

THE-WIZZ: clustering redshift estimation for everyone

C. B. Morrison,^{1,2★} H. Hildebrandt,¹ S. J. Schmidt,³ I. K. Baldry,⁴ M. Bilicki,⁵
A. Choi,^{6,7} T. Erben¹ and P. Schneider¹

¹Argelander-Institut für Astronomie, Auf dem Hügel 71, D-53121 Bonn, Germany

²Department of Astronomy, University of Washington, Box 351580, Seattle, WA 98195, USA

³Department of Physics, University of California, Davis, One Shields Ave., Davis, CA 95616, USA

⁴Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool L3 5RF, UK

⁵Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, the Netherlands

⁶Scottish Universities Physics Alliance, Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK

⁷Center for Cosmology and Astro Particle Physics, The Ohio State University, 191 W. Woodruff Avenue, Columbus, OH 43210, USA

Accepted 2017 February 7. Received 2017 February 3; in original form 2016 September 28

ABSTRACT

We present THE-WIZZ, an open source and user-friendly software for estimating the redshift distributions of photometric galaxies with unknown redshifts by spatially cross-correlating them against a reference sample with known redshifts. The main benefit of THE-WIZZ is in separating the angular pair finding and correlation estimation from the computation of the output clustering redshifts allowing anyone to create a clustering redshift for their sample without the intervention of an ‘expert’. It allows the end user of a given survey to select any subsample of photometric galaxies with unknown redshifts, match this sample’s catalogue indices into a value-added data file and produce a clustering redshift estimation for this sample in a fraction of the time it would take to run all the angular correlations needed to produce a clustering redshift. We show results with this software using photometric data from the Kilo-Degree Survey (KiDS) and spectroscopic redshifts from the Galaxy and Mass Assembly survey and the Sloan Digital Sky Survey. The results we present for KiDS are consistent with the redshift distributions used in a recent cosmic shear analysis from the survey. We also present results using a hybrid machine learning–clustering redshift analysis that enables the estimation of clustering redshifts for individual galaxies. THE-WIZZ can be downloaded at <http://github.com/morriscb/The-wizz/>.

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

1 INTRODUCTION

Current and future photometric galaxy surveys are designed to measure the properties and evolution of galaxies as well as constrain cosmological parameters and the properties of the Universe. In order to enable this, accurate and unbiased estimates of galaxy redshifts are required to extract the maximum amount of information. Until recently, redshift information in photometric surveys was only gained through spectroscopic follow-up or photometric redshifts (photo- z s) from multi-band photometry. Many techniques exist for deriving photo- z s (see Hildebrandt et al. 2010, for a partial review); however, all these techniques rely on a calibration set of spectroscopic redshifts that is representative of the survey galaxy population. Such a sample of spectra is only possible for the shallowest surveys and still requires a significant amount of telescope time. For future deep, large-area surveys such as the Large Synoptic

Survey Telescope¹ (LSST), a sample of representative spectra will be even more difficult. Such challenges are presented in Newman et al. (2015).

An alternative and complementary method to photo- z s is that of clustering redshift estimation (clustering- z s). Clustering redshifts make use of the fact that galaxies with unknown redshifts reside in the same structures as galaxies that have known redshifts. Thus, spatial cross-correlations can be used to estimate the redshift distribution of the sample with unknown redshifts. The basic method bins the sample with known redshifts in z and then spatially cross-correlates each of these bins against the unknown sample. The amplitude of the resultant correlation can then be used to estimate the amount of redshift overlap and thus the redshift distribution of the sample with unknown redshifts. One of the first suggestions of such a method can be seen in Schneider et al. (2006) with the formalism for this method written out in Newman (2008) and later

*E-mail: morrison.chrisb@gmail.com

¹ <http://www.lsst.org/>

generalized in Schmidt et al. (2013) and Ménard et al. (2013) with quadratic estimators laid out in McQuinn & White (2013) and Johnson et al. (2017). The method has some drawbacks from sensitivity to galaxy bias both from the reference sample with known redshifts and the sample with unknown redshifts that can affect clustering- z s. However, suggestions to mitigate this bias exist in the literature (Newman 2008; Ménard et al. 2013; Schmidt et al. 2013).

Cross-correlation techniques are beginning to be applied to real data (Rahman et al. 2015, 2016b; Choi et al. 2016; Rahman, Ménard & Scranton 2016a; Scottez et al. 2016; Hildebrandt et al. 2017; Johnson et al. 2017) with an eye towards future surveys. A failing of this method, however, is that the current implementations of clustering redshifts are not as easy to use as their photo- z counterparts and nominally require spatial correlations to be run and rerun for each galaxy subsample of interest. This is a time-consuming process and could limit clustering redshift's adoption by the larger community. Suggestions exist such as producing clustering- z s in colour-colour space cells (Rahman et al. 2016a; Scottez et al. 2016), but this will have limitations for some samples and precludes the ability to weight galaxies in the clustering redshift estimation in the same way as in a given analysis or utilize additional information after the correlations are run for each cell. A more flexible method that separates the spatial correlation computation from the act of creating clustering redshifts would be ideal.

In this paper, we present THE-WIZZ,² a method for estimating redshift distributions from clustering designed for ease of use by survey end users. THE-WIZZ separates the difficult step of finding close angular pairs from the act of creating a clustering redshift estimate. In this way, the correlations between close pairs can be run once by the survey data pipeline and then an end user can create a clustering redshift estimate for their unique subsample of galaxies in a matter of, in some cases, seconds. THE-WIZZ can add to the legacy of galaxy surveys by producing a stable data product that can continue to be used by the astronomy community without a large amount of specialized software, much like how photo- z s are used today. THE-WIZZ can of course also be used by individuals with any data overlapping a spectroscopic sample allowing them to produce clustering- z s quickly and easily.

This paper is laid out as follows. In Section 2, we give an overview of the method and software including showing how it can be used in the context of a galaxy survey. Section 3 explains the data products we use to test THE-WIZZ. In Section 4, we show the resultant clustering redshift estimates and present a novel method of colour-redshift mapping made possible by the speed of THE-WIZZ. Section 5 discusses these redshift estimates and THE-WIZZ in the context of current surveys. Finally in Section 6, we present our conclusions with an eye towards future surveys such as LSST, *Euclid*³ and the Wide Field Infrared Survey Telescope⁴ (WFIRST). Throughout this analysis, we use the *Wilkinson Microwave Anisotropy Probe* 5 (Komatsu et al. 2009) cosmology for consistency between the code we use for spatial pair finding, STOMP,⁵ and THE-WIZZ. The choice of cosmology will, however, have little effect on the resultant clustering- z s (Newman 2008; Matthews & Newman 2010).

2 METHOD OVERVIEW

The methodology of THE-WIZZ is to separate the computationally intensive step of pair finding and angular correlation estimation

from the creation of a clustering redshift estimate for a given galaxy sample of unknown redshift, allowing for fast computation and re-computation of the output clustering redshift estimate. We do this by pre-computing and storing all pairs between a galaxy sample with known redshifts (hereafter known as the reference sample) and catalogue of galaxies with unknown redshifts (hereafter, the unknown sample) within a fixed physical radius around the reference galaxy. This is similar in concept to fast correlation codes pre-computing data structures for quick pair finding/correlation estimation. End users can then simply select their desired subsample from the unknown sample catalogue and match the catalogue indices of their sample into the data file containing the pairs using the provided software. THE-WIZZ then takes care of all the book keeping and produces a properly normalized estimate of the subsample's overdensity as a function of redshift that can then be converted into a clustering redshift estimate or estimated probability density function (PDF).

THE-WIZZ is thus extremely powerful for use within survey collaborations and as a legacy, value-added catalogue data product for users of the survey's data in the future. The software is designed to make creating clustering redshifts for any unknown sample nearly as easy as selecting in photo- z . This is especially powerful in the context of survey collaborations as each working group will likely have their own selections and weighting scheme for optimal signal-to-noise within the context of the science they are interested in. Without THE-WIZZ, this would require computation of the angular correlations and clustering- z s for each unknown sample in question, and if the samples a working group was using ever changed, the clustering- z s would have to be computed all over again. THE-WIZZ circumvents this problem by effectively computing the correlations for all galaxies in the unknown sample against the reference sample simultaneously, collapsing these measurements into a clustering redshift only when called with a user-specified subsample of galaxies. Additional data in newly observed areas can be easily appended in this data structure without having to rerun the full sample. The only time the pair finding portion of THE-WIZZ would have to be rerun is if the photometric detection catalogue of the survey were to fundamentally change (e.g. new detection algorithm, new thresholding, increased survey depth). THE-WIZZ makes use of mostly widely available and well-supported packages including the PYTHON-based astronomy library ASTROPY⁶ (Astropy Collaboration et al. 2013), making it even easier for end users to get set up and started.

Fig. 1 shows the data flow through THE-WIZZ with pre-computation of the correlations on the left and the user-selected catalogue on the right. The lower panel shows the part of THE-WIZZ that takes the pre-computed angular correlations and matches them with a catalogue of their specific unknown sample to produce an output clustering redshift distribution. This is the part of THE-WIZZ that the majority of users will see.

The remaining portion of this section goes into depth about the internals of both the pre-computation step PAIR_MAKER and the final clustering redshift estimation creation step PDF_MAKER.

