A&A 642, A200 (2020)
https://doi.org/10.1051/0004-6361/202038835

Astronomy & Astrophysics

Testing KiDS cross-correlation redshifts with simulations

J. L. van den Busch[1,2], H. Hildebrandt[1], A. H. Wright[1], C. B. Morrison[3], C. Blake[4], B. Joachimi[5], T. Erben[2], C. Heymans[6,1], K. Kuijken[7], and E. N. Taylor[7]

[1] Ruhr-University Bochum, Astronomical Institute, German Centre for Cosmological Lensing, Universitätsstr. 150, 44801 Bochum, Germany
e-mail: jlvdb@astro.rub.de
[2] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[3] Department of Astronomy, University of Washington, Box 351580, Seattle, WA 98195, USA
[4] Centre for Astrophysics & Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia
[5] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
[6] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[7] Leiden Observatory, Leiden University, PO Box 9513, 2300RA Leiden, The Netherlands

Received 3 July 2020 / Accepted 17 August 2020

ABSTRACT

Measuring cosmic shear in wide-field imaging surveys requires accurate knowledge of the redshift distribution of all sources. The clustering-redshift technique exploits the angular cross-correlation of a target galaxy sample with unknown redshifts and a reference sample with known redshifts. It represents an attractive alternative to colour-based methods of redshift calibration. Here we test the performance of such clustering redshift measurements using mock catalogues that resemble the Kilo-Degree Survey (KiDS). These mocks are created from the MICE simulation and closely mimic the properties of the KiDS source sample and the overlapping spectroscopic reference samples. We quantify the performance of the clustering redshifts by comparing the cross-correlation results with the true redshift distributions in each of the five KiDS photometric redshift bins. Such a comparison to an informative model is necessary due to the incompleteness of the reference samples at high redshifts. Clustering mean redshifts are unbiased at $|\Delta z| < 0.006$ under these conditions. The redshift evolution of the galaxy bias of the reference and target samples represents one of the most important systematic errors when estimating clustering redshifts. It can be reliably mitigated at this level of precision using auto-correlation measurements and self-consistency relations, and will not become a dominant source of systematic error until the arrival of Stage-IV cosmic shear surveys. Using redshift distributions from a direct colour-based estimate instead of the true redshift distributions as a model for comparison with the clustering redshifts increases the biases in the mean to up to $|\Delta z| \sim 0.04$. This indicates that the interpretation of clustering redshifts in real-world applications will require more sophisticated (parameterised) models of the redshift distribution in the future. If such better models are available, the clustering-redshift technique promises to be a highly complementary alternative to other methods of redshift calibration.

Key words. cosmology: observations – surveys – large-scale structure of Universe – galaxies: distances and redshifts

1. Introduction

Weak gravitational lensing (WL) experiments have been established as one of the most sensitive cosmological probes (e.g. Bartelmann & Schneider 2001; Troxel et al. 2018; Hikage et al. 2019; Hildebrandt et al. 2020a). The aim of these experiments is to statistically probe the distribution and evolution of large-scale matter structures by studying the effect of their gravitational field on the propagation of light. The main observable effect – coherent distortions in the images of background galaxies – is very weak and can only be observed statistically from large samples of galaxies with well-measured shapes. In this regime, tight control of systematic errors is essential throughout the analysis (Mandelbaum 2018).

These efforts include a precise, unbiased calibration of the source redshift distribution for very large galaxy samples (e.g. Hoyle et al. 2018; Tanaka et al. 2018; Wright et al. 2020). Complete spectroscopy is unfeasible for such surveys consisting of tens of millions of faint galaxies so that secondary estimates for the redshift distributions are required, for example redshifts from multicolour photometry known as photometric redshifts (photo-z; see Salvato et al. 2019 for a review). Weak lensing by the large-scale-structure of the Universe (a.k.a. cosmic shear) is a statistical effect, integrated along the line-of-sight so that redshift precision for individual galaxies is not critically important. Typically, the sources used for the measurement are divided into several so-called tomographic redshift bins. It is the redshift distributions of the sources in these bins that are required to model the observed signals. These distributions, and most importantly their mean redshifts, need to be estimated with very high accuracy, which is challenging with photometric redshifts alone due to degeneracies in colour-redshift space and incompleteness of spectroscopic template libraries.

In order to meet the stringent requirements on the accuracy of the mean redshifts (as opposed to accuracy of the individual redshifts per se), modern cosmic shear surveys have developed different methods of redshift calibration. If a survey contains a sub-sample of galaxies with spectroscopic redshift (spec-z) measurements, this sub-sample can be used to estimate the redshift distribution of the full survey. However, it is in principle required that this sub-sample is representative of the full WL source sample. If that is not the case, some re-weighting of the spec-z

sample can – under certain conditions – still yield an unbiased estimate of the true redshift distribution (Lima et al. 2008). This re-weighting method, implemented via k-nearest-neighbour matching (kNN) or self-organised-maps (SOM), has been used widely in the WL literature (Bonnett et al. 2016; Masters et al. 2016; Hildebrandt et al. 2017, 2020a; Wright et al. 2020; Buchs et al. 2019).

A complementary approach to estimate redshift distributions with the help of a spec-z calibration sample uses galaxy angular cross-correlation measurements (Newman 2008; Matthews & Newman 2010; Schmidt et al. 2013; Ménard et al. 2013; McQuinn & White 2013). Galaxy samples that overlap in redshift show some correlation of their angular positions on the sky. Hence, a measurement of this angular cross-correlation function can yield an estimate of the redshift distribution of an unknown sample if the other sample has accurately known redshifts. The great appeal of this method is that it does not require a reference spec-z sample that is representative of the unknown target sample so long as both overlap in redshift (Newman 2008). For example, a bright reference sample can be used to estimate the redshift distribution of a faint target sample via cross-correlations as bright galaxies cluster with faint galaxies; they are both tracers of the same underlying matter field. This makes such a cross-correlation approach especially attractive for faint target samples that are hard to study spectroscopically in a representative way. Here we will concentrate on this approach, a method that is also dubbed "clustering redshifts".

Any measurement of galaxy clustering – and hence also clustering redshift measurements – is affected by the unknown bias of the galaxies with respect to the underlying matter density field, which depends on galaxy type and evolves with redshift. Clustering redshifts need to be corrected for this bias before they can be used to estimate redshift distributions for cosmological inference. While the absolute value of the galaxy bias is not important for the clustering-redshift method, any redshift evolution of the galaxy bias within either the reference or the target samples will introduce a skew in the estimates of the redshift distributions. Thus, this bias evolution needs to be estimated and corrected for. While this is straightforwardly done for the reference sample, by estimating its angular auto-correlation function and exploiting its precise redshift information, such a correction is not possible for the target sample. Instead, self-consistency relations can be formulated that estimate the galaxy bias evolution of the target sample by dividing this sample into redshift slices of different width.

The aim of this work is to assess the performance of the clustering-redshift methodology and the bias corrections described above. We use the optical and infra-red weak lensing surveys KiDS (Kilo-Degree Survey, Kuijken et al. 2015) and VIKING (VISTA Kilo Degree, Edge et al. 2013) as reference to create a simulated, realistic target galaxy sample, as well as a set of simulated spectroscopic calibration samples which are used as a reference in the clustering redshift measurements. These mock catalogues, based on the MICE simulation (Fosalba et al. 2015a,b; Crocce et al. 2015; Carretero et al. 2015; Hoffmann et al. 2015), include effects such as gravitational lensing, evolving galaxy bias and spectroscopic selection effects which allows us to quantify the impact of these systematics on the clustering redshift measurements. The cross-correlation measurements are analysed with an updated methodology similar to the one presented in Hildebrandt et al. (2020a). We quantify the residual biases in such a clustering redshift experiment by comparing to the known redshift distributions of the simulated mock catalogues. This yields realistic best-practice solutions that can be used with contemporary and future cosmic shear surveys.

The paper is organised as follows. In Sect. 2, we present the simulated datasets that form the basis of this work, in particular the detailed creation of the catalogues. We note that these mock catalogues have already been used to validate the DIR and SOM redshift calibration methods in Wright et al. (2020) and Joudaki et al. (2020). The theory behind and implementation of the clustering-redshift technique for this work and differences to previous KiDS clustering redshift measurements are covered in Sect. 3. Results are presented in Sect. 4 and discussed in Sect. 5 before the paper is summarised in Sect. 6.

2. Simulated data

To assess the performance of the KiDS clustering-redshift methodology, we must construct a simulated dataset that is similarly complex as the observational data, particularly with respect to photometric properties and selection effects. In this section we describe the construction of such realistic mock galaxy samples that closely resemble the KiDS+VIKING-450 (KV450, Wright et al. 2019) cosmic shear sample (Sect. 2.1) and spectroscopic reference samples (Sect. 2.2), which we subsequently use to verify our clustering-z methodology. We start from an existing galaxy mock catalogue and add properties like a photometry realisation, galaxy weights and photometric redshifts (photo-z) in a post-processing pipeline. This pipeline represents a blueprint for the construction of future KiDS mock samples and is publicly available[1].

The basis for our mock creation is the MICE simulation (Fosalba et al. 2015a). MICE is a dark matter-only simulation generated in a box of width $L = 3.1\,\mathrm{h}^{-1}\,\mathrm{Gpc}$, thereby allowing construction of a light cone that covers an octant of the sky. The simulation assumes a flat Λ-CDM cosmological model, with $\Omega_m = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_\mathrm{b} = 0.044$, $\sigma_8 = 0.8$, and $h = 0.7$. The simulation traces the evolution of $\sim 6.9 \times 10^{10}$ particles with mass $2.9 \times 10^{10}\,\mathrm{h}^{-1} M_\odot$, from an initial redshift of $z_\mathrm{init} = 100$ to the present day. This high particle density allows the simulation to match (to within a few percent, Fosalba et al. 2015a) theoretical predictions for matter clustering even on small scales ($k \sim 1\,h\,\mathrm{Mpc}^{-1}$); scales which are of particular interest for clustering redshift measurements.

