mags, GAMA-depth ANNz2 solution	-1.94×10^{-3}	-1.91×10^{-4}	0.0328	0.0219
ugri ISO magnitudes	-2.45×10^{-3}	-3.26×10^{-4}	0.0360	0.0259
ugri AUTO magnitudes	-2.72×10^{-3}	-3.83×10^{-4}	0.0384	0.0278
ugri GAaP magnitudes +				
KiDS `FLUX_APER_10`_*ugri*	-1.98×10^{-3}	-2.71×10^{-4}	0.0321	0.0214
KiDS `FLUX_APER_14`_*ugri*	-1.97×10^{-3}	-2.49×10^{-4}	0.0323	0.0214
KiDS `FLUX_APER_25`_*ugri*	-2.00×10^{-3}	-2.84×10^{-4}	0.0327	0.0214
KiDS `FLUX_APER_{4, ..., 100}`_*r*	-1.85×10^{-3}	-1.91×10^{-4}	0.0315	0.0214
KiDS `A`, `B` linear semi-axes	-1.96×10^{-3}	-2.54×10^{-4}	0.0318	0.0213
KiDS six GAaP *ugri* colours	-2.13×10^{-3}	-3.99×10^{-4}	0.0321	0.0210
WISE *W1*, *W2* fluxes	-1.44×10^{-3}	1.28×10^{-4}	0.0314	0.0209
WISE *W1*, *W2* fluxes, observational errors used[d]	-1.55×10^{-3}	5.28×10^{-5}	0.0313	0.0209
WISE *W1*, *W2*, *W3* fluxes	-1.61×10^{-3}	-1.47×10^{-5}	0.0312	0.0206
WISE *W1*, *W2*, *W3* fluxes, observational errors used[d]	-1.54×10^{-3}	3.45×10^{-5}	0.0311	0.0205
VIKING *zYJHK_s* fluxes	-1.58×10^{-3}	-2.22×10^{-5}	0.0312	0.0202
ugri GAaP magnitudes +				
KiDS `A`, `B` + `FLUX_APER_10`_*ugri*	-1.71×10^{-3}	-1.16×10^{-4}	0.0311	0.0209
KiDS `FLUX_APER_{4, ..., 100}`_*r* + GAaP colours	-1.74×10^{-3}	-1.75×10^{-4}	0.0309	0.0206
KiDS `FLUX_APER_10`_*ugri* + GAaP colours	-1.71×10^{-3}	-1.23×10^{-4}	0.0313	0.0205
KiDS `A`, `B` + GAaP colours	-1.59×10^{-3}	-7.61×10^{-5}	0.0305	0.0204
KiDS `A`, `B` + WISE *W1*, *W2* fluxes	-1.45×10^{-3}	5.68×10^{-5}	0.0303	0.0205
KiDS GAaP colours + WISE *W1*, *W2* fluxes	-1.60×10^{-3}	-6.87×10^{-5}	0.0307	0.0202
KiDS `A`, `B` + VIKING fluxes	-1.70×10^{-3}	-1.25×10^{-4}	0.0307	0.0201
KiDS `A`, `B` + WISE *W1*, *W2*, *W3* fluxes	-1.43×10^{-3}	4.33×10^{-5}	0.0300	0.0200
KiDS GAaP colours + VIKING fluxes	-1.72×10^{-3}	-1.53×10^{-4}	0.0311	0.0199
KiDS GAaP colours + WISE *W1*, *W2*, *W3* fluxes	-1.50×10^{-3}	-9.24×10^{-6}	0.0303	0.0198
VIKING fluxes + WISE *W1*, *W2*, *W3* fluxes	-1.18×10^{-3}	2.47×10^{-4}	0.0302	0.0197
ugri GAaP magnitudes +				
KiDS `A`, `B` + `FLUX_APER_{4, ..., 100}`_*r* + GAaP colours	-1.73×10^{-3}	-1.93×10^{-4}	0.0302	0.0201
KiDS `A`, `B` + `FLUX_APER_4`_*ugri* + GAaP colours	-1.77×10^{-3}	-2.24×10^{-4}	0.0303	0.0201
KiDS `A`, `B` + `FLUX_APER_10`_*ugri* + GAaP colours	-1.69×10^{-3}	-1.67×10^{-4}	0.0304	0.0200
KiDS `A`, `B` + GAaP colours + VIKING fluxes	-1.59×10^{-3}	-1.07×10^{-4}	0.0299	0.0194
KiDS `A`, `B` + GAaP colours + WISE *W1*, *W2*, *W3* fluxes	-1.59×10^{-3}	-1.24×10^{-4}	0.0293	0.0192
KiDS GAaP colours + VIKING fluxes + WISE *W1*, *W2*, *W3* fluxes	-1.32×10^{-3}	7.55×10^{-5}	0.0296	0.0191
KiDS `A`, `B` + VIKING fluxes + WISE *W1*, *W2*, *W3* fluxes	-1.29×10^{-3}	7.42×10^{-5}	0.0290	0.0188
ugri GAaP magnitudes +				
KiDS `A`, `B` + GAaP colours + VIKING fluxes + WISE *W1*, *W2*, *W3* fluxes	-1.33×10^{-3}	3.57×10^{-5}	0.0288	0.0186
KiDS `A`, `B` + `FLUX_APER_10`_*ugri* + GAaP colours + VIKING fluxes	-1.75×10^{-3}	-2.46×10^{-4}	0.0299	0.0195
KiDS `A`, `B` + `FLUX_APER_10`_*ugri* + GAaP colours + WISE *W1*, *W2*, *W3* fluxes				
	-1.61×10^{-3}	-1.68×10^{-4}	0.0291	0.0190
ugri GAaP magnitudes +				
KiDS `A`, `B` + `FLUX_APER_10`_*ugri* + GAaP colours + VIKING fluxes + WISE *W1*, *W2*, *W3* fluxes				
	-1.56×10^{-3}	-1.48×10^{-4}	0.0287	0.0186

Notes. All the experiments used the same training and test sets of $\langle z \rangle = 0.23$; what was varied were the photometric parameters used for the photo-z derivation, as given in the first column. See text for details of the photometric parameters. [a] SMAD is the scaled median absolute deviation, converging to standard deviation for Gaussian distributions. [b] BPZ is independent of the training sets – the numbers are given for comparison (for the same GAMA test sample). These statistics are based on the KiDS pipeline solution. [c] ANNz2 overall KiDS-DR3 solution from Sect. 4.4 calculated for the same GAMA test set as in all the experiments of this table. [d] Errors on all the input parameters were directly used in photo-z derivation instead of relying on the internal error model of ANNz2. See Sects. 2.1 and 5.2 for details.