2.1 PAIR_MAKER

The left-hand side of the data flow diagram shown in Fig. 1 shows the input and outputs of the THE-WIZZ program called PAIR_MAKER. The program does the majority of the calculations involved in masking catalogues, finding pairs and generating random points. The

² Available at <http://github.com/morriscb/The-wizz/>

³ <http://sci.esa.int/euclid/>

⁴ <http://wfirst.gsfc.nasa.gov/>

⁵ Available at <http://github.com/ryanscranton/astro-stomp/>

⁶ Available at <http://www.astropy.org>

THE-WIZZ: Data Flow

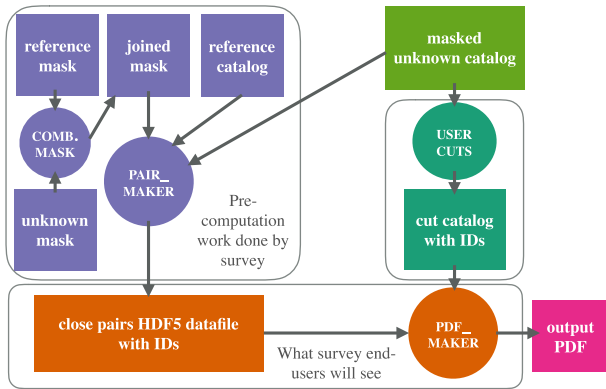


Figure 1. Flow chart of the inputs and output of THE-WIZZ. In the upper left, we have the work done by an individual survey in spatially masking the catalogue, running PAIR_MAKER and creating THE-WIZZ’s output HDF5 data file. The upper right shows a user selecting a sample from the masked catalogue for their own work. The lower portion is the end user matching their specific sample into the data file using PDF_MAKER and producing a resultant clustering redshift estimate for their sample without having to run any cross-correlations.

method we utilize is described in detail in Schmidt et al. (2013) and Ménard et al. (2013). We will describe some of our modifications and generalizations to that method in this subsection. The PAIR_MAKER software utilizes the STOMP spherical pixelization library for masking and pair finding. Further information on STOMP can be found in Scranton et al. (2002, 2005).

Our first step in using PAIR_MAKER is creating a combined mask of the area covered by both the unknown and reference samples. STOMP stores these resultant maps in a hierarchical pixel format that allows for storage of pixels at different resolutions for optimal file size and quick spatial searches. This masking map contains the observed area of the survey with bright stars, bad pixels, etc. masked out. Once this map is created, we can load both our reference and unknown samples with their indices into THE-WIZZ. PAIR_MAKER then creates a searchable quadtree on the STOMP hierarchical pixels using the unknown sample and also generates random points on the mask if requested and stores them in a similar quadtree.

Data: reference catalogue, unknown catalogue, R_{\min} , R_{\max} , randoms, data file

Result: Stored, sorted unknown ids and number of randoms, per reference object with meta-data

for *object in reference catalogue* **do**

```

  find object spatial region;
  find pair ids and distances within  $R_{\min} - R_{\max}$ ;
  sort ids and distances by id;
  find  $n$  randoms within  $R_{\min} - R_{\max}$ ;
  store sorted ids and distances in data file;
  store  $n$  randoms in data file;
  store object redshift in data file;
  store object spatial region in data file;

```

end

Algorithm 1: Pseudo-code describing the main loop of PAIR_MAKER.

Unlike Matthews & Newman (2010), Choi et al. (2016) or Johnson et al. (2017) in which they measure and fit the full angular correlation function in bins of redshift for the reference sample, we only measure the correlation amplitude in a single bin in projected radius. STOMP allows for easy conversion from a fixed radial bin in physical radius to an angular bin on the sky given the redshift of the reference object and an assumed cosmology. We use this binning for all measurements with THE-WIZZ. STOMP then finds all the pixels at a fixed resolution that cover this annulus and uses these pixels to search the unknown quadtree. THE-WIZZ then stores the unique index of each unknown object as well as the inverse distance from the reference object to the pixel centre. One of the key aspects of the method presented in Schmidt et al. (2013) and Ménard et al. (2013) is the signal-matched filtering and weighting of the galaxy pairs by the inverse projected physical distance to the reference object. Newman (2008) and Matthews & Newman (2010) show that clustering redshift measurements depend only weakly on the assumed cosmology so the clustering- z_S THE-WIZZ produces can be assumed to be general. This pair information is then stored for each unmasked reference object. If requested, the software repeats the process for the random sample but only stores the total number of randoms found and the sum of their inverse distances rather than storing the information for individual galaxies. These steps are sketched out in pseudo-code in Algorithm 1.

PAIR_MAKER can then repeat this process for a large number of requested radial bins using the same masked data set and randoms. For this analysis, we use a binning of $R = 100\text{--}1000$ kpc. We also computed bins similar to those of Schmidt et al. (2013) with physical radius bins of $R = 3\text{--}30$; $30\text{--}300$; $300\text{--}3000$ kpc. We combine some of these bins as $R = 3\text{--}300$; $30\text{--}3000$ kpc. Note that bins of abutting radii are not completely independent from one another. Given the coarse pixelization of STOMP in finding pairs, the bins are likely to overlap slightly (i.e. the $R = 3\text{--}300$ kpc bin is not simply the addition of its two child bins). Computing multiple scales allows the user to find the correct compromise between reduced sensitivity to non-linear galaxy bias (large scales) and signal-to-noise (smaller scales). The choice will likely depend on the sample used. Combining multiple scales could also be used as in the quadratic estimators of McQuinn & White (2013) and Johnson et al. (2017).

As an aside, it should be stated that any functional form of weighting by distance is possible with THE-WIZZ. Weighting by the inverse distance is conceptually simple and close to the roughly expected power-law scaling of the correlation galaxy function of $\gamma = 1.8$. The software allows for simple modifications of this weighting scheme and can be extended to any weighting as a function of projected physical distance. This weight function could be modified, for instance, to a similar weighting of McQuinn & White (2013) or Johnson et al. (2017) that attempts to optimally weight for number of galaxies and mitigation of non-linear scales. We leave it up to the user to decide what is best for their analysis with the default behaviour being inverse distance.

The output from the pair finding processes is stored in a custom data structure in HDF5 format for later use with PDF_MAKER. Intricate knowledge of this format or how it is used is not required to utilize THE-WIZZ. The unique indices of each unknown object as well as their inverse distance from the reference object are stored in sorted arrays for each reference object, for each scale considered. Several other data products are stored per reference object such as its redshift and the number of randoms around the object. We attempt to reduce the final file size through lossless data compression. In the end, the final size of the data files depends on the scales requested, number of reference objects and the density of the unknown sample. As an

example, the data file created for the analysis we present in Section 4 is roughly equal in size compared to the input unknown catalogue masked to the area covered by the reference sample. Currently, this size comes from using very simple and straightforward techniques and data structures to output the resultant pairs. This ratio of input catalogue to output is likely to improve as better and more efficient techniques are applied to the storage of the data.

2.1.1 Notes on STOMP regions

STOMP contains powerful internal methods for creating regions on the sky for spatial bootstrapping and jackknifing. Such regionation can be difficult given a complex survey mask and the requirement that regions be equal area and regular in shape. These regions are extremely useful for mitigating the effects of observing strategy and the density systematics that come with them. STOMP allows for the creation of regions that are roughly square and equal area. STOMP regions are what THE-WIZZ uses to compute spatial bootstrap errors on the clustering- z s and are thus extremely important. THE-WIZZ also uses said regions to significantly speed up the pair matching in PDF_MAKER. One should specify regions that are a compromise between observational errors and the scales desired. For instance, it may not be possible to run a scale that is larger than the size of individual pointings in a multi-epoch survey. The user of the software is encouraged to experiment with this variable for their own survey.

2.2 PDF_MAKER

PDF_MAKER is the part of THE-WIZZ that the large majority of users will interact with. It is the portion of the codebase that takes the resultant HDF5 data file created from PAIR_MAKER and combines it with the user's subsample and returns the clustering redshift estimation. The right-hand side of Fig. 1 shows the work a user of PDF_MAKER will perform in preparing to utilize THE-WIZZ for creating clustering redshifts. The user selects a subsample of galaxies from the same catalogue that was masked and used in PAIR_MAKER. The user then invokes PDF_MAKER with this subsample and the HDF5 data file output from PAIR_MAKER to create a clustering redshift estimate for their specific subsample as in the lower portion of Fig. 1. At the run time of PDF_MAKER, the user requests one of the scales stored in the HDF5, PAIR_MAKER data file and a redshift binning. PDF_MAKER then computes the natural estimator of overdensity (Davis & Peebles 1983)

$$\delta(z_i) = \frac{D_r D_u(z_i)}{D_r R(z_i)} - 1, \quad (1)$$

where $D_r D_u(z_i)$ are the pairs between the reference (r) and unknown (u) sample in redshift bin z_i . $D_r R(z_i)$ are the pairs between the reference sample and random positions drawn from the same mask as the unknown sample. During this calculation, the number of randoms are properly scaled to the requested subsample and any weights requested for the unknown sample are applied (e.g. shape weights, detection efficiency, photometric redshift posterior probabilities).

THE-WIZZ minimizes the amount of time spent matching pairs from the user-specified subsample by sorting the IDs of the subsample and matching them into the, already sorted, IDs stored around each reference object using a binary search tree. Algorithm 2 shows in pseudo-code the steps PDF_MAKER performs. The software also makes use of two methods of spatially locating the pairs for matching. First, THE-WIZZ takes advantage of the fact that many source detection programs return IDs that are partially sorted spatially. For

Data: sample catalogue, data file

Result: overdensity around each reference object

```

for ref-obj in data file do
    load stored unkn-ids around ref-object;
    rescale n randoms around ref-object to match sample;
    set n unkn-objs around ref-obj to zero;
    for unkn-id in sample catalogue do
        binary search for unkn-id in unkn-ids around ref;
        if unkn-ids contains unkn-id then
            | add 1 to n unkn-objs;
        end
    end
    store unkn-objs divided and scaled n randoms;
end

```

Algorithm 2: Pseudo-code describing the main loop of PDF_MAKER.

instance, SExtractor (Bertin & Arnouts 1996) returns IDs that are ordered in increasing y -axis position and then increasing x -axis position such that a sub-selection of increasing, ordered IDs will be localized between a x -min and x -max and thus localized spatially. The software recognizing this results in a speed-up of the analysis by a moderate amount, but spatially sorted IDs are not required by the code. Secondly, the software masks for the independent STOMP regions stored in the HDF5 pair file assuming that the input subsample likewise has information on the STOMP regions. For the data we use and clustering- z s we show in Section 4.1, THE-WIZZ will spend of the order of tens of seconds in calculating clustering- z s for scales less than 300 kpc and of the order of minutes for larger scales for a fixed number of cores. This allows users to compute and re-compute clustering- z s for any given sample in a tractable amount of time. The software for PDF_MAKER can also make use of multiple cores ensuring scalability to even larger data sets.

THE-WIZZ computes its errors through spatial bootstrapping utilizing the STOMP regions that were previously calculated in PAIR_MAKER. Thanks to these independent regions and clever book keeping, THE-WIZZ can compute thousands of bootstrap realizations and calculate errors nearly instantly. THE-WIZZ even allows for the storage of intermediate data products such as the overdensities in each region and the individual bootstrap samples, allowing one to propagate errors in the clustering redshift estimate into any analysis that utilizes the clustering redshift distributions THE-WIZZ produces.