An additional strength of the MICE simulation is the availability of a synthetic galaxy and halo catalogue[2] for the full light-cone. This catalogue was generated by identifying halos using a Friends-of-Friends algorithm (Crocce et al. 2015), and subsequently populating these halos with galaxies using a mixture of halo occupation distribution (HOD) and halo abundance matching (HAM) techniques (Carretero et al. 2015) up to a redshift of $z \approx 1.4$. This redshift limit implies that we are not be able to model the tails of the redshift distribution of the KV450 cosmic shear sample, which extends beyond $z = 1.4$. However, the incompleteness of our spectroscopic reference samples (see Sect. 2.2) would make a clustering redshift calibration of these tails difficult. For all analyses in this work, we use the second version of this galaxy catalogue, which we simply refer to as MICE2. This catalogue provides galaxy positions, shapes, stellar masses, and simulated photometry for many photometric bandpasses, such as those utilised by *Euclid* (Laureijs et al. 2011), Sloan Digital Sky Survey (SDSS, York et al. 2000), the Dark

[1] https://github.com/KiDS-WL/MICE2_mocks
[2] Distributed on https://cosmohub.pic.es/ (Carretero et al. 2017).

Energy Survey (DES, Flaugher et al. 2015), and the VIKING survey (Edge et al. 2013). As a result, the MICE2 galaxy catalogue comes pre-packaged with simulated photometry in filters similar (or identical) to those used in KV450 ($ugriZYJHK_s$).

As the MICE2 catalogue was constructed with gravitational lensing applications in mind, it includes both shear (split in two components, γ_1 and γ_2) and convergence (κ) information at the position of each galaxy (Fosalba et al. 2015b). Besides the shape, gravitational lensing also affects the position and magnifies the flux of galaxies through changes in the observed solid angle. Therefore, both true and lensed galaxy positions are listed in MICE2. Photometry contained within the catalogue does not, however, include a consideration for the effects of magnification. We utilise this fact to investigate the impact of magnification on our redshift calibration methodology: we perform our analysis using catalogues with lensed positions and magnified fluxes, and then again using a catalogue containing true positions and unmagnified fluxes (see Sect. 4.2). Our method for implementing flux magnification is provided below.

The methods applied to the underlying dark matter-only simulation to obtain the MICE2 galaxy catalogue and our sample selection strategies (see Sect. 2.1) are similar to the efforts to design the Buzzard Flock synthetic sky catalogues (DeRose et al. 2019) for DES. Some key differences are that MICE has a higher particle density and therefore mass resolution, whereas the Buzzard mocks implement gravitational lensing with full ray-tracing opposed to MICE2 which computes lensing observables using the Born approximation.

2.1. KV450 mock catalogue

The combined KV450 dataset, including object detection, forced optical and infrared photometry, and photometric redshifts (photo-z), is described in detail in Wright et al. (2019). The primary strength of this dataset lies with the addition of the infrared $ZYJHK_s$-bands, from the VIKING survey (Edge et al. 2013; Venemans et al. 2015), to the KiDS optical $ugri$ dataset (de Jong et al. 2017), thereby significantly improving the performance of the photo-z (particularly at $z \gtrsim 0.9$) obtained via template-fitting with BPZ (Bayesian Photometric Redshift, Benítez 2000).

The effective survey area of the KV450 dataset is 341.3 deg^2 (Wright et al. 2019). This sample is limited to sources with successful photometric estimates, made using the Gaussian Aperture and PSF (GAAP, Kuijken 2008) photometric pipeline, in all nine bands. GAAP is a technique that allows to accurately measure colours by accounting for differences in the point spread function (PSF) in each filter. This is achieved by measuring fluxes with a filter-dependent, spatially varying kernel that Gaussianises the PSF. All sources are assigned lensing weights obtained from *lensfit* (Miller et al. 2007, 2013; Fenech Conti et al. 2017; Kannawadi et al. 2019). This weighting effectively selects extended sources with r-band apparent magnitudes in the interval $20 \lesssim r \lesssim 25$ (Wright et al. 2019), resulting in an effective surface density (Heymans et al. 2012) of $n_{\rm eff} = 7.38\,\mathrm{arcmin}^{-2}$. Furthermore, in Hildebrandt et al. (2020a) we selected sources for cosmic shear tomography as being those within the photo-z window $0.1 < Z_B \leq 1.2$, where Z_B is the photo-z point-estimate returned by BPZ. We split these sources into five non-overlapping tomographic bins with boundaries $Z_B \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.2\}$.

The mock sample that we create based on MICE2 mimics the same selection function, object weights, and photometric redshifts, thereby allowing us to apply the same tomographic photo-z selection on the mocks as in the real KV450 dataset. This requires us to generate photometric realisations, which match the KiDS photometric noise properties, based on the MICE2 model magnitudes. We can do this using, per filter, the observed median photometric depth and PSF (see Table 1).

Table 1. Median limiting magnitudes and PSF FWHMs for each filter in the KV450 dataset.

Filter	PSF FWHM (arcsec)	GAAP magnitude limit (1σ, AB)
u	1.0	25.5
g	0.9	26.3
r	0.7	26.2
i	0.8	24.9
Z	1.0	24.9
Y	1.0	24.1
J	0.9	24.2
H	1.0	23.3
K_s	0.9	23.2

We require only a relatively small fraction of the full MICE2 octant to match the area of the KV450 footprint. Due to the way in which the lightcone was constructed from the simulation box, the completeness of the MICE2 galaxy catalogue varies with position (see Fosalba et al. 2015a). Therefore we select a rectangular region between $35° < \mathrm{RA} < 55°$ and $6° < \mathrm{Dec} < 24°$, which has a reportedly high completeness down to $i_{\rm DES} = 24.0$ mag. Using this subset as a basis, we construct our mock KV450 photometric sample by applying the following series of steps: application of evolution corrections[3], application of flux magnification, construction of photometric apertures, realisations of photometric noise, assignment of shear-measurement weights, and computation of photometric redshifts. The MICE2 completeness limit, quoted above, is nominally brighter than the corresponding KV450 i-band magnitude limit (see Table 1). However, the shear-measurement weights preferentially select bright objects such that the completeness limit is not an issue for our sample selection.

We start our mock sample construction by first selecting, from MICE2, the raw simulated photometry (which is noiseless, uncorrected for evolution and magnification, and expressed in AB apparent magnitudes per filter X: $m_X^{\rm raw}$) which uses photometric band-passes X that are most similar to those used in KV450. For the OmegaCAM $ugri$-bands and VISTA Z-band we use the provided SDSS $u'g'r'i'z'$-band fluxes. For the VISTA Y-band we use the DES y-band. The VISTA JHK_s-bands are provided natively within the catalogue.

Following the recommendation of Fosalba et al. (2015b), we apply a redshift dependent evolution correction to all fluxes within the MICE2 catalogue:

$$m_X^{\rm evo}(z_{\rm true}) = m_X^{\rm raw} - 0.8\,[\arctan(1.5\,z_{\rm true}) - 0.1489]\,, \quad (1)$$

where $m_X^{\rm evo}$ is the evolution-corrected apparent magnitude in filter X, $m_X^{\rm raw}$ is the raw simulated apparent magnitude in filter X, and $z_{\rm true}$ is the true redshift of the source. We then approximate the effect of magnification on our fluxes, again following the recommendation of Fosalba et al. (2015b, see Eq. (21)), with the correction:

$$m_X^{\rm mag} = m_X^{\rm evo} - 2.5\log_{10}(1 + \delta\mu), \quad (2)$$

[3] The evolutionary correction ensures a better match in the galaxy number density between MICE2 and observational data.

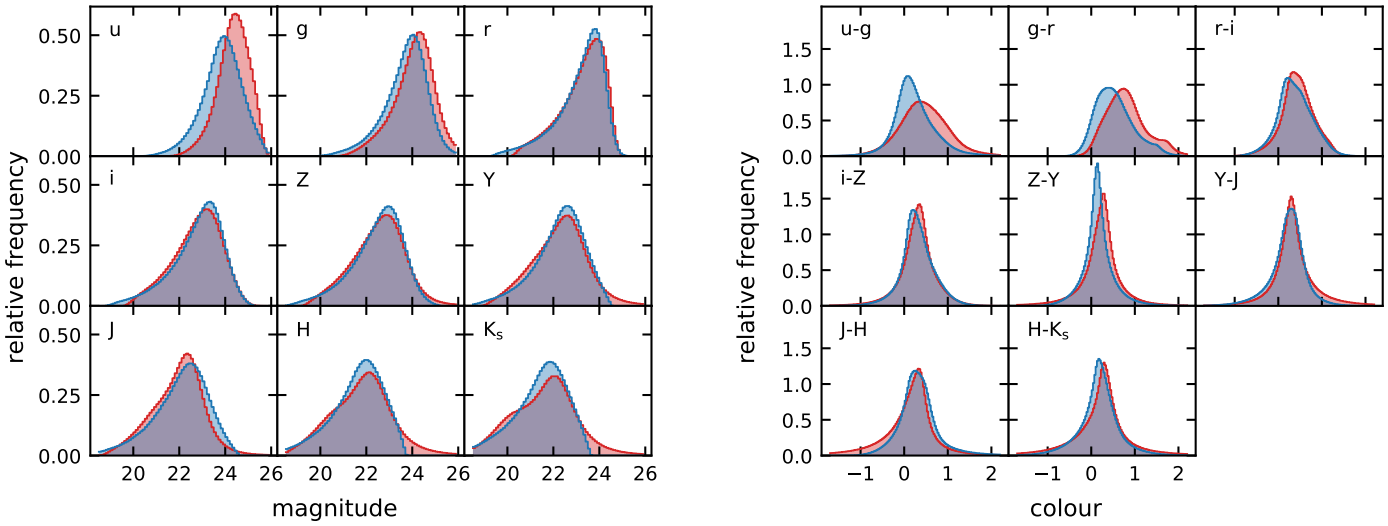

Fig. 1. Comparison of the colours and magnitudes for the KV450 (red) and MICE2 mock (blue) cosmic shear samples. All histograms are unweighted, but objects with zero *lensfit* weight are excluded from both samples. The simulated distributions match the data well, except that MICE2 fluxes tend to be bluer overall.

where $\delta\mu \approx 2\kappa$ in the weak lensing limit. We then define the "true" magnitude of each source in band X, $m_X^{\rm true}$, as either $m_X^{\rm mag}$ (for the magnified case) or $m_X^{\rm evo}$ (for the unmagnified case).