An important differentiation between PAIR_MAKER and PDF_MAKER is that the latter does not require that STOMP be installed or run. PDF_MAKER uses very few non-standard PYTHON packages and those it does use can be easily installed through PIP or come with an installation of the popular ANACONDA⁷ distribution of PYTHON. Data products produced from THE-WIZZ's PAIR_MAKER can then be widely distributed as a value-added catalogue product with end users only needing to use PDF_MAKER to produce robust clustering- z s. This is the main power of THE-WIZZ in enabling science with clustering redshifts. In the remainder of the paper, we will show how the flexibility of THE-WIZZ enables unique uses of clustering redshifts, such as producing clustering redshift estimates for individual galaxies using machine learning.

2.3 Bias mitigation

Properly mitigating the effect of galaxy bias in clustering redshift estimates is essential to using these redshift distributions in any

⁷ <http://www.continuum.io/why-anaconda/>

scientific analysis. There is a large amount of literature on this topic, and we will not go into this in depth as it is not the focus of this paper. *THE-WIZZ* does not currently implement a technique for mitigating the effect galaxy bias, leaving the choice up to the user. In general, galaxy bias mitigation techniques can be thought of as a post-processing applied to the output of *THE-WIZZ*, even those of e.g. Newman (2008) and McQuinn & White (2013). We will point out however that *THE-WIZZ* is perfectly suited for many of the literature techniques suggested. One specific example is the technique of Schmidt et al. (2013) and Ménard et al. (2013), which shows that by pre-selecting a narrow redshift range of unknown objects, one can attempt to mitigate the effect of galaxy bias. Indeed, Rahman et al. (2016b) showed that this is the case when selecting Sloan Digital Sky Survey (SDSS) galaxies in narrow photo- z bins and summing the individual clustering- z s to create the clustering- z s for magnitude-limited samples. Rahman et al. (2016a,b) and Scottez et al. (2016) showed that one can use selections in galaxy colour to achieve a similar effect. *THE-WIZZ* is ideal for producing such pre-selected clustering- z s as it enables the redshift estimation of any subsample of the unknown galaxy sample considered. Since the galaxy bias removal is a post-processing step, function forms of the galaxy bias could be provided along with the data files to run *THE-WIZZ* for a given set of data. This could be very powerful in enabling science for end users and add to legacy value in the context of a galaxy survey.

For the results shown in Section 4.1, we implement a simplified version of these bias mitigations that is similar to that shown in fig. 5 of Schmidt et al. (2013) and the ‘no bias’ photo- z sampling from Rahman et al. (2016b). This simplified bias mitigation pre-selects in narrow redshift bins using photo- z s. For narrow redshift distributions, the galaxy bias evolution is close to a constant over the peak of the redshift distribution. Clustering- z s using these narrow selections can then be summed together, creating a clustering- z measurement for a larger redshift range that has much of the effect of galaxy bias mitigated. This assumes that the evolution of the galaxy cross-bias between the unknown and reference samples is smooth and well behaved, an assumption that is likely broken when, for instance, the reference sample’s selection changes [e.g. switching from luminous red galaxies (LRGs) to quasi-stellar objects (QSOs)]. This can be thought of as a first-order correction to the galaxy bias. Precision cosmology measurements will likely need to further mitigate the effects of galaxy bias using the spectroscopic bias evolution for example (see Rahman et al. 2016b; Scottez et al. 2016); however, for analyses that require less precision, this simplified method is an ideal way of using clustering- z s in a straightforward manner. If one does not have access to photo- z s, colour or brightness cuts could also be employed as long as they represent fairly narrow selections in redshift. We follow the pre-selection in photo- z bias mitigation technique in Section 4.1.

3 DATA

We use several different sets of reference data and one set of unknown data in demonstrating the capabilities of *THE-WIZZ*. The data come from the large spectroscopic catalogues of the SDSS and the Galaxy and Mass Assembly (GAMA) survey. Throughout this analysis, we use photometric data with unknown redshifts from the Kilo-Degree Survey (KiDS). The data we use are an excellent test bed for the *THE-WIZZ*’s ability to scale to future, high data volume surveys.

3.1 Photometric, unknown data

The photometric data we use come from KiDS. KiDS represents a large-area lensing survey that shows *THE-WIZZ*’s ability to scale to future data sets such as LSST, *Euclid* and WFIRST.

3.1.1 The Kilo-Degree Survey (KiDS)

The ongoing Kilo-Degree Survey⁸ (KiDS; de Jong et al. 2015) is a 1500 deg² survey observed with OmegaCAM on the VLT Survey Telescope (VST) in SDSS-like u, g, r, i bands down to 5σ limiting magnitudes of 24.3, 25.1, 24.9 and 23.8 AB, respectively. The survey is designed for weak lensing and has a median seeing of better than 0.7 arcsec in the r band. Further details on the survey can be found in de Jong et al. (2015), Kuijken et al. (2015) and Hildebrandt et al. (2017). For this analysis, we use catalogues and automated masks of bright stars and image defects produced by Astro-WISE (Valentijn et al. 2007; Begeman et al. 2013) and THELI (Erben et al. 2005; Schirmer 2013). Magnitudes and colours are produced using GaaP, a seeing Gaussianization process that produces consistent aperture photometry across the different observed bands (Kuijken 2008; Kuijken et al. 2015). Initial detection catalogues for photometry use SExtractor (Bertin & Arnouts 1996). For photometric redshifts, we use a modified version of the Bayesian Photometric Redshifts (Benítez 2000, BPZ) code as described in Hildebrandt et al. (2012, 2017).

The data we use from KiDS are a currently non-public, early data product dubbed KiDS-450. This iteration of the survey has an area of roughly 450 deg² and covers all GAMA fields in all four KiDS bands. The survey is also covered by spectra from the north Galactic cap of the SDSS. After applying the full masking that intersects with the northerly GAMA and SDSS fields, we have a total area of ~ 170 deg². In this analysis, we mimic the cuts described in Hildebrandt et al. (2017) for comparison to the redshift distributions shown therein. We utilize the shape weights produced by *Lensfit* (Miller et al. 2013; Fenech Conti et al. 2016) as weights for each object to further mimic this selection. These weights also act as a magnitude limit, returning low and zero weights for galaxies with $r > 25$. The cuts we make also exclude all galaxies with $r < 20$. In total, our sample of photometric, unknown objects is 3959 558 total galaxies with weight > 0 . We make a further cut for the analysis we present in Section 4.2, additionally requiring that the GaaP magnitude in each band has a value of > 0 assuring that each magnitude is observed (not necessarily detected) for each object. For this sample, GaaP values of 99, defined as non-detected in any band, are replaced with the limiting magnitude in that band.

3.2 Spectroscopic, reference data

We use two spectroscopic surveys for our reference sample, the GAMA survey and SDSS Data Release 12 (DR12). The distribution as a function of redshift for both surveys within the overlapping area of KiDS is shown in Fig. 2. As seen in the figure, GAMA dominates for a low redshift while SDSS dominates for a higher redshift. In total, there are 135 567 galaxies in the sample we use spanning a redshift range of $0.01 < z < 7.0$. At redshifts $z > 1.0$, we rely almost exclusively on spectroscopic QSOs from the SDSS DR12 catalogue. In addition to using the quality cuts provided by each

⁸ <http://kids.strw.leidenuniv.nl/>

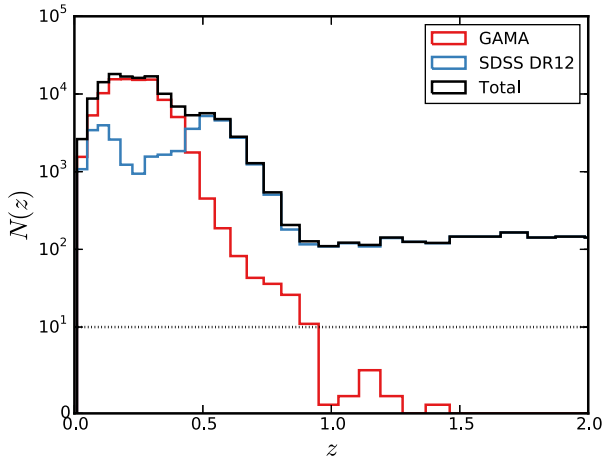


Figure 2. Symmetric-log plot of the number of spectra overlapping with the current KiDS coverage as a function of redshift. The GAMA survey is the dominant sample for low redshifts with SDSS dominating for high redshift. The large number of objects above $z = 1.0$ are spectroscopic QSOs. The data are binned linearly in $\ln(1+z)$ and follow the exact binning that we will later use in our clustering redshifts. Above the dotted line, the data are plotted logarithmically in $N(z)$, below they are plotted linearly. For this plot, we show galaxies in the GAMA catalogue that have spectral redshifts from SDSS as SDSS galaxies.

survey, we also reject all spectra within a 2 arcsec radius of each other to remove duplicate objects.

3.2.1 GAMA survey

We make use of non-public spectroscopic data, dubbed GAMA-II, from the GAMA survey⁹ (Driver et al. 2009; Baldry et al. 2014; Liske et al. 2015). GAMA is a magnitude-limited spectroscopic survey covering over 286 deg^2 on the Anglo-Australian Telescope using the AAOmega multi-object spectrograph. The GAMA survey is designed to study galaxy and mass evolution at low and intermediate redshifts; however, the spectra can also be used in cross-correlation studies such as this one. For our analysis, we make use of the equatorial fields of GAMA that overlap with KiDS dubbed G09, G12 and G15 corresponding to their RA centre. These fields are primarily observed to a limiting magnitude of $r < 19.8$ over 180 deg^2 . We select spectroscopic redshifts from the survey that satisfy their ‘main sample’ criteria ($\text{SURVEY_CLASS} \geq 3$) and have a redshift quality value of $nQ \geq 3$. GAMA contains galaxies from other spectroscopic surveys including SDSS to reach its level of completeness. We reject any galaxy from the SDSS catalogue that is within 2 arcsec of a GAMA catalogue galaxy to avoid duplicate redshifts between the two catalogues. We make use of the GAMA redshift completeness masks to exclude bad area from the survey and limit the area we must search for pairs in. After masking for the KiDS and GAMA combined area, we have 101 deg^2 with a total number of 94 694 unique spectroscopic galaxies from the GAMA catalogue.