We now derive "observed" photometric quantities (magnitudes and uncertainties) per-filter, by adding representative photometric noise to the values of $m_X^{\rm true}$. This requires us to match the photometric signal-to-noise ratio (S/N) of the simulations, again per-filter, to that which is observed in KiDS and VIKING imaging. In order to simulate realistic photometric errors, we need to estimate an effective aperture size for each galaxy. We first derive the half-light radius $R_{{\rm E},i}$ from the tabulated disk and bulge radii as well as the bulge-to-total flux ratio of the two component Sérsic (1963) profiles of the MICE2 galaxies. Based on this projected galaxy size we compute a corresponding photometric aperture the KiDS pipeline would apply: The per-filter aperture major ($a_{X,i}^{\rm ap}$) and minor axes ($b_{X,i}^{\rm ap}$) are

$$a_{X,i}^{\rm ap} = \sqrt{\sigma_{{\rm PSF},X}^2 + (2.5R_{{\rm E},i})^2} \quad \text{and} \tag{3}$$

$$b_{X,i}^{\rm ap} = \sqrt{\sigma_{{\rm PSF},X}^2 + \left(2.5\frac{b_i^{\rm int}}{a_i^{\rm int}}R_{{\rm E},i}\right)^2}, \tag{4}$$

where $\sigma_{{\rm PSF},X}$ is the filter X PSF standard deviation (derived from the FWHM, see Table 1), and the intrinsic major-to-minor axis ratio $a_i^{\rm int}/b_i^{\rm int}$ of source i. Using the aperture area $A_{X,i}^{\rm ap} = \pi a_{X,i}^{\rm ap} b_{X,i}^{\rm ap}$, the S/N values of $S/N_{X,i}$ can now be computed as:

$$S/N_{X,i} = 10^{-0.4\left(m_{X,i}^{\rm true} - m_X^{\rm lim}\right)}\sqrt{\frac{\pi\sigma_{{\rm PSF},X}^2}{A_{X,i}^{\rm ap}}}\,k, \tag{5}$$

where $m_X^{\rm lim}$ is the observed GAAP magnitude limit in filter X (see Table 1) and k is a free scaling parameter. We compare the distributions of S/N_X in the simulation with the observed signal-to-noise distributions, as determined from the correspondingly measured GAAP magnitude. By varying the scaling parameter, we find that a good match between the simulated and observed distributions is achieved at $k \approx 1.5$ in all filters. We then compute Gaussian uncertainties in magnitude and generate observed magnitudes:

$$m_{X,i}^{\rm obs} = m_{X,i}^{\rm true} + x \quad \text{with } x \sim \mathcal{N}\left(0, \frac{2.5}{\ln 10}\frac{1}{S/N_{X,i}}\right). \tag{6}$$

We use these magnitudes in Eq. (5), substituting $m_{X,i}^{\rm true}$ with $m_{X,i}^{\rm obs}$, to derive the observed S/N per filter and source, $S/N_{X,i}^{\rm obs}$. Finally, we perform a quasi-detection in each band, only keeping magnitude information for sources with $S/N_{X,i}^{\rm obs} > 1.0$. We note that the precise implementation of this detection limit is not critical due to the implementation of the shape measurement weights that limit our analysis to significantly higher S/N anyway.

All cosmic-shear sources in KV450 have an associated shape-measurement weight, which is assigned based on the confidence of their shape measurement. As we do not have simulated imaging for the MICE2 dataset, we cannot recreate this value from the simulations directly. Instead we exploit the fact that these weights are strongly correlated with magnitude (particularly the r-band magnitude, as this is the band which is used for galaxy shape measurement). Therefore, to assign shear measurement weights to simulated sources, we simply perform a kNN matching in 9-dimensional magnitude space between the simulated and real photometric datasets; the simulated data then inherit the shear weight of the nearest neighbour in the data. In cases where there is no nearest neighbour match within a 1.0 mag Euclidean radius, we assign a weight of zero[4]. This assignment produces well behaved shape-measurement weights, while also encoding any strong selection effects that are present in the data but overlooked in our simulation setup.

The resulting colour and magnitude distributions for the mock and real photometric data are shown in Fig. 1. These distributions match the data fairly well, except for the u and g-bands; MICE2 galaxies are typically bluer than we see in the data. The mock galaxy magnitudes do not reproduce the faint tails of the near-infrared photometry seen in KV450, which we attribute to stronger depth variations seen in the VIKING data, which are not modelled in our analytic prescription (Eq. (5)).

[4] This situation arises in about 2% of the cases, in particular for the faintest galaxies with $r \gtrsim 24$.

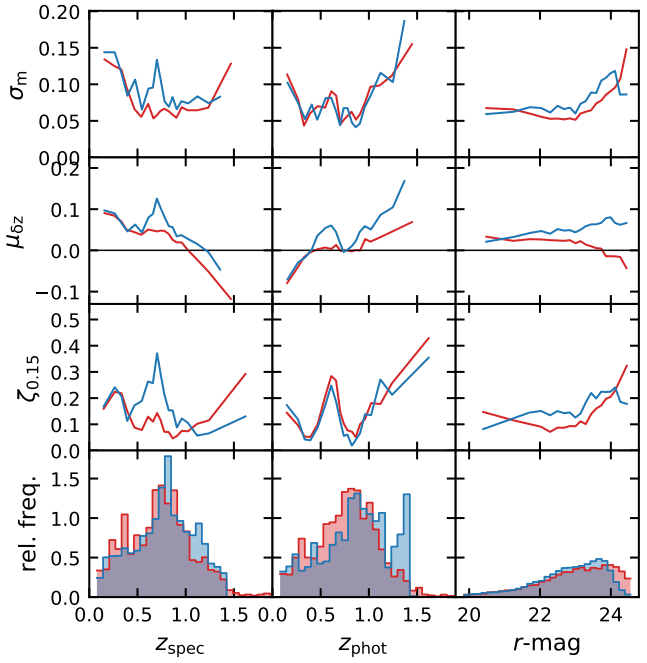

Fig. 2. Comparison of statistics of the scaled photo-z bias $\delta z = (z_B - z_{spec})/(1 + z_{spec})$ of the KV450 (red) and MICE2 mock (blue) datasets. The galaxies used for this compilation originate from DEEP2, VVDS, and zCOSMOS. The columns *from left to right* show statistics as a function of spectroscopic redshift, photometric, and r-band magnitude (shown as histograms in the *bottom row*). The rows *from top to bottom* show the median-absolute-deviation (σ_m), mean ($\mu_{\delta z}$), and the outlier fraction ($\zeta_{0.15}$, the fraction of objects with $\delta z > 0.15$).

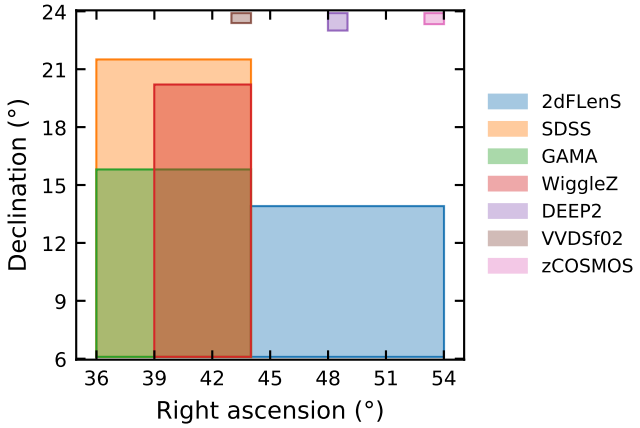

Fig. 3. Distribution of the mock spectroscopic samples within the mock KV450 footprint (indicated by the axes limit).

Finally, we perform a photo-z estimation for the simulated data using our observed magnitudes and magnitude uncertainties. To ensure consistency with the data, we implement the same photo-z code (i.e. BPZ) and setup; that is, we use the same redshift prior (Raichoor et al. 2014) and template set (Capak 2004) as described in Wright et al. (2019). A difference between the simulations and the data, however, lies in the redshift priors applied to the simulations. As we know a priori that the mock catalogue only contains galaxies within $0.07 \lesssim z \lesssim 1.41$, we limit the BPZ redshift-prior to this range. We also reject objects if their aperture photometry cannot be measured in the i-band, the reference magnitude for the redshift prior, to avoid spurious photo-z estimates.

The resulting mock photometric redshift properties are well matched to the KV450 data as shown in Fig. 2. Following Wright et al. (2019), we measure statistics of the scaled photo-z bias distribution, $\delta z = (z_B - z_{spec})/(1 + z_{spec})$, as a function of spectroscopic redshift, photometric redshift, and r-band apparent magnitude. We find agreement between the mocks and data in each of the test statistics explored: distribution scatter (σ_m; estimated using the normalised median-absolute-deviation from median, nMAD), mean bias ($\mu_{\delta z}$), catastrophic outlier fraction ($\zeta_{0.15}$; the fraction of sources with $|\delta z| > 0.15$), and relative counts. In particular, we highlight the similarities between the scatter and outlier fractions seen as a function of photo-z, which are closely matched between the mocks and the data.

2.2. Spectroscopic mock catalogues

The KV450 photometric dataset has spatial overlap with a rich set of spectroscopic surveys, which are needed for optimal redshift calibration. For a detailed discussion of all available spectroscopic data which intersect the KV450 footprint, we refer the interested reader to Hildebrandt et al. (2020a). Here we briefly summarise the spectroscopic selection functions and how to implement these on the MICE2 mocks for the subset of the datasets which we use for clustering redshift measurements. These samples can be divided in two categories. First, there are four wide-area spectroscopic surveys with a combined overlap with KV450 of 212.0 deg^2, namely the Sloan Digital Sky Survey (SDSS, Alam et al. 2015), the 2-degree Field Lensing Survey (2dFLenS, Blake et al. 2016), the Galaxy and Mass Assembly (GAMA, Driver et al. 2011) survey, and the WiggleZ Dark Energy Survey (WiggleZ, Drinkwater et al. 2010). Secondly, KiDS relies on deep spectroscopic surveys, which each have a post-masking overlap area with KiDS of less than 1 deg^2, to calibrate the high-redshift portions of the redshift distributions; the DEEP2 Galaxy Redshift Survey (DEEP2, Newman et al. 2013), the VIMOS VLT Deep Survey (VVDS, Le Fèvre et al. 2013), and the Cosmic Evolution redshift survey (zCOSMOS, Lilly et al. 2009). As with the KV450 mock galaxy sample, we aim to create mock spectroscopic samples with similar complexity as seen in the actual spectroscopic data.

We start our mock spectroscopic source generation by simplifying the complex footprints of each of our spectroscopic samples by selecting rectangular regions which respect their size and the various overlaps between each other and with the true KV450 data. These footprints can be seen in Fig. 3, demonstrating the various spatial overlaps between the different samples in our spectroscopic compilation and matching (in area) those of the corresponding data samples within a few percent.

Next, we apply the spectroscopic target selection functions (where possible) and adjustments (where necessary) to obtain the closest match possible between our mock and data spectroscopic redshift distributions. These selection functions are all magnitude- and/or colour-dependent, and are variously based on imaging that is either shallower (2dFLenS, GAMA, SDSS, and WiggleZ) or deeper (DEEP2, VVDS, and zCOSMOS) than the KV450 imaging. As a result, we are able to use the mock KV450 photometry (i.e. $m^{obs}_{X,i}$) to perform spectroscopic selections for the former samples, but revert to the noiseless photometry (i.e. $m^{mag}_{X,i}$) for the selection of the latter samples. A tabular summary of these selection functions and additional plots for VVDS and zCOSMOS can be found as supplementary material in Appendix A.

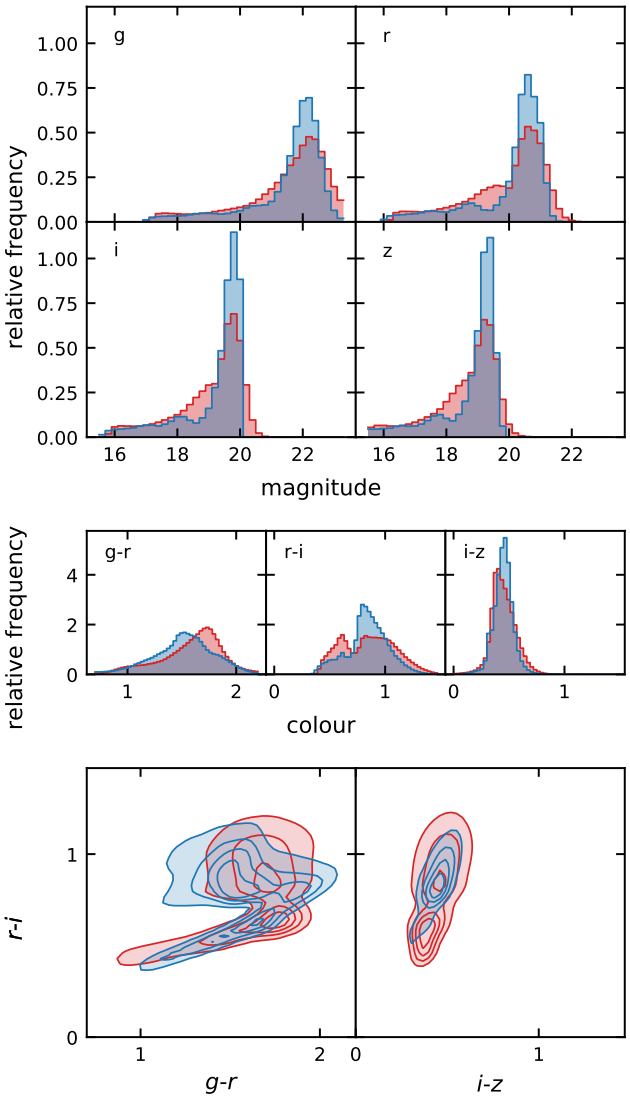

Fig. 4. Comparison of colour and magnitude distributions for the simulated (blue) and observed (red) BOSS LOWZ and CMASS datasets.

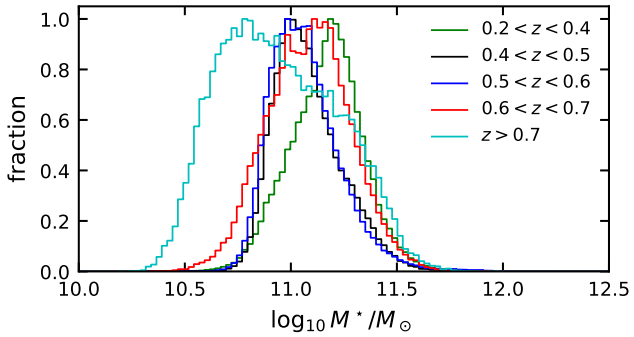

Fig. 5. MICE2 stellar mass distributions for the BOSS mock sample in bins of redshift, normalised to their peak value. The comparison to Fig. 10 in Maraston et al. (2013) reveals significantly lower stellar masses in the highest redshift bin.

2.2.1. SDSS

SDSS covers large parts of the northern KiDS fields. Our sample consists of galaxies from the SDSS Main Galaxy Sample (Strauss et al. 2002), BOSS (Baryon Oscillation Spectroscopic Survey, Dawson et al. 2013), and the QSO sample (Schneider et al. 2010) at high redshifts. The selections applied to MICE2 for each of these samples are given in Table A.1, and differences with respect to the literature are justified in the following.

First, we approximate the main galaxy sample by selecting objects with $m^{\rm obs}_{r,i} < 17.7$, which is slightly brighter than the literature limit of 17.77 (Strauss et al. 2002). We attribute this small discrepancy to the different magnitude definitions used in SDSS and our mock KV450 dataset (i.e. simple Petrosian vs our $2.5R_e$ aperture fluxes). Our updated selection gives a better match between the mock and simulated redshift distributions for the SDSS main galaxy sample.

The BOSS LOWZ and CMASS samples use a set of auxiliary colours, based on *gri*-band model magnitudes and optimised for the selection of luminous red galaxies (LRGs), which are defined as:

$$c_\parallel = 0.7\,(g-r) + 1.2\,(r-i-0.18); \quad (7)$$
$$c_\perp = (r-i) - (g-r)/4 - 0.18; \text{ and} \quad (8)$$
$$d_\perp = (r-i) - (g-r)/8. \quad (9)$$

These auxiliary colours can be directly computed for our MICE2 sources given the available bandpasses: $m^{\rm obs}_{X,i}$ for $X \in gri$. We are therefore able to apply the literature colour-cuts for the LOWZ and CMASS samples (Dawson et al. 2013), albeit again with some minor modifications. These modifications are applied to construct better matches between the simulations and data in colour and magnitude space: see Fig. 4 for a comparison of these values between our mocks and the data.

To verify the fidelity of our BOSS selection, we compare the distributions of stellar mass, in bins of redshift, between our mock sample (see Fig. 5) and those of Maraston et al. (2013, see their Fig. 10). At redshifts below $z \sim 0.6$ our BOSS sample is in good agreement with Maraston et al. (2013), except for a constant 0.2 dex systematic offset between the two samples. This systematic offset is not surprising, given the systematic differences between the Maraston et al. (2013) stellar population synthesis models and those of Bruzual & Charlot (2003), which are used to estimate MICE2 stellar masses. Sources with $z > 0.7$ in our mock sample, however, have a significantly lower stellar mass than the corresponding BOSS galaxies in Maraston et al. (2013). Whereas in Maraston et al. (2013) this redshift bin has the highest mean stellar mass, we find that high-redshift MICE2 BOSS galaxies have the lowest mean stellar mass; an offset of ~0.5 dex in mass. This in turn suggests that the biasing of these galaxies will differ in the simulations compared to the data. We expect this to have a minor impact on the results presented here. It could however be of importance in other applications of this dataset which are more sensitive to the absolute value of the galaxy bias.

There is no obvious way to implement the SDSS QSO sample within MICE2, as the simulation does not include active galactic nuclei in the galaxy catalogue construction. We therefore approximate the selection of quasars using physical properties. We assume that quasars are triggered exclusively in dense environments ($M_{\rm halo} > 1{\times}10^{13}\,M_\odot$), are always hosted by central galaxies (`flag_central` = 1), and their hosts have high stellar masses $M_{\rm stellar} > 1 \times 10^{11}\,M_\odot$. This selection yields a sample with both a number density and redshift distribution similar to those found in the observed QSO sample.

2.2.2. 2dFLenS

2dFLenS is similar to the BOSS sample, but occupies the southern KiDS fields and consists of three subsamples with different selection functions: the low-z, the mid-z, and the high-z sample. The selection criteria for each of these samples are given in Table A.2. The low-z and mid-z selections are very similar to the BOSS LOWZ and CMASS selections, and are based on the same complement of *gri*-magnitudes and auxiliary colours (i.e. Eqs. (7)–(9), Blake et al. 2016). We apply these cuts again with minor parameter fine-tuning to better reproduce the observed redshift and colour/magnitude distributions. The high-z sample, however, contains additional selections incorporating fluxes from the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010) $W1$-band (central wavelength $\bar{\lambda} = 3.6\,\mu$m), which is not available in MICE2. We approximate this selection using the VIKING K_s-band (central wavelength $\bar{\lambda} = 2.15\,\mu$m) in place of $W1$. While not exact, comparisons between the redshift distributions in MICE2 and the 2dFLenS data suggest that this approximation is suitable for the selection accuracy required in this work. Furthermore, 2dFLenS applies a density downsampling after selecting targets according to the criteria in Table A.2, resulting in approximately half the density of the corresponding BOSS LRG samples. This downsampling is partly comprised in our modified selection, which yields an approximately 30% higher density than the 2dFLenS data sample. We consider this difference sufficiently small for our purposes that we do not implement an additional downsampling for the MICE2 sample.

2.2.3. GAMA

GAMA is a flux-limited spectroscopic survey distributed over three equatorial fields that overlap with the northern KiDS fields. The survey data used here (Driver et al. 2011; Liske et al. 2015) is highly complete (>98%) to the magnitude limit of $r = 19.8$ mag, with the bulk of galaxies residing at redshifts of $z_{\rm spec} \lesssim 0.4$. We find that a single $r < 19.87$ selection is best suited to reproduce the redshift distribution of GAMA in our MICE2 mocks.