3.2.2 SDSS DR12

We make use of spectra from the 12th data release from the SDSS¹⁰ (York et al. 2000; Eisenstein et al. 2011; Alam et al. 2015). SDSS not

only adds low- and intermediate-redshift spectra but also spectroscopic QSOs that allow us to produce clustering- z s out to very high redshift. For our purposes, we make use of all galaxy spectra from the survey that overlap with KiDS. This nominally includes galaxies from the SDSS main sample (Strauss et al. 2002), Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013) galaxies both from the LOWZ and CMASS samples, and the aforementioned QSOs (Ross et al. 2012). The galaxies we utilize are those defined as ‘Science Quality’ from the SkyServer catalogue and have a redshift quality selection with $z\text{Warning} = 0$ (Bolton et al. 2012). As stated previously, objects are also checked for duplication between SDSS and GAMA. The mask we use for this analysis comes from converting the SDSS Mangle polygons into STOMP format. This STOMP map was previously used in the analyses of Schmidt et al. (2015) and Rahman et al. (2015, 2016b). The overlapping area between the current coverage of KiDS and SDSS/BOSS is 170 deg^2 containing 40 873 objects in total.

4 CLUSTERING REDSHIFTS

In this section, we show clustering- z s produced by THE-WIZZ from various subsamples of KiDS as the unknown samples with the reference data coming from SDSS and GAMA. This is not the first time a clustering redshift technique has been applied to the KiDS data, with lower redshift and smaller galaxy sample results shown in Hildebrandt et al. (2017) and Johnson et al. (2017) using different clustering redshift estimators and wider binning in photo- z . We compare to the results of Hildebrandt et al. (2017) in Section 5.1.

We show the power of THE-WIZZ in producing clustering- z s using the same catalogues and data files. The clustering- z s for a given unknown data sample are all produced from the same data; we only change how we select the given subsamples used. Throughout this section, the PAIR_FINDER portion of the software is only run once. We start in Section 4.1 by producing clustering redshifts using photometric redshift peak probability (z_B) as a pre-selection as suggested in Ménard et al. (2013) and Schmidt et al. (2013) and shown in Rahman et al. (2016b) and Scottez et al. (2016). Then in Section 4.2, we introduce a novel technique where we estimate the redshifts of individual objects using a k -dimensional spatial search tree (kdTree) based method that allows us to select the k -nearest neighbours to an object in colour-magnitude space, run the software on those neighbours and produce a clustering redshift estimate for individual galaxies. For all the analyses we present in this section, we use the same randoms for use in the natural estimator. These are drawn to have a size of 10 times the total photometric sample. This means that every subsample has a large number of randoms compared to the number of objects used in the subsample.

We estimate errors and covariances by spatially bootstrapping 1000 times over 279 independent spatial regions as defined by STOMP. The spatial regions approximate the individual 1 deg^2 pointings from KiDS. We also separate the analysis into two parts: computing the overdensities in regions where SDSS and GAMA both overlap KiDS (representing 166 regions) and the regions where only SDSS overlaps KIDS (113 regions). We then combine these regions by spatial bootstrap that smoothly joins the two surveys’ redshift overlap. Throughout this analysis, we measure the cross-correlation amplitude on physical scales between $R = 0.1$ and 1.0 Mpc . We bin the reference galaxies in redshift with 50 bins equally spaced in $\ln(1+z)$ from $0.01 < z < 6.0$. We only plot and normalize the data to a redshift of $z = 2.0$ for clarity and to compare to the redshift distributions from Hildebrandt et al. (2017). Because of

⁹ <http://www.gama-survey.org/>

¹⁰ <http://www.sdss.org/>

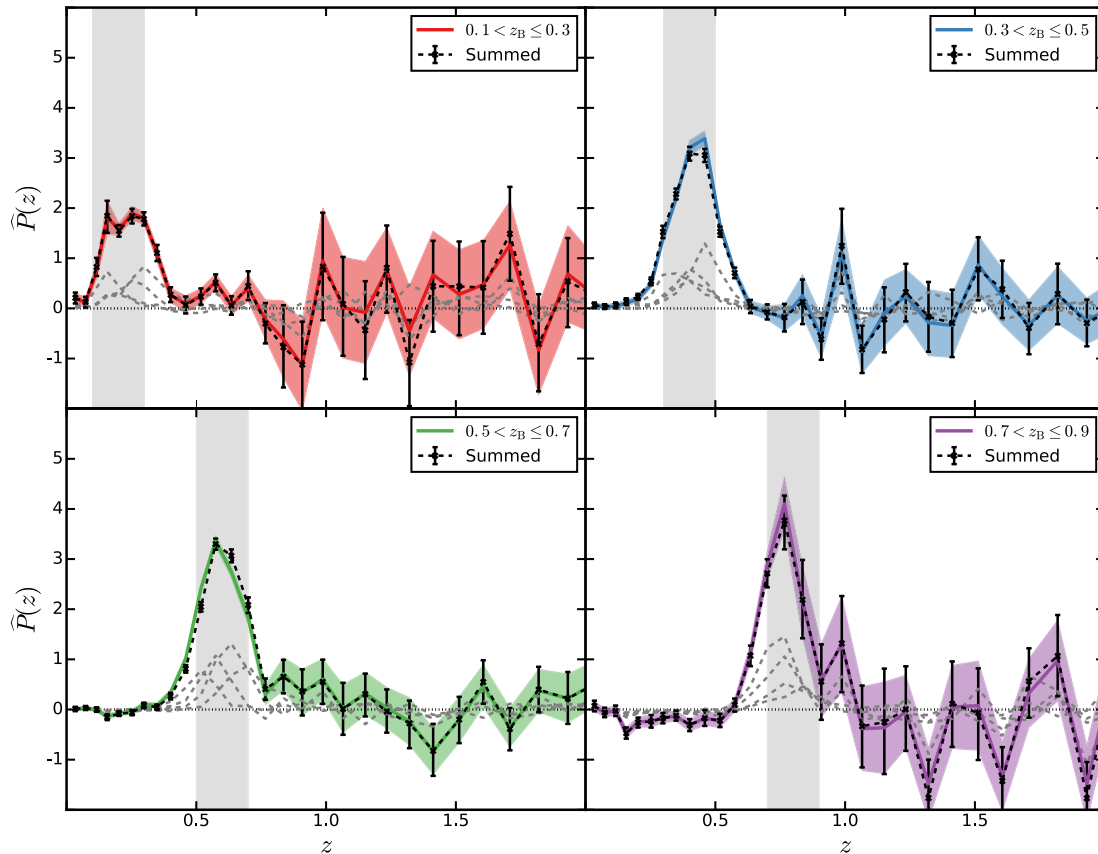


Figure 3. Raw and summed clustering- z s produced by `THE-WIZZ` using objects from KiDS selected in z_B as the unknown sample and GAMA and SDSS spectra as the reference sample normalized into an estimated PDF. Coloured bands are clustering- z s from selections in z_B mimicking the bins of Hildebrandt et al. (2017, CS bins). The light grey regions show the selection in photo- z . Grey dashed lines are the cluster- z s produced by dividing the CS bins into four sub-bins with $\Delta z_B = 0.05$ normalized by their number of objects relative to the CS bin. The grey dashed lines appear to sum up to the CS bin as a function of redshift suggesting that the galaxy bias in the clustering redshift estimate is well behaved. Black data points are the resultant clustering- z from normalizing, summing and averaging the individual spatial bootstraps of the sub-bins into the full CS bin. The bins were all selected from the same catalogue and use the same `THE-WIZZ` data file demonstrating how clustering- z s can be quickly created for a variety of samples using `THE-WIZZ`.

measurement noise, spatially dependent survey systematics and changes in unknown galaxy selection function, some points in the clustering- z s are negative. We treat these negative points by inverse variance averaging them with neighbouring bins until all points are positive definite. Without this smoothing, negative points with large values and error bars will bias the norm. This is also true of points with large positive values and error bars. This allows us to properly convert the clustering- z s into PDF estimates assuming that the bias is well mitigated. We compute the normalization using this adaptive smoothing but plot the data as measured. Normalizations to transform the overdensities into an estimated redshift PDF are computed using a trapezoidal sum with fixed end points of $z = 0.01, 2.0$.

4.1 Photo- z selection

Ménard et al. (2013) and Schmidt et al. (2013) demonstrate that one way to mitigate the effect of galaxy bias in clustering redshifts is to utilize colour or photometric redshift information to pre-select a sample of galaxies in a narrow redshift range, making the galaxy bias as constant as possible. One can then add clustering- z s of these pre-selected samples together using their relative numbers to produce the redshift distribution for a larger sample where the effect of the bias has been mitigated. Rahman et al. (2016b) and Scottez et al. (2016) apply this method to real data from SDSS and the Canada–

France–Hawaii Telescope Legacy Survey, respectively, and show that indeed the bias is mitigated by using these narrow redshift samples. This technique works best when the resultant redshift distributions are singularly peaked and narrow. If the distribution is found to have long tails in z or is multiple peaked, the galaxy bias mitigation will not be as robust.

The design of `THE-WIZZ` enables this kind of clustering- z and bias mitigation very simply. Subsamples can be selected and re-selected without having to rerun any correlations, significantly increasing the ease at which this method can be applied. We apply a simplified version of the bias mitigation of the previously mentioned clustering- z analyses that we describe in detail in Section 2.3. We attempt to recreate the photometric redshift distributions from the KiDS-450 cosmic shear (hereafter referred to as the CS bins) results (Hildebrandt et al. 2017) as a test of `THE-WIZZ`. This is a sample of four redshift bins with a width of $\Delta z_B = 0.2$, spanning the range of $0.1 < z_B \leq 0.9$ selected by the peak of the redshift posterior, z_B . We further divide each of these bins into four smaller photo- z sub-bins with a width of $\Delta z_B = 0.05$. This selection is pushing the limits of the errors of the photo- z s that are similar in size or slightly larger than $\Delta z_B = 0.05$ for some redshifts.