2.2.4. WiggleZ

WiggleZ is the deepest of the wide area surveys overlapping with KV450, extending to $z_{\rm spec} \lesssim 1.1$. This dataset spans the gap between the wide and deep spectroscopic datasets. The WiggleZ selection function (Drinkwater et al. 2010) consists of a number of cuts which both include and exclude certain parts of the far-UV to near-IR colour-colour space, and are summarised in Table A.3. As we do not have mock UV photometry in the MICE2 catalogues, we are unable to fully reproduce the literature selection function of WiggleZ. This deficiency means that our initial selection function produces a markedly different redshift distribution in the mocks compared to the data. To accurately model the redshift distribution of WiggleZ we are therefore required to - after performing all the possible selections - apply a direct matching of the simulation and data redshift distributions. We do this via a direct redshift-dependent downsampling of our initial WiggleZ sample to match the simulations to the data.

2.2.5. DEEP2

DEEP2 has been covered by KiDS- and VIKING-like observations in two fields at RA $\approx$ 02 h and 23 h. Sources in these fields are pre-selected by colour to target galaxies in the range $0.75 \lesssim z_{\rm spec} \lesssim 1.5$. This is achieved by cuts in the Johnson *B-R*, *R-I* colour-colour space (Newman et al. 2013). These selections are listed in Table A.4, and the resulting colour- and magnitude-distributions are shown in Fig. 6. Johnson magnitudes are available within MICE2, and so we are able to apply this selection directly to the simulations. When we apply these cuts to MICE2, however, we find significant contamination by low redshift objects in the sample, and a clear shift of the cut-off redshift from $z > 0.75$ to $z > 0.65$. Informed by the redshift distribution of the *B-R*, *R-I* colour-colour space within MICE (shown graphically in Fig. 7), we adjust two of the colour cuts to obtain redshift and colour distributions that are closer to those observed in the DEEP2 observations. The influence of our updated colour-cuts in DEEP2 can be seen in Fig. 8, which shows clearly that our updated colour cuts yield a closer match to the observed DEEP2 redshift distribution. We hypothesise that the need for different colour cuts in MICE2 is partially due to our use of noiseless magnitudes in the sample definition, but moreover is because of an extension of the colour-redshift space, at intermediate redshifts, to more negative colours than is seen in the data. Finally, we model spectroscopic incompleteness within the

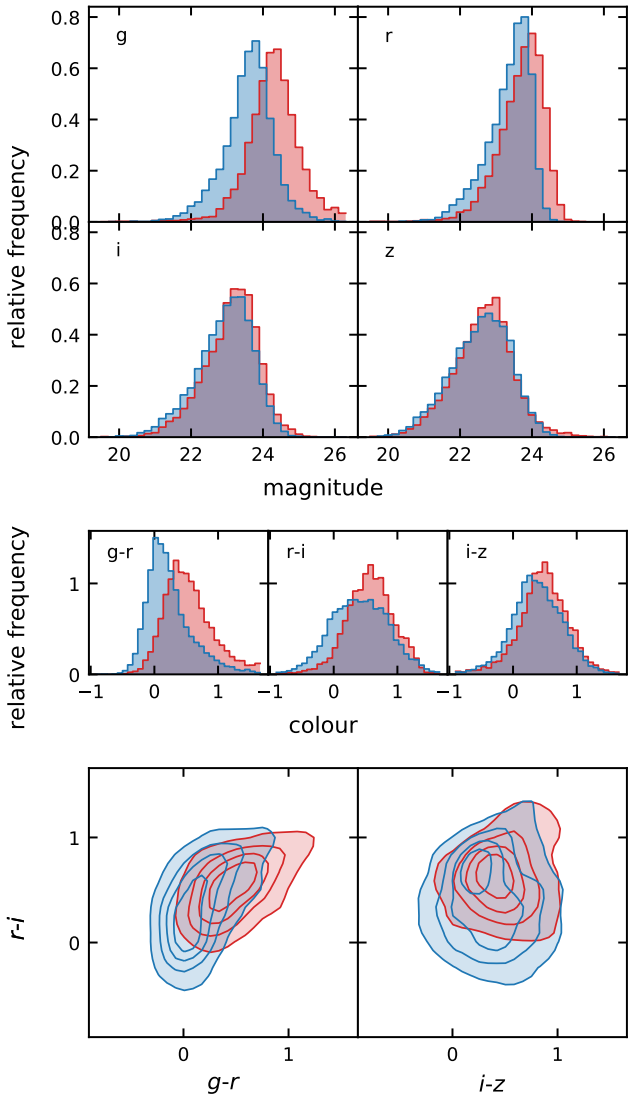

Fig. 6. Comparison of colour and magnitude distributions for the simulated (blue) and observed (red) DEEP2 datasets.

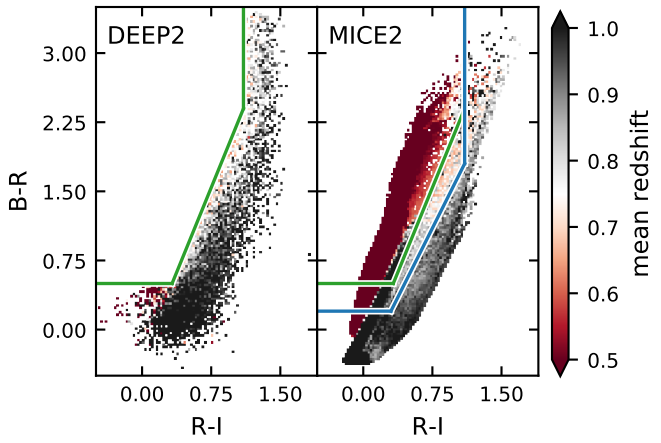

Fig. 7. *Left*: DEEP2 spectroscopic data in $R-I$-$B-R$ colour space with the original target selection indicated in green, *right*: MICE2 galaxies (based on noiseless model magnitudes) with the original and the fiducial data selection indicated in blue.

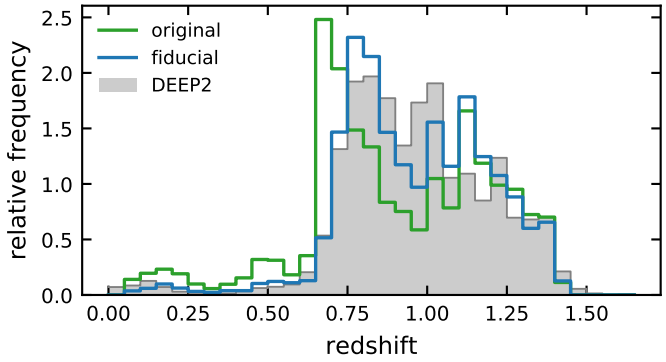

Fig. 8. DEEP2 redshift distribution in comparison to the redshift distribution of MICE2 galaxies selected with the original colour cut (Newman et al. 2013, green) and our fiducial colour cut (blue).

DEEP2 sample using the redshift-completeness function presented in Newman et al. (2013), which shows a clear decrease in the fraction of sources with high-confidence ($nQ \geq 3$) redshifts with increasing R-band magnitude. Finally, we randomly downsample the remaining sample to ∼60% of its initial size, to obtain a similar number of objects as found in the observational data.

2.2.6. VVDS

The VVDS field at RA $\approx$ 02 h is selected via a combination of both a wide-field selection and a deep-drill selection realised by simple magnitude cuts which add additional spectra over the whole $0 < z_{\rm spec} \lesssim 1.3$ redshift range. However, the deep sample is overwhelmingly dominant in this field, and so we opt to simulate this selection only. The deep sample is defined by a simple Johnson I-band magnitude limit, $I < 24.0$. We find that implementing this limit in the simulations, without modification, results in a sample well matched to the observations. We implement the literature spectroscopic success rate for VVDS (Le Fèvre et al. 2005, Figs. 13.a/b and 16) as a function of both I-band magnitude and true-redshift. These selections are performed independently; that is, we assume no correlation between the sources removed in the selection. Finally, the VVDS spectroscopic sample has a roughly 25% completeness in the 2hr field; we find, however, that a ∼17% completeness is required to

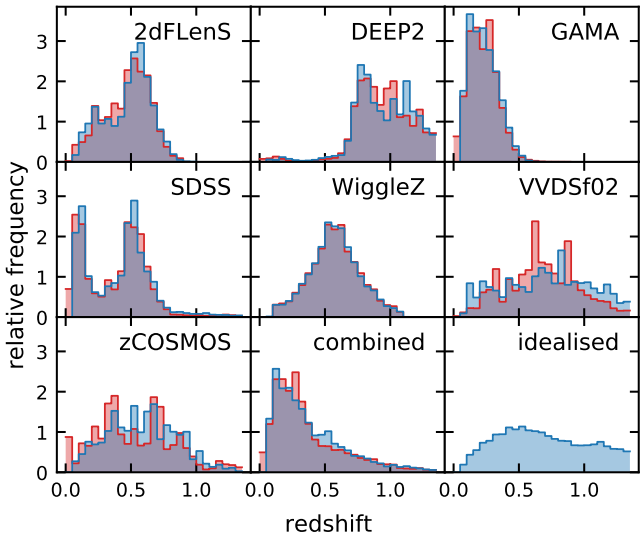

Fig. 9. Redshift distributions of the spectroscopic data catalogues (red) compared to their corresponding MICE2 mock counterparts (blue). The idealised sample in the *bottom right panel* is used for benchmarking purposes only.

reproduce (in MICE2) the number density of VVDS spectra seen in the observations. Hence, we down-sample the VVDS mock catalogue to that density. A comparison of observed and simulated colours/magnitudes in VVDS can be found in Fig. A.1.

2.2.7. zCOSMOS

As with the VVDS sample, our zCOSMOS sample is a combination of two distinct subsamples: the public "bright" and a proprietary "deep" sample. The deep sample in zCOSMOS preferentially targets objects at $z > 1.5$, which is beyond the maximum redshift available in MICE2. Therefore, we opt to simulate only the bright selection, which is defined by Johnson I-band magnitudes in the range $15.0 < I < 22.5$. We apply this selection as-is to MICE2, finding good agreement in the colours and redshift distributions between the simulations and the observations. We apply the spectroscopic success rate for zCOSMOS, which is given as a joint function of I-band magnitude and redshift (Lilly et al. 2009, Fig. 3). Finally, we randomly down-sample the resulting spectra by 33%, to match the observed zCOSMOS spectroscopic number density. Figure A.2 shows zCOSMOS and our mock sample in colour and magnitude space. We note the absence of the deep sample creates a clear dearth of spectra at faint magnitudes, but does not seem to systematically bias the colour-colour space in our zCOSMOS simulation.

2.2.8. Idealised spectroscopic sample

Finally we create an idealised mock spectroscopic sample, defined by selecting MICE2 sources with $r < 24.0$ that lie within the footprint of our 2dFLenS and SDSS mock galaxy samples. This sample is then sparse-sampled to a number density that is ∼10% of the mock KV450 number density. Such an idealised sample allows the computation of clustering redshifts for our mock KV450 sample excluding the influence of spectroscopic selection functions, and therefore allows us to estimate the influence of these selections on our results.

The final redshift distributions of all our mock spectroscopic compilations are shown in Fig. 9. The mock redshift istributions

Table 2. Comparison of galaxy densities and median redshifts of the spectroscopic data and mock samples.

Survey	$n_{\mathrm{MICE}}/\mathrm{arcmin}^{-2}$	$n_{\mathrm{spec}}/\mathrm{arcmin}^{-2}$	$z_{\mathrm{MICE}}^{\mathrm{median}}$	$z_{\mathrm{spec}}^{\mathrm{median}}$
2dFLenS	0.020	0.015	0.51	0.50
GAMA	0.289	0.292	0.21	0.22
SDSS (all)	0.052	0.047	0.46	0.54
WiggleZ	0.063	0.072	0.57	0.57
DEEP2	2.96	2.96	0.96	0.96
VVDS-02h	2.54	2.52	0.74	0.69
zCOSMOS	4.56	4.58	0.59	0.52

agree well with their data counterparts for both mean and median statistics (excluding those whose data distributions exhibit considerable tails above $\max(z_{\mathrm{sim}}) = 1.4$) to within $|\Delta z| \lesssim 0.05$ (see Table 2). These samples are quite different from Scottez et al. (2018), who also use MICE2 to study the performance of clustering redshifts, since the ones presented here reflect some of the complications that arise in practice with spectroscopic reference samples, such as realistic spatial overlaps or redshift incompleteness, which results in more complex selection functions.

3. Redshifts from cross-correlations

Due to the gravitational clustering of matter in the Universe, the positions of objects which reside in a common volume (i.e. in the same large-scale structure) are highly correlated. Conversely, the positions of objects from disparate volumes/structures are uncorrelated (except for a small contribution from magnification, see e.g. Gatti et al. 2018). We can use this fact to constrain the redshift distributions of an ensemble of extragalactic sources by measuring the amplitude of their angular cross-correlation with a tracer sample of galaxies with known redshifts. This approach to redshift distribution estimation is known as "cross-correlation redshifts" or simply "clustering redshifts".

3.1. Basic formalism

In the literature there are several different approaches to clustering redshift estimation (e.g. Newman 2008; McQuinn & White 2013; Ménard et al. 2013). In this work we follow an approach similar to Johnson et al. (2017), described briefly here.

Consider two samples of extragalactic objects which overlap in three dimensions:

1. A reference sample (s) with known angular positions and redshift distribution $n_{\mathrm{s}}(z)$ obtained from secure point redshift estimates (typically spectroscopic redshifts).

2. A target sample (p), also with known angular positions but without precise redshift information (typically selected photometrically); the redshift distribution of this sample, $n_{\mathrm{p}}(z)$, is what we wish to recover.

The angular cross-correlation of sources within the reference and target samples, at a fixed reference-sample redshift z and separation angle θ, can be estimated by projection along the line of sight:

$$w_{\mathrm{sp}}(\theta, z) = b_{\mathrm{s}}(\theta, z) \int_0^\infty \mathrm{d}z'\, n_{\mathrm{p}}(z')\, b_{\mathrm{p}}(\theta, z')\, \xi\left[R(\theta, z, z'), z\right], \quad (10)$$

where $n_{\mathrm{p}}(z)$ is the redshift probability distribution of the target sample and $b_{\mathrm{s}}(\theta, z)$ and $b_{\mathrm{p}}(\theta, z)$ are terms for the scale dependent redshift evolution of the linear galaxy bias in both samples. Finally, $\xi(R, z)$ is the matter auto-correlation function at redshift z and comoving, 3-dimensional separation

$$R(\theta, z, z') = \sqrt{\left[\chi(z) - \chi(z')\right]^2 + \left[f_K(z')\,\theta\right]^2}\,. \quad (11)$$

Here, $\chi(z)$ is defined as the radial comoving distance and $f_{\mathrm{K}}(z)$ as the comoving angular diameter distance to a given redshift z.

In practise we compute the cross-correlation in narrow bins of redshift of which each has a width of Δz (this bin width can vary with redshift). In the following we will make a series of assumptions that allow us to simplify Eq. (10) significantly. First, we assume that the redshift distributions and bias evolution terms $n_{\mathrm{s}}(z)$, $n_{\mathrm{p}}(z)$, $b_{\mathrm{s}}(z)$ and $b_{\mathrm{p}}(z)$ are constant over the interval of each redshift bin. In presence of significant sample variance, strongly varying sample selections, or insufficiently fine redshift binning, this assumption is likely violated, potentially causing biases in the recovered redshift distribution $n_{\mathrm{p}}(z)$. Secondly, we assume that the redshift bins have sufficient radial extent such that neighbouring bins are uncorrelated. This allows us to reduce the integration limits in Eq. (10) to a single bin and combined, these two assumptions yield

$$w_{\mathrm{sp}}(\theta, z) = n_{\mathrm{p}}(z)\, b_{\mathrm{s}}(\theta, z)\, b_{\mathrm{p}}(\theta, z) \int_z^{z+\Delta z} \mathrm{d}z'\, \xi\left[R(\theta, z, z'), z\right]. \quad (12)$$

The remaining integral is then simply the angular matter auto-correlation function of that particular bin, $w_{\mathrm{mm}}(\theta, z)$. Finally, we express the angular separation θ in terms of the projected physical scale $r = \theta\chi/(1+z)$ at given redshift z using the flat-sky and small-angle approximations:

$$w_{\mathrm{sp}}(r, z) = n_{\mathrm{p}}(z)\, b_{\mathrm{p}}(r, z)\, b_{\mathrm{s}}(r, z)\, w_{\mathrm{mm}}(r, z). \quad (13)$$

In a similar manner we can derive terms for the reference and target sample angular autocorrelation functions. Analogous to Eq. (10) we define

$$w_{\mathrm{ss}}(\theta, z) = b_{\mathrm{s}}(\theta, z) \int_z^{z+\Delta z} \mathrm{d}z'\, n_{\mathrm{s}}(z')\, b_{\mathrm{s}}(\theta, z')\, \xi\left[R(\theta, z, z'), z\right]. \quad (14)$$

Applying the same assumptions as in Eq. (12) yields

$$w_{\mathrm{ss}}(\theta, z) = \frac{b_{\mathrm{s}}^2(\theta, z)}{\Delta z} \int_z^{z+\Delta z} \mathrm{d}z'\, \xi\left[R(\theta, z, z'), z\right], \quad (15)$$

where we have substituted $n_{\mathrm{s}}(z) \equiv 1/\Delta z$ since the redshift distribution does not vary over Δz. Again, we identify the integral as the angular matter auto-correlation function $w_{\mathrm{mm}}(\theta, z)$ of that particular bin and rearrange to obtain an expression for the reference sample bias evolution

$$b_{\mathrm{s}}(\theta, z) = \sqrt{\Delta z\, \frac{w_{\mathrm{ss}}(\theta, z)}{w_{\mathrm{mm}}(\theta, z)}}. \quad (16)$$

By repeating this approach we obtain an analogous term for the target sample bias evolution (substituting s → p in Eqs. (14)–(16)).

After expressing the angles in projected physical separation we can express the bias terms in Eq. (13) by the sample auto-correlation functions and solve for our redshift distribution of interest:

$$n_{\mathrm{p}}(z) = \frac{w_{\mathrm{sp}}(r, z)}{\sqrt{\Delta z^2\, w_{\mathrm{ss}}(r, z)\, w_{\mathrm{pp}}(r, z)}}. \quad (17)$$

In summary, it is possible to estimate an unknown redshift distribution $n_p(z)$ by measuring the cross-correlation of p with our tracer sample of known redshift s. However, this redshift estimate is degenerate with the redshift evolution of the bias factors, $b_s(z)$ and $b_p(z)$, which can be, in principle, measured through the individual sample auto-correlation functions w_{ss} and w_{pp}.

The simple relation, presented in Eq. (17), requires some non-trivial assumption, such as linear galaxy bias, which is typically violated on small scales, and parameterising the redshift distributions and bias terms through step functions. Depending on the chosen redshift binning and the clustering of the reference and target sample, this can introduce significant systematic errors. Furthermore, it is very challenging to correct for the target sample bias, since measuring w_{pp} requires binning the target sample into the same narrow redshift bins Δz that we use to slice the reference sample. This would require accurate redshift point estimates for all galaxies in the target sample. We give a brief overview of the most common literature approaches to bias mitigation below.

3.2. Cross-correlation methods and bias mitigation

Newman (2008) and Matthews & Newman (2010) parameterise galaxy bias by modelling the correlation functions with power laws:

$$\xi(r, z) = \left(\frac{r}{r_0(z)}\right)^{-\gamma(z)}, \tag{18}$$

where r_0 is the correlation length, and γ defines the shape of the correlation function. Assuming linear biasing, the cross-correlation can be written as $\xi_{sp} = \sqrt{\xi_{ss}\,\xi_{pp}}$ and $r_{0,sp}$ and γ_{sp} can be calculated from the parameters of the auto-correlation functions. Newman (2008) obtains $r_{0,ss}(z)$ and $\gamma_{ss}(z)$ by fitting the reference sample's auto-correlation function, measured in bins of redshift, and an average value for γ_{pp} by fitting the angular auto-correlation function of the target sample. Since $r_{0,pp}$ cannot be measured without redshift information, the authors assume that it is constant which allows to break the degeneracy of the redshift distribution and the galaxy bias. They apply an iterative approach to obtain an estimate for $r_{0,pp}$ averaged over the redshift baseline of the target sample:

1. Make an initial guess for $r_{0,pp}$ and compute $r_{0,sp}(z)$ and $\gamma_{sp}(z)$.
2. Estimate the redshift distribution $n_p(z)$ by fitting the measured cross-correlation with the power-law model.
3. De-project the angular auto-correlation using Eq. (12) and the redshift distribution from step 2, and thereby obtain a new guess for $r_{0,pp}$.
4. Repeat steps 2 and 3 until convergence is reached.