Fig. 3 shows the clustering redshifts produced by running `THE-WIZZ` on the CS bins plotted as coloured bands. The CS bins are normalized to a sum of 1 over the range $z = 0.01–2.0$. The

clustering- z of the CS bin agrees largely with previous results from Hildebrandt et al. (2017), especially the detection of a significant tail to high redshift in the $0.1 < z_B \leq 0.3$ bin. In addition, the high-redshift bins appear to be largely free of low-redshift interlopers, which is again in agreement with the KiDS-450 results. We show a direct comparison to the distributions in Hildebrandt et al. (2017) in Section 5.1. The clustering- z s also agree in overall shape and peak position compared to the previous results. The sub-bins, shown as grey dashed lines, are normalized to the number of objects they contain relative to their corresponding CS bin in addition to sum normalization. These bins are single peaked, normalizable and appear to properly sum to the full CS bin. This suggests that the galaxy bias is already fairly constant across the redshift range shown. If the bias were evolving strongly with redshift for either the reference or unknown sample, one would likely see discrepancies between the sub-bins and CS bins.

In order to add the sub-bins together to create the CS bins as well as the larger $0.1 < z_B \leq 0.9$, total spanning bin, we make use of the 1000 bootstrap samples. First, we ensure that the spatial bootstraps we draw for each subsample are the same so we can compute proper errors and covariances both within each CS bin we create from the summed sub-bins and also covariances between the summed CS bins. We first compute one normalization from the average of the spatial bootstraps for each sub-bin. This is due to each bootstrap realization being too noisy to properly compute a normalization. We compute all of the normalizations in a range $z = 0.01\text{--}2.0$ except for the $0.7 < z_B < 0.9$. Here we use $z = 0.3\text{--}2.0$ as our normalization range due to the significant low-redshift negative peak causing the computation to not converge. The redshift distributions of Hildebrandt et al. (2017) show no significant amplitude at $z < 0.3$ so we do not expect this cut to bias our results significantly. We apply these sub-bin normalizations to each of the sub-bins' spatial bootstraps and also multiply by the number of galaxies in each sub-bin sample for each spatial bootstrap. We then sum these bootstraps together to create a new set of 1000 spatial bootstraps for each summed CS and total bin. We then compute the median, low-side and high-side errors by calculating the 16th, 50th and 84th percentiles from the spatial bootstraps. We do this as the percentiles are much more stable than the simple mean and variance. We also calculate the mean and median of each of the redshift bins. The mean is calculated in the same way as the normalization using a trapezoidal sum while the median is taken as the point where the cumulative density function has a value of 50 per cent. We compute mean and median on the averaged, positive definite clustering- z s for each bootstrap and then compute the same percentiles as mentioned previously for central values and errors.

The black data points in Fig. 3 show the results of this process for each of the four CS summed bins. The summed data have slightly larger error bars than that of the CS clustering- z s largely due to the extra normalization step during the addition. The shapes of the clustering- z s between the summed and CS clustering- z s are similar but there are slight differences. These differences mainly show up in the $0.3 < z_B \leq 0.5$ bin. This bin finds its peak slightly shifted to higher redshift relative to the CS clustering- z . There are also slight differences in the peak of each of the other bins. We show the total, z_B spanning bin in Fig. 4. This bin has its low-redshift amplitude increased and intermediate redshifts suppressed in the summed clustering- z relative to the raw clustering- z . If one assumes that the galaxy bias is increasing with redshift for both the reference and unknown samples, this is the trend one should expect as the bias exaggerates the amplitude of the clustering- z s at high redshift compared to the truth. This is also reinforced by the fact that the

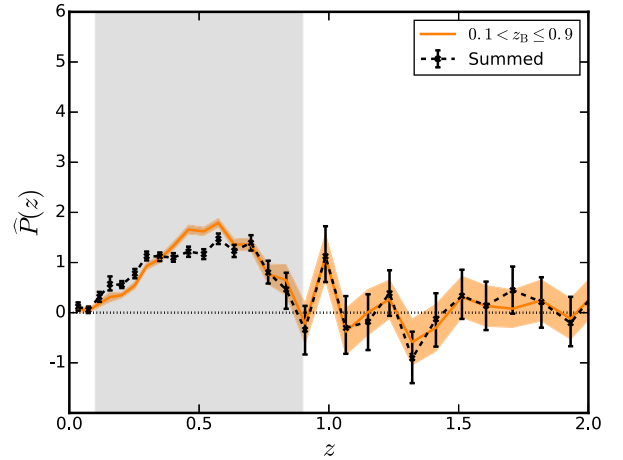


Figure 4. Raw and summed clustering- z s for the total, spanning bin of $0.1 < z_B \leq 0.9$. The orange coloured band shows the clustering- z result from running THE-WIZZ on the full $0.1 < z_B \leq 0.9$ sample. The black data points are the clustering- z created from summing the spatial bootstraps of the clustering- z s from the 16, $\Delta z_B = 0.05$ sub-bins into the total bin. The light grey region shows the z_B selection. We sum the clustering- z s together in this manner to mitigate galaxy bias in the clustering- z s as suggested in Ménard et al. (2013) and Schmidt et al. (2013). These works show how clustering- z s of wide distributions in z are much more susceptible to the effect of galaxy bias than narrow selections. The difference between the raw and summed clustering- z s shows the effect of this bias mitigation. The design of THE-WIZZ is well suited for bias mitigation strategies such as this.

peak positions of the redshift bins with narrower distributions are largely unchanged between the summed and CS clustering- z s. The increase in noise at higher redshift is likely twofold. First, there are many fewer reference galaxies at these redshifts. Secondly, the increase in reference bias accentuates any marginal correlation at these redshifts, increasing the noise. The latter is partially mitigated by the narrow redshift bins but would likely require explicit removal for the reference bias to be completely accounted for.

Table 1 summarizes the single point statistics we measure for the original and summed CS clustering- z s. For the majority of bins, both the mean and median of the clustering- z s are consistent within their error bars between the raw CS clustering- z and the summed version.

4.2 Colour selection with kdTrees

We can also utilize the colours of the unknown sample objects themselves to determine a mapping from colour to redshift rather than relying on photo- z to make the mapping for us. Such selections have been carried out in Ménard et al. (2013), Rahman et al. (2016a,b) and Scottez et al. (2016) to select narrow distributions in colour and thus narrow distributions in redshift. We can then compute the clustering redshifts of these colour selected samples for a colour-colour cell and create an estimate for most subsamples of galaxies by assigning them to a cell.

A major failing of this method is the high-dimensional partitioning of colour-colour space and efficiently populating each partition. A way around this is to use machine learning to reduce the dimensionality of the problem. This dimensionality reduction allows for the use of any galaxy quantity that correlates with redshift to arbitrary complexity within the limits of the machine learning algorithm chosen. We chose a relatively simple method by creating a kdTree

Table 1. Properties and summary statistics of the clustering redshift estimates for the KiDS-450 CS sample using both the clustering- z from running on the CS sample directly (z_{CS}) and after summing the clustering- z sub-bins from the CS sample (z_{sum}). z_{HH} are the median and mean results from the spectroscopic re-weighting scheme presented by Hildebrandt et al. (2017) in their table 1 as z_{DIR} .

Bin	z_{B} range	No. of unknown objects	Median(z_{CS})	$\langle z_{\text{CS}} \rangle$	Median(z_{sum})	$\langle z_{\text{sum}} \rangle$	Median(z_{HH})	$\langle z_{\text{HH}} \rangle$
1	$0.1 < z_{\text{B}} \leq 0.3$	1199 854	$0.596^{+0.462}_{-0.258}$	$0.751^{+0.176}_{-0.366}$	$0.357^{+0.574}_{-0.062}$	$0.698^{+0.308}_{-0.354}$	0.418 ± 0.041	0.736 ± 0.036
2	$0.3 < z_{\text{B}} \leq 0.5$	940 381	$0.464^{+0.032}_{-0.015}$	$0.488^{+0.058}_{-0.025}$	$0.458^{+0.022}_{-0.022}$	$0.473^{+0.218}_{-0.023}$	0.451 ± 0.012	0.574 ± 0.016
3	$0.5 < z_{\text{B}} \leq 0.7$	951 747	$0.633^{+0.018}_{-0.015}$	$0.683^{+0.138}_{-0.048}$	$0.649^{+0.028}_{-0.018}$	$0.670^{+0.117}_{-0.049}$	0.659 ± 0.003	0.728 ± 0.010
4	$0.7 < z_{\text{B}} \leq 0.9$	867 576	$1.255^{+0.372}_{-0.363}$	$0.969^{+0.313}_{-0.264}$	$1.276^{+0.310}_{-0.300}$	$0.985^{+0.269}_{-0.214}$	0.829 ± 0.004	0.867 ± 0.006
1–4	$0.1 < z_{\text{B}} \leq 0.9$	3959 558	$0.604^{+0.037}_{-0.037}$	$0.704^{+0.113}_{-0.074}$	$0.606^{+0.051}_{-0.044}$	$0.643^{+0.155}_{-0.071}$	–	–

in a space defined by several galaxy properties. This is similar to colour-space re-weighting techniques for photo- z s described in Lima et al. (2008) and Cunha et al. (2009). By using this kdTree, we can create clustering redshifts for individual galaxies by matching a single galaxy into the kdTree and measuring a clustering- z on the self-similar objects that the kdTree returns. This method can be extremely useful for survey users interested in individual or small samples of unique galaxies. It is also possible to use this cluster- z as a prior for Bayesian-based photo- z methods. This will also be useful for predicting clustering- z s in surveys that are observed with similar band pass filters but contain little or no spectroscopic overlap by matching their objects into the survey with spectra.

For this analysis, we use a sub-set of the full catalogue, limiting ourselves to the KiDS area intersecting both SDSS and GAMA. We also ensure that the objects are ‘observed’ in each band that is the GaaP magnitude returns > 0 . In total, roughly 2.8 million galaxies remain in this sample compared to the 4 million previously used. We use the three GaaP colours $u - g$, $g - r$, $r - i$ as well as the r -band magnitude as the space to create our kdTree in. We treat these colours and magnitudes similarly to that of BPZ where non-detections in a given band are replaced with the limiting aperture magnitude in the appropriate band. We create the four-dimensional kdTree after we standardize the colours and r magnitude to have mean of 0 and variance of 1. This regularizes the tree and prevents dimensions with large variance from dominating the Euclidean distances and therefore the computed neighbours. We make use of the package CKDTREE from SCIPY¹¹ to create the kdTree.

For each unknown object, we then match the same properties we created the kdTree with into the tree and then return the nearest 4096 unknown objects with similar properties as identified by the kdTree. Fig. 5 shows the median distance of the 4096 objects to each input, matched object in colour-magnitude space for the samples we consider. These data can be used as a quality statistic, removing objects that were matched with large distances relative to the rest of objects matched in. The distances plotted in this figure are plotted in standard deviations relative to the normalized colour-magnitude distributions. It is likely that objects matched with a median distance of larger than 1σ are not well represented in colour-magnitude space and will produce inaccurate clustering- z s. We can then input these 4096 objects into THE-WIZZ to produce a clustering redshift estimate for the individual object we matched into the kdTree. We select 4096 objects as it gives us relatively stable statistics per STOMP region ($N \sim 24$) and is not so wide as to have too many of the median distances beyond 1σ in colour-magnitude space.

To test our method, we first select galaxies with photo- z values of $z_{\text{B}} = 0.2, 0.4, 0.6, 0.8$, the mid-points of the CS bins. We make this

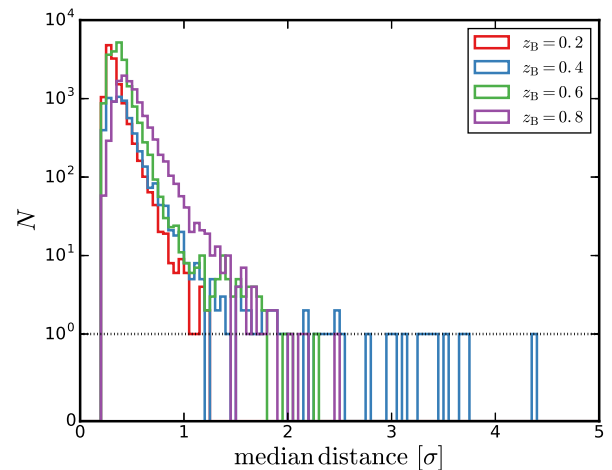


Figure 5. Symmetric-log plot of the histograms of the median distance in normalized units of the 4096 objects matched to the input object from the kdTree. The different curves are for each of the photo- z selected galaxy samples we use to test the method. That most of the median distances are below a value of 1σ away in colour-magnitude space lends credence that the self-similar galaxies selected from the kdTree are representative of the input galaxy. Above the dotted line, the data are plotted logarithmically, and below they are plotted linearly.

selection to have a rough idea of what the redshift is before creating the clustering- z s for comparison. The kdTree is created from the full catalogue and has no direct knowledge of redshift. We create single galaxy clustering- z s for each of the 12 654, 5968, 20 292, 10 882 galaxies in each sample, respectively. Fig. 6 shows the individual clustering- z s for each galaxy in these samples as grey lines. Darker regions represent redshifts where the clustering- z s are similar. The data points and bars shown are the median ± 34 per cent showing the dispersion of the individual clustering- z s to give a sense of how the clustering- z s are distributed. For clarity, we normalize the individual distributions to a redshift range of $z = 0.01$ – 1.0 when plotting. The relatively few unknown objects we use and the few high-redshift QSOs in this footprint prevent interpretation of the clustering- z s beyond $z = 1.0$.

5 DISCUSSION

The clustering redshifts shown in the previous section demonstrate the flexibility of THE-WIZZ in producing clustering- z s for any subsample of the data without having to rerun any two-point correlations. In this section, we will discuss our results specifically in the context of the redshift distributions used in Hildebrandt et al. (2017) and some non-cosmology applications.

¹¹ <http://www.scipy.org/>

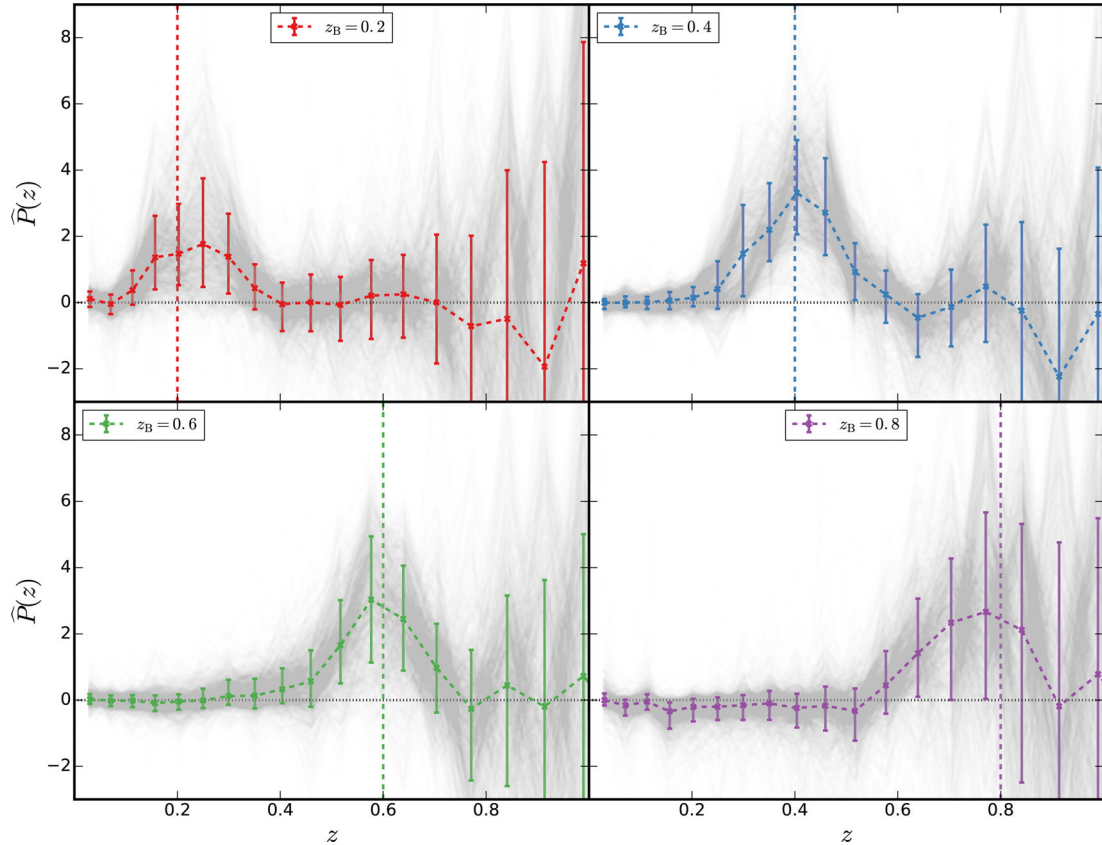


Figure 6. Single galaxy clustering redshift estimates for each of the samples we compared against. The grey lines are the individual clustering- z s produced by finding self-similar galaxies in colour–magnitude space for a random subsample of input objects from the photo- z selection. The coloured data points are the median \pm 34 per cent dispersion of the full sample of input object clustering- z s. We use photo- z z_B to select a test sample but create the kdTree and self-similar galaxies using the full catalogue with no explicit redshift information. The peak redshift found by this method agrees with the photo- z estimate validating this hybrid machine learning–cluster- z approach.

5.1 Comparison to KiDS-450 CS

In Fig. 7, we directly compare the data points from the KiDS-450 CS results with those of clustering- z s from THE-WIZZ. Overall, the clustering- z distributions presented here confirm the distributions shown in Hildebrandt et al. (2017). The canonical redshift distribution from KiDS-450, the z_{DIR} distribution that re-weights spectroscopic galaxies in colour space to account for non-representative spectra in calibrating photo- z s, agrees with our results with some allowance for sample variance. In addition, a previous, pre-THE-WIZZ clustering- z code (CC in Fig. 7) also confirms our redshift distribution. This CC clustering- z was created using only 1.6 deg^2 of spectra from zCOSMOS (Lilly et al. 2009) and DEEP2 (Newman et al. 2013) and uses the Newman iteration (Newman 2008) to mitigate bias suggesting again that the bias of the clustering- z s presented here is well behaved.

When comparing the mean and median redshifts of the summed clustering- z s to the redshift estimates from the direct calibration, z_{DIR} , we find gross agreement between the two methods. Comparing to Hildebrandt et al. (2017, table 1), we find that both the means and medians are consistent to within 1σ or 2σ when summing the square of the admittedly large errors. These errors come largely from uncertainty in the amplitude of the high-redshift tail in the clustering- z s. There are only 100 reference spectra per redshift bin for this part of the clustering- z leading to the large errors. There are several problems with the clustering- z s presented that we discuss here.

Around $z = 1.0$, there is a feature in the clustering redshifts that shows up in each of the clustering- z s shown in Fig. 3. Given its position in redshift, this feature likely comes from the switch between SDSS galaxies and QSOs. GAMA is also not contributing any more galaxies at this point as seen in Fig. 2. This negative correlation could be suppressing the measurement of the high-redshift tail that is seen in the z_{DIR} method of Hildebrandt et al. (2017). These negative correlations were also seen in the cross-correlation technique used in that paper and seem to be a feature of the data rather than a failing of THE-WIZZ. Negative amplitudes can be caused by incorrect masking or by extremely dense large-scale structure as shown in Rahman et al. (2015). We did not account for these overdensities by ‘cleaning’ in this work. Rahman et al. (2015) show that this cleaning is required to remove excess positive correlation and excess noise at redshifts $z < 0.2$. With THE-WIZZ however we do not observe such excess correlation when comparing it to the software of Rahman et al. (2015) when using the data. This could be due to the code of Rahman et al. (2015) using signal-matched filter weights in θ rather than R_{physical} . This adds an extra factor in of the angular diameter distance in the signal-matched weights Rahman et al. (2015) and similar codes use that may accentuate overdensities at low redshift causing these excess amplitudes. When necessary, however, the Rahman et al. (2015) cleaning step can simple be thought of as a pre-processing step on the reference sample before it is input into THE-WIZZ. It does not require any change to the algorithm.