Newman (2008) restricts the cross-correlation measurements to scales of $2 < r < 10$ Mpc to avoid the highly non-linear biasing regime where the assumption $\xi_{sp} = \sqrt{\xi_{ss}\,\xi_{pp}}$ might no longer be valid.

McQuinn & White (2013) and an extension by Johnson et al. (2017) build on Newman (2008)'s approach and construct an estimator that optimally weights the correlation scales to improve the S/N of the recovered clustering redshift distribution.

Ménard et al. (2013) and Schmidt et al. (2013) demonstrate that using small scales for the correlation measurements is extremely valuable, as these scales carry the strongest correlation signal. They conclude that the systematic errors introduced by violating the assumption of linear, deterministic bias are outweighed by an improved S/N when measuring on scales $r < 1$ Mpc. Furthermore, they suggest using a single-bin correlation measurement:

$$\bar{w}_{sp}(z) = \int_{r_{min}}^{r_{max}} \mathrm{d}r\, W(r)\, w_{sp}(r, z), \tag{19}$$

where $W(r) \propto r^{\beta}$ is a weight function. For $\beta = -1$ this amounts to weighting galaxy pairs by their inverse separation distance, which further increases the S/N and the sensitivity to galaxy bias by up-weighting the smallest scales.

The authors also suggest two methods of bias mitigation for the reference and target samples. First, the reference sample auto-correlation function yields an estimate of the bias evolution with redshift, when measured on the same scales and with the same weighting as the cross-correlation function (see Eq. (17)):

$$\tilde{n}_p(z) = \frac{\bar{w}_{sp}(z)}{\sqrt{\Delta z\, \bar{w}_{ss}(z)}} = n_p(z)\sqrt{\Delta z\, \bar{w}_{pp}(z)}, \tag{20}$$

where the barred correlation functions are scale-weighted according to Eq. (19). Here we have defined $\tilde{n}_p(z)$ as clustering redshift distribution corrected for the reference sample bias. Secondly, they note the correlation between the width of the target sample redshift distribution and its sensitivity to the galaxy bias redshift evolution. By pre-selecting narrow redshift bins (e.g. using magnitudes, colours or photometric redshifts) the impact of the redshift-evolution of the bias can be reduced. Additionally, Newman (2008)'s iterative bias technique can be applied to each bin individually to recover the step-wise redshift evolution of the bias.

Finally, Davis et al. (2018) parameterise the bias of the target sample through a simple power-law:

$$\mathcal{B}_\alpha(z) = (1+z)^{\alpha} \propto \sqrt{\Delta z\, \bar{w}_{pp}(z)}. \tag{21}$$

The normalisation of this parameterisation is arbitrary; it is degenerate with the normalisation of the resulting redshift distribution.

3.3. Bias mitigation with self-consistency

In Hildebrandt et al. (2020a) we adopted a method to constrain the bias of the target sample using an analytical model (e.g. Eq. (21)), which is based on ideas developed in Schmidt et al. (2013), Ménard et al. (2013), and Morrison et al. (2017). The method leverages the response of clustering redshifts to the target sample bias as a function of the target sample redshift distribution width.

3.3.1. Method

We split the target sample into N_{bin} (preferentially narrow) bins using a secondary redshift indicator, such as photometric redshifts. Then we measure clustering redshifts $\tilde{n}_{p,j}(z)$ for each of these bins. If the bias evolves with redshift, the sum of these binned measurements must differ from a clustering redshift measurement $\tilde{n}_{p,tot}(z)$ (Eq. (20)) over the full target sample:

$$\sum_{j=1}^{N_{bin}} \mathcal{W}_j\, \tilde{n}_{p,j}(z) \neq \tilde{n}_{p,tot}(z), \tag{22}$$

where $\mathcal{W}_j$ is the total weight of the jth tomographic bin, obtained by summing over the individual weights of all galaxies contained in that bin. This is because each of the $\tilde{n}_{p,j}(z)$ is

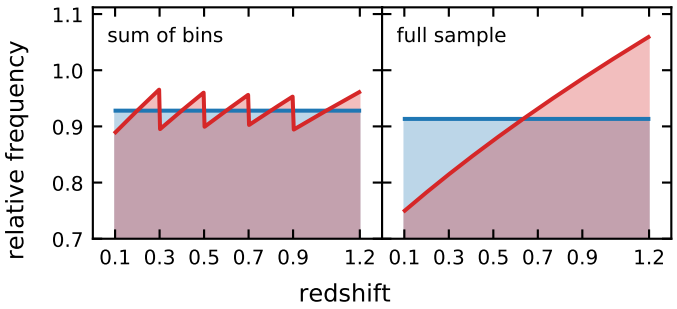

Fig. 10. Toy-model showing the impact of the bias evolution on clustering redshifts for wide redshift distribution compared to measurements on narrow redshift bins.

normalised individually and this normalisation factor depends on the mean bias amplitude per bin and allows, in principle, to break the degeneracy between redshift and bias evolution. A toy-model demonstration of this effect can be seen in Fig. 10. We assume a flat redshift distribution in all bins, shown in blue, and a redshift evolution of the sample bias of $\mathcal{B}_\alpha(z) \propto (1+z)^\alpha$, where we set $\alpha = 0.5$. This results in the recovered, biased clustering redshift distributions shown in red. The left- and right-hand side of this figure correspond to the left- and right-hand side terms of Eq. (22). Whereas the full sample has the complete bias evolution imprinted, the normalisation of each redshift bin leads to a sawtooth-shaped redshift distribution.

These differences between the full sample redshift distribution and the sum of the redshift bins allows us to constrain the bias evolution, if we have a sufficiently accurate model. We can estimate the model parameter α of the bias model $\mathcal{B}_\alpha(z)$ by minimising

$$\Delta(\alpha) = \int \mathrm{d}z \left[\mathrm{norm}\left(\frac{\tilde{n}_{\mathrm{p,tot}}(z)}{\mathcal{B}_\alpha(z)}\right) - \sum_{j=1}^{N_\mathrm{bin}} \mathcal{W}_j \, \mathrm{norm}\left(\frac{\tilde{n}_{\mathrm{p},j}(z)}{\mathcal{B}_\alpha(z)}\right)\right]^2 , \quad (23)$$

where $\mathrm{norm}[f(z)] = f(z) \,/\, \int_0^\infty \mathrm{d}z' \, f(z')$.

The advantage of this approach is that it allows us to correct for any combination of biases in clustering redshifts, $\bar{b}_\mathrm{p}(z)$, $\bar{b}_\mathrm{s}(z)$ or the combination of both. The bias model, however, must be fairly simplistic since the amount of information that can be extracted from Eq. (23) is small and depends on total number of bins and the degree of overlap of their respective redshift distributions. The greater the overlap between the redshift bins is, the stronger is the correlation between their redshift distributions and the harder it is to constrain the bias evolution. Furthermore, this approach assumes that all bins follow the same universal bias evolution. This is not true in general, since any redshift binning will select a different population of galaxies which cluster differently. This effect is small for the tomographic redshift bins of the KV450 mock galaxy sample, but may be larger for more complex sample selections. Finally, we note that this mitigation approach can be applied to a variety of the existing clustering-redshift methods (see overview in Sect. 3.2) as a post-processing step.

We will call this mitigation approach "self-consistent bias mitigation" (SBM) in the following.

3.3.2. Tests on mock data

We apply the SBM to both, clustering redshifts obtained from the idealised reference sample as well as from the compilation of realistic reference samples, to assess whether it is able to recover

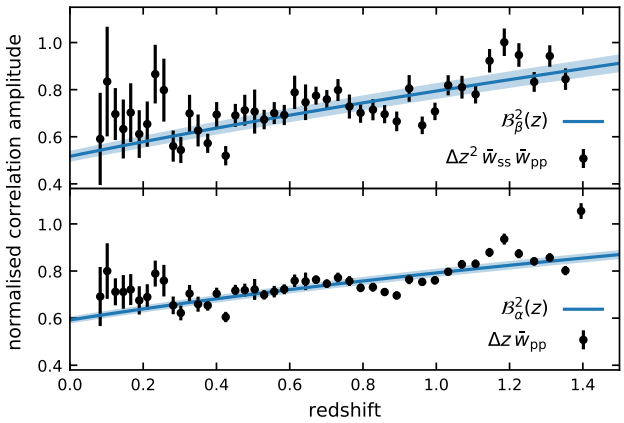

Fig. 11. Bias model (blue lines; corresponding to the bold-face numbers in Table 3) fitted directly to the unaccounted bias terms of the raw (black data points; *top panel*) and the reference sample bias corrected clustering redshifts (black data points; *bottom panel*) for the idealised setup. The correlation amplitudes are re-normalised.

Table 3. Best-fit values for the bias model parameter β (modelling the reference and target sample bias), α (modelling only the target sample bias), and γ (null test).

Method	Setup	β	α	γ
SBM	Idealised	-0.03 ± 0.11	-0.09 ± 0.10	-0.26 ± 0.10
	realistic	0.50 ± 1.05	0.88 ± 1.05	0.82 ± 1.04
Direct fit	idealised	$0.31 \pm 0.03^{(\star)}$	$0.21 \pm 0.01^{(\star)}$	–
	realistic	-0.06 ± 0.04	0.18 ± 0.02	–

Notes. The top group summarises the fit results from the SBM, the bottom group from direct fits to the auto-correlation function terms. The two values marked by ($\star$) correspond to the blue model fits in Fig. 11.

the bias evolution terms correctly. In either case we expect the SBM best fit to match $\sqrt{\Delta z \, \bar{w}_\mathrm{pp}(z)}$ (see Eq. (20)) when applied to $\tilde{n}_\mathrm{p}(z)$, where we have already corrected the reference sample bias by measuring its autocorrelation function. We conduct additional tests, namely fitting the raw cross-correlations $\bar{w}_\mathrm{sp}(z)$ with

$$\mathcal{B}_\beta(z) = (1+z)^\beta \propto \sqrt{\Delta z^2 \, \bar{w}_\mathrm{ss}(z) \, \bar{w}_\mathrm{pp}(z)} \quad (24)$$

and fitting the fully bias-corrected clustering redshifts $n_\mathrm{p}(z)$ (Eq. (17)) with $\mathcal{B}_\gamma(z) = (1+z)^\gamma$. The latter serves as a null test, since there should be no residual bias evolution, hence $\mathcal{B}_\gamma(z) \approx 1$ and $\gamma \approx 0$.