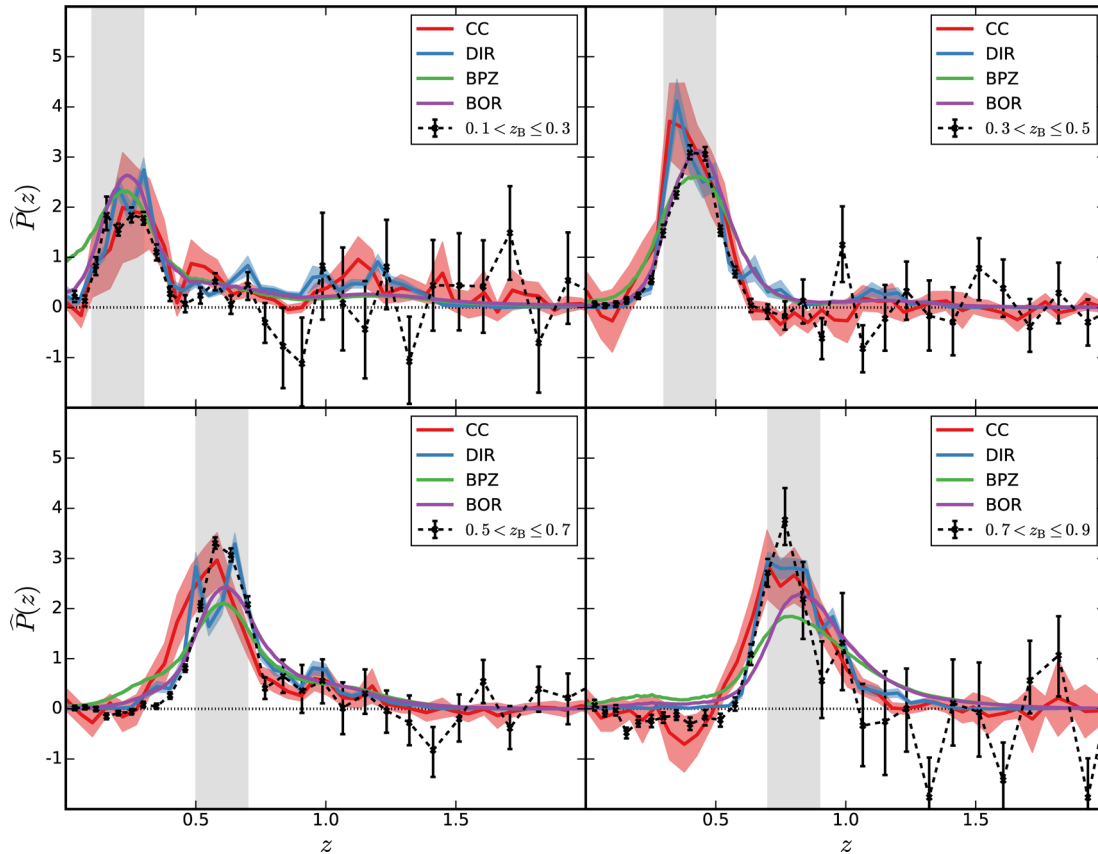


Figure 7. Comparison of the summed CS bins with the redshift distributions of Hildebrandt et al. (2017). CC (red) refers to the cross-correlation technique used in that work with spectroscopic galaxies coming from the zCOSMOS (Lilly et al. 2009) and DEEP2 (Newman et al. 2013) redshift surveys, BPZ (green) refers to the summed posteriors from BPZ, DIR (blue) refers to the direct recalibration scheme described in that work and BOR (purple) refers to the posteriors re-weighted as described in Bordoloi, Lilly & Amara (2010). Overall, the peak position and amplitudes agree with the fiducial method of the KiDS CS sample (DIR).

Another way to cause these negative correlations is if the normalizations of the area and randoms are slightly off. The density of galaxies in KiDS changes from pointing to pointing largely due to variations in average seeing. We try to account for these density variations by using the *STOMP* regions; however, these regions are not perfectly matched to the pointings and as such could cause the computation of the average density over the survey to be incorrect, leading to negative correlations. This could be fixed in the future by accounting for such systematics in a way similar to that of Morrison & Hildebrandt (2015) or Leistedt et al. (2016). These methods produce weight maps that can be used to weight the unknown objects similar to how we use shape weights in this analysis. Such weight maps should be used for high-precision analyses to account for selection effects.

Small discrepancies between these results and Hildebrandt et al. (2017) could also come from the galaxy bias not being completely removed from the samples. The lowest CS redshift bin shows a significant second peak in the redshift distribution around $z = 0.5$. Schmidt et al. (2013) show specifically how such a multi-peak distribution causes ambiguity in the peaks' relative heights that may not be fully corrected by the subsample bins as many of them also exhibit this second peak and tail. The other bins have similar problems but not to the extent of the lowest redshift bin. In the future, it may be necessary to apply a further bias calibration to the clustering redshifts such as the Newman iteration (Newman 2008). Another option is to apply a self-calibration technique to the clustering red-

shifts. This could take the form of constraining the relative bias by applying a corrective function to the summed sub-bins and the raw bins. The true galaxy bias should be a function that corrects both the sum of the sub-bins and the raw bins to the same value assuming that the bias does not change rapidly between sub-bins. We can then constrain the bias by fitting a function that brings the sum of the sub-bins and the raw bins into agreement. This could be considered a second-order correction on the bias after using the summing technique.

5.2 kdTree, single galaxy redshifts

The method of mapping colour–magnitude relations using the hybrid machine learning–clustering– z method shows definite promise. Without adding redshift information explicitly to the kdTree, we were able to use only colour–magnitude information to confirm the photo- z s. The peak of the single galaxy clustering- z s follows that of the photo- z extremely well and with a width of the clustering- z peak being no worse than $\Delta z = 0.1$ for the majority of the objects. For context, Hildebrandt et al. (2017) show that their calibrated photo- z s for similar samples have an error of $\Delta z = 0.05(1 + z)$. The single galaxy clustering- z s also identify some objects that have significantly narrower distributions. These single galaxy clustering- z s created with *THE-WIZZ* can not only be used as a stand-alone redshift estimate, but also to identify objects in colour–magnitude space where their photo- z s are expected to be more reliable.

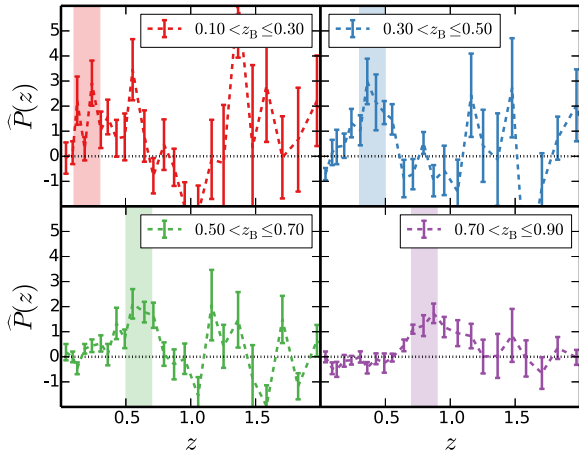


Figure 8. Raw clustering- z_s produced by THE-WIZZ using objects from KiDS selected in z_B as the unknown sample and zCOSMOS spectra as the reference sample normalized into an estimated PDF. Data points coloured and dashed with error bars are selections in z_B to mimic the bins of Hildebrandt et al. (2017) (CS bins). The coloured bands show the redshift selection. These clustering- z_s show that we can still use clustering redshift estimation even for small footprint surveys. The differences between these clustering- z_s and the previously shown results come from the smaller sample of reference redshifts with $z > 1.0$ and galaxy bias in the reference sample. It should be said that these results cannot be directly compared to the clustering- z_s shown in Hildebrandt et al. (2017) as different scales were used and the data were corrected for galaxy bias using the Newman iterative method (Newman 2008) in that study.

The ability to measure clustering- z_s for individual galaxies is a powerful tool for astronomers interested in properties of individual or small samples of galaxies rather than statistical cosmology. We have shown how one can use the measured photometric properties of galaxies to select a sample of self-similar objects to estimate a single object clustering- z . Clustering- z_s of this kind can be useful for galaxies where it is difficult to estimate photo- z_s such as those without a representative training set (this applies to all photo- z methods) or those for which galaxy spectral templates are not well understood or unavailable. This will be very common in future deep surveys such as LSST and WFIRST where faint galaxies will still likely be unable to have their redshift confirmed spectroscopically even by future 30 m class telescopes. These kinds of single galaxy clustering- z_s can be used as training information for photo- z_s for very faint objects with no redshift information in place of spectroscopic redshift. Samples with no optical counterpart, such as submillimetre galaxies (SMGs), also fall into this category and can be identified in redshift without the use of representative spectra. THE-WIZZ can use any catalogue parameter that correlates with galaxy type or redshift such as morphology, concentration, etc. not just flux and colour making it a very general tool.

5.3 Discussion on area for clustering- z_s

Clustering redshifts are largely considered a tool for current and future large-area surveys. However, what is important for signal-to-noise in clustering- z_s is not area but the product of number of reference objects and density of unknown objects. As such, they can be used on very small area surveys with the caveat that one should be mindful of using very small scales and possible sample variance due to the small survey footprint. In Fig. 8, we show cluster- z_s for the CS bins using only 0.8 deg^2 from the intersection of KiDS and the Cosmic Evolution Survey (COSMOS; Scoville et al. 2007)

using a non-public zCOSMOS (Lilly et al. 2009) catalogue kindly provided to the KiDS team for photo- z verification. The clustering- z_s still show significant signal especially the higher CS bins despite the small area. For the $z_B > 0.3$ bins, we clearly detect the redshift peaks and similar tails to that of the previous results. We use the same $R = 0.1\text{--}1.0 \text{ Mpc}$ radius as in the previous results. The effect of the bias of the spectroscopic sample can be clearly seen in the larger amplitude at $z > 1.0$ compared to the results from SDSS and GAMA.

As an extreme case, we point the reader to Schrabback et al. (2016) where THE-WIZZ was used to create clustering- z_s from 3D-Hubble Space Telescope survey data. These data covered only 0.16 deg^2 but still gave robust results thanks to the density of grism spectra and objects. This feature of clustering- z_s will be very useful in mapping colour-redshift relations out to high redshift using dense spectral and photometric fields with many filters such as COSMOS. Efforts to map colour-redshift space such as Masters et al. (2015) that used self-organizing maps to map a *Euclid* survey like colour-redshift space can benefit from clustering redshifts, mapping out redshift degeneracies in photo- z methods where more spectral or filter coverage will be required. In addition to this, clustering- z_s can be used in such high-redshift pencil beam surveys to estimate the redshift distributions of non-optically detected objects such as SMGs.

6 CONCLUSIONS

In this work, we have presented THE-WIZZ, an open source clustering redshift estimation code designed to add legacy value to current and future photometric and spectroscopic surveys. The software attempts to make using clustering redshifts as easy as photometric redshifts are by separating out the step of computing the two-point, cross-correlation statistics required for computing a clustering redshift for a given sample from creating a final clustering- z . THE-WIZZ is designed for ease of use by end users of current and future surveys and produces clustering redshifts for any subsample of objects without the intervention of a clustering redshift ‘expert’.

We have shown robust results from both pre-selecting objects from a catalogue (in this case photo- z) and from a hybrid machine learning-clustering redshift method using kdTrees in colour-magnitude space. The results from the photo- z selection reinforce other work that showed how pre-selecting objects in narrow redshift regions helps mitigate the effect of galaxy bias in clustering redshifts (Ménard et al. 2013; Schmidt et al. 2013; Rahman et al. 2016b; Scottez et al. 2016). The kdTree clustering redshift method also shows robust results for estimating the redshift of individual galaxies. Such clustering redshifts are very interesting for survey users studying individual or small samples of objects and could possibly be used as priors for future photometric redshift codes. Assuming that one can measure narrow-peaked redshift distributions for a sample of individual objects, one could use this sample as a training set for photo- z_s . This will be especially useful for high-redshift, faint objects that will likely not have observable spectra even on future 30 m class telescopes.

THE-WIZZ will be an extremely useful clustering redshift code for future photometric surveys such as LSST, *Euclid* and WFIRST given its speed and flexibility. These surveys are planning to rely at least in part on clustering redshifts to reach the precision required of their redshift distributions (Newman et al. 2015) and a public code such as THE-WIZZ can fit perfectly into these surveys’ collaborative software development environments. Future spectroscopic efforts

such as the Dark Energy Spectroscopic Instrument¹², Prime Focus Spectrograph¹³ and the 4 m Multi-Object Spectroscopic Telescope¹⁴ will soon provide the reference catalogues needed for this goal.

The code will likely require an amount of optimization in both the `PAIR_MAKER` and `PDF_MAKER` modules to minimize the size of the data file and reduce the run time per core further. However, given the relatively simple nature of the algorithm, we have no doubt such optimizations will be found. We will continue to develop `THE-WIZZ` over the years before these surveys begin taking data to maximize its impact. Future development may also include a web-based tool to allow users to request the distributions for a given sample or incorporating `THE-WIZZ` into SQL catalogue requests. These tools will be extremely useful and will speed up the adoption of clustering redshifts within the community.

`THE-WIZZ` will continue to be developed on GitHub.¹⁵ If for some reason GitHub closes or `THE-WIZZ` is moved to a new repository, contact the authors¹⁶ for the location of the current repository.

ACKNOWLEDGEMENTS

CBM would like to thank the members of the KiDS weak lensing group who helped debug the code and those that hosted the KiDS WL Busy Week at the Lorentz Centre in Leiden where much of this software was written: Catherine Heymans, Henk Hoekstra, Konrad Kuijken and Massimo Viola. We thank the International Space Science Institute (ISSI) for hosting CBM and HH during which time some of the software development took place. We also thank Mubdi Rahman for providing us with his `STOMP` map of SDSS. We thank Alexander Karim, Benjamin Magnelli and Jeffrey Newman for their helpful discussions. We also thank the zCOSMOS team for making their full, non-public redshift catalogue available to us. We thank the anonymous referee for their comments.

CBM and HH are supported by the DFG Emmy Noether grant Hi 1495/2-1. SJS is funded by grants from the Association of Universities for Research in Astronomy (AURA) and the Heising-Simons Foundation. MB is supported by the Netherlands Organization for Scientific Research, NWO, through grant number 614.001.451, and by the European Research Council through FP7 grant number 279396. AC acknowledges support from the European Research Council under the FP7 grant number 240185. PS is supported by the Deutsche Forschungsgemeinschaft in the framework of the TR33 ‘The Dark Universe’.

This material is based upon work supported in part by the National Science Foundation through Cooperative Agreement 1258333 managed by the Association of Universities for Research in Astronomy (AURA), and the Department of Energy under Contract No. DE-AC02-76SF00515 with the SLAC National Accelerator Laboratory. Additional LSST funding comes from private donations, grants to universities and in-kind support from LSSTC Institutional Members.

Based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017 and 177.A-3018, and on data products produced by Target/OmegaCEN, INAF-OACN, INAF-OAPD and the KiDS production team, on behalf of the KiDS consortium.

GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including *GALEX* MIS, VST KiDS, VISTA VIKING, *WISE*, *Herschel*-ATLAS, GMRT and ASKAP providing UV-to-radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO and the participating institutions.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the US Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society and the Higher Education Funding Council for England. The SDSS website is <http://www.sdss.org/>.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory and the University of Washington.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation and the US Department of Energy Office of Science. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington and Yale University.

Author Contributions. The authorship list is given in three groups: the lead authors (CBM), followed by two alphabetical groups. The first alphabetical group includes those who are key contributors to both the scientific analysis and the data products. The second group covers those who have either made a significant contribution to the data products or to the scientific analysis.

REFERENCES

- Alam S. et al., 2015, *ApJS*, 219, 12
 Astropy Collaboration et al., 2013, *A&A*, 558, A33
 Baldry I. K. et al., 2014, *MNRAS*, 441, 2440

¹² <http://desi.lbl.gov/>

¹³ <http://sumire.ipmujp/en/>

¹⁴ <http://www.4most.eu/>

¹⁵ <http://github.com/morriscb/The-wizz/>

¹⁶ E-mail: morrison.chrisb@gmail.com

- Begeman K., Belikov A. N., Boxhoorn D. R., Valentijn E. A., 2013, *Exp. Astron.*, 35, 1
- Benítez N., 2000, *ApJ*, 536, 571
- Bertin E., Arnouts S., 1996, *A&AS*, 117, 393
- Bolton A. S. et al., 2012, *AJ*, 144, 144
- Bordoloi R., Lilly S. J., Amara A., 2010, *MNRAS*, 406, 881
- Choi A. et al., 2016, *MNRAS*, 463, 3737
- Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, *MNRAS*, 396, 2379
- Davis M., Peebles P. J. E., 1983, *ApJ*, 267, 465
- Dawson K. S. et al., 2013, *AJ*, 145, 10
- de Jong J. T. A. et al., 2015, *A&A*, 582, A62
- Driver S. P. et al., 2009, *Astron. Geophys.*, 50, 5.12
- Eisenstein D. J. et al., 2011, *AJ*, 142, 72
- Erben T. et al., 2005, *Astron. Nachr.*, 326, 432
- Fenech Conti I., Herbonnet R., Hoekstra H., Merten J., Miller L., Viola M., 2016, *MNRAS*, preprint ([arXiv:1606.05337](https://arxiv.org/abs/1606.05337))
- Hildebrandt H. et al., 2010, *A&A*, 523, A31
- Hildebrandt H. et al., 2012, *MNRAS*, 421, 2355
- Hildebrandt H. et al., 2017, *MNRAS*, 465, 1454
- Johnson A. et al., 2017, *MNRAS*, 465, 4118
- Komatsu E. et al., 2009, *ApJS*, 180, 330
- Kuijken K., 2008, *A&A*, 482, 1053
- Kuijken K. et al., 2015, *MNRAS*, 454, 3500
- Leistedt B. et al., 2016, *ApJS*, 226, 24
- Lilly S. J. et al., 2009, *ApJS*, 184, 218
- Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, *MNRAS*, 390, 118
- Liske J. et al., 2015, *MNRAS*, 452, 2087
- McQuinn M., White M., 2013, *MNRAS*, 433, 2857
- Masters D. et al., 2015, *ApJ*, 813, 53
- Matthews D. J., Newman J. A., 2010, *ApJ*, 721, 456
- Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, preprint ([arXiv:1303.4722](https://arxiv.org/abs/1303.4722))
- Miller L. et al., 2013, *MNRAS*, 429, 2858
- Morrison C. B., Hildebrandt H., 2015, *MNRAS*, 454, 3121
- Newman J. A., 2008, *ApJ*, 684, 88
- Newman J. A. et al., 2013, *ApJS*, 208, 5
- Newman J. A. et al., 2015, *Astropart. Phys.*, 63, 81
- Rahman M., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., 2015, *MNRAS*, 447, 3500
- Rahman M., Ménard B., Scranton R., 2016a, *MNRAS*, 457, 3912
- Rahman M., Mendez A. J., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., Budavári T., 2016b, *MNRAS*, 460, 163
- Ross N. P. et al., 2012, *ApJS*, 199, 3
- Schirmer M., 2013, *ApJS*, 209, 21
- Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, *MNRAS*, 431, 3307
- Schmidt S. J., Ménard B., Scranton R., Morrison C. B., Rahman M., Hopkins A. M., 2015, *MNRAS*, 446, 2696
- Schneider M., Knox L., Zhan H., Connolly A., 2006, *ApJ*, 651, 14
- Schrabback T. et al., 2016, *MNRAS*, preprint ([arXiv:1611.03866](https://arxiv.org/abs/1611.03866))
- Scottez V. et al., 2016, *MNRAS*, 462, 1683
- Scoville N. et al., 2007, *ApJS*, 172, 1
- Scranton R. et al., 2002, *ApJ*, 579, 48
- Scranton R. et al., 2005, *ApJ*, 633, 589
- Strauss M. A. et al., 2002, *AJ*, 124, 1810
- Valentijn E. A. et al., 2007, in Shaw R. A., Hill F., Bell D. J., eds, *ASP Conf. Ser. Vol. 376, Astronomical Data Analysis Software and Systems XVI*. Astron. Soc. Pac., San Francisco, p. 491
- York D. G. et al., 2000, *AJ*, 120, 1579

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.