We validate these SBM parameter estimates by directly fitting the bias model to the autocorrelation terms, i.e. fitting the right hand side terms with the left hand side models of Eqs. (20) and (24). These autocorrelation terms and direct fits are presented in Fig. 11. The correlation measurements reveal that the bias evolution is small for all the mock data samples. The galaxy bias of the target sample increases by ≈25% over the MICE2 redshift baseline. Furthermore, the direct fitting shows that the power-law bias model (Eq. (21)) is able to recover the global trend of the bias evolution, despite having only a single free parameter. The best-fit values for α, β and γ obtained from both, the SBM and the direct auto-correlation function fits, are summarised in Table 3.

The results we obtain from the idealised mock sample using the SBM are not in agreement with the direct fits and instead predict negative values in all three cases. The values themselves are

relatively small though because the idealised sample is purely magnitude limited and hence should not show a very strong bias evolution. This example illustrates the limitations of the SBM. The impact of this on the mean redshifts is discussed in Sect. 4.1.

In case of the realistic mocks we find that α, determined via the SBM, is of order unity but, due to the large parameter uncertainty, consistent with the value expected from the idealised setup. We note that the uncertainties reported in Table 3 are significantly larger than those reported in Hildebrandt et al. (2020a). This results from `the-wizz` error estimates adopted by Hildebrandt et al. (2020a), which we find underestimate the true uncertainty on the measured correlation functions at high redshifts. Our revised pipeline `yet_another_wizz` (see Sect. 3.4 below) provides accurate error estimates that serve to increase the overall uncertainty on α.

3.4. Implementation

We have used clustering redshifts to obtain the redshift distributions of the KiDS-450 and the KiDS+VIKING-450 datasets in previous works (Hildebrandt et al. 2017, 2020a). The major difference to the latter is that we are using an improved implementation of the Schmidt et al. (2013) clustering-redshift method.

Previously, we used `the-wizz`[5] (Morrison et al. 2017) for clustering redshifts and a `python`-binding[6] of the spherical pixelation library `STOMP` (Scranton et al. 2002) for efficient angular correlation measurements. The issue with this approach is that `STOMP` (and therefore `the-wizz`) approximates the correlation annulus, in which the pairs are counted, by selecting entire `STOMP` pixels. Since we compute correlations on constant physical scales, the area of the annulus may change discretely with position and redshift, depending on the `STOMP` pixel resolution and the angular diameter corresponding to the given projected physical separation. This can cause biases in the recovered redshift distribution. During the development stage of this project, the latest version of `the-wizz`, which no longer depends on `STOMP`, was not yet available. Therefore we implemented a simplified version of `the-wizz`, called `yet_another_wizz`[7] to circumvent the pixelation problems. Similar to `the-wizz`, this code measures the correlation on a single annulus of fixed projected physical separation r_p, weighted by the inverse distance of the pairs. Even though the calculation of $r_\mathrm{p}(z)$ requires assuming a cosmological model, the results are only very weakly dependent on the choice of cosmological parameters. The inverse distance weight significantly increases the S/N of the correlation amplitude (Schmidt et al. 2013). Similar approaches have been used e.g. by McQuinn & White (2013) or Alarcon et al. (2020) to optimise the correlation signal.

We compute the single-bin correlation amplitude (see Eq. (19)) from pair counts using the Davis-Peebles correlation estimator (Davis & Peebles 1983)

$$\bar{w} = \frac{n_\mathrm{R} \int_{r_\mathrm{min}}^{r_\mathrm{max}} \mathrm{d}r\, W(r)\, \mathrm{DD}(r)}{n_\mathrm{D} \int_{r_\mathrm{min}}^{r_\mathrm{max}} \mathrm{d}r\, W(r)\, \mathrm{DR}(r)} - 1 \tag{25}$$

with an inverse distance weight $W(r) \propto r^{-1}$. DD is the number of (ordered) pairs[8] between reference and target sample galaxies and DR between reference galaxies and target sample random points. The ratio $n_\mathrm{R}/n_\mathrm{D}$ (sum over the galaxy weights divided by the sum over the random point weights) accounts for the average density of the data catalogue and its typically much larger random representation. In general it would be favourable to use the Landy-Szalay estimator (Landy & Szalay 1993) instead of the Davis-Peebles estimator, albeit this would come at the cost of increased computational complexity. We find that the difference between the estimators in our analysis is negligible compared to other systematic errors, such as the evolving target sample bias. Finally, we choose to measure all angular correlations on scales from $r_\mathrm{min} = 100\,\mathrm{kpc}$ to $r_\mathrm{max} = 1000\,\mathrm{kpc}$. This range is a good trade-off between high signal-to-noise and measuring on scales with highly non-linear biasing.

The whole clustering-redshift pipeline as well as our bootstrap resampling-based method for combining cross-correlation measurements from different reference samples and covariance estimation is described in detail in Appendix B.

4. Results

In this section we report the results of applying our method to the KV450 simulations. We focus on the effects of galaxy bias, the choice of measurement scales, and the effect of lensing magnification.

4.1. Impact of galaxy bias

There are different degrees of bias correction that can be applied before getting a redshift estimate $n_\mathrm{CC}(z)$ from cross-correlation measurements:

1. The raw cross-correlation with no correction ($n_\mathrm{CC} = \bar{w}_\mathrm{sp}$),
2. Mitigating the reference sample bias using its auto-correlation function ($n_\mathrm{CC} = \tilde{n}_\mathrm{p}$, Eq. (20)),
3. Additionally mitigating the target sample bias with the SBM ($n_\mathrm{CC} = \tilde{n}_\mathrm{p}/\mathcal{B}_\alpha$), and
4. Mitigating the target sample bias using its auto-correlation function ($n_\mathrm{CC} = n_\mathrm{p}$, Eq. (17), only possible on mock data using the true redshifts).

It is difficult to directly compare each of these four redshift estimates. Due to negative correlation amplitudes, originating from systematic effects and statistical noise, it is dangerous to interpret cross-correlation-derived redshift estimates directly as probability distributions. This prevents us from using them directly in cosmological applications such as cosmic shear, unless we model the redshift distributions such that negative amplitudes are mitigated.

We implement this modelling by adopting the true target sample redshift distribution $p(z)$ as a model and fit it to the clustering redshift estimate $n_\mathrm{CC}(z)$, allowing a free normalisation amplitude A and a free shift parameter Δz. We minimise

$$\chi^2 = \left[n_\mathrm{CC}(z) - A\, p(z + \Delta z)\right]^T C^{-1} \left[n_\mathrm{CC}(z) - A\, p(z + \Delta z)\right], \tag{26}$$

where we shift $p(z)$ according to Δz and apply the binning of the cross-correlation to the model redshift distribution. We perform these fits jointly for all tomographic bins[9] to capture the full correlation of the shift parameters Δz_i. These shift parameters $\Delta z_i \approx \langle z_\mathrm{CC}\rangle_i - \langle z_\mathrm{true}\rangle_i$ (presented in Sect. 4.1.1) give us a direct estimate of the systematic shifts in the four different redshift estimates listed above, which may originate from evolving galaxy

[5] https://github.com/morriscb/the-wizz
[6] https://github.com/jlvdb/astro-stomp3
[7] https://github.com/jlvdb/yet_another_wizz
[8] Each pair is weighted by the product of the individual weights of the partners.
[9] See Sect. 3.4 for a summary of the covariance estimation recipe.

Table 4. Shift fit parameters for different bias correction methods for clustering redshifts obtained using the idealised and realistic spectroscopic mock samples using the true redshift distributions as fit model.

Setup	$n(z)$-type	$100 \times \Delta z_1$	$100 \times \Delta z_2$	$100 \times \Delta z_3$	$100 \times \Delta z_4$	$100 \times \Delta z_5$	χ^2	$n_{\rm dof}$
Idealised	$\bar{w}_{\rm sp}$	$-0.08^{+0.11}_{-0.13}$	$-0.12^{+0.09}_{-0.11}$	$-0.40^{+0.12}_{-0.08}$	$-0.79^{+0.13}_{-0.20}$	$-0.07^{+0.20}_{-0.18}$	263.0	215
	$\tilde{n}_{\rm p}$	$-0.22^{+0.08}_{-0.14}$	$-0.15^{+0.10}_{-0.10}$	$-0.51^{+0.12}_{-0.12}$	$-0.80^{+0.12}_{-0.18}$	$-0.26^{+0.16}_{-0.19}$	261.5	215
	$\tilde{n}_{\rm p}/\mathcal{B}_\alpha$	$-0.15^{+0.12}_{-0.12}$	$-0.14^{+0.09}_{-0.10}$	$-0.45^{+0.13}_{-0.10}$	$-0.70^{+0.16}_{-0.14}$	$-0.19^{+0.21}_{-0.23}$	249.5	215
	$n_{\rm p}$	$-0.24^{+0.12}_{-0.11}$	$-0.16^{+0.10}_{-0.09}$	$-0.68^{+0.10}_{-0.13}$	$-0.82^{+0.10}_{-0.17}$	$-0.76^{+0.41}_{-0.33}$	282.0	215
Realistic	$\bar{w}_{\rm sp}$	$-0.32^{+0.47}_{-2.10}$	$0.74^{+0.19}_{-0.17}$	$0.01^{+0.20}_{-0.39}$	$-0.41^{+0.44}_{-0.49}$	$-2.55^{+2.27}_{-2.89}$	234.6	100
	$\tilde{n}_{\rm p}$	$-0.44^{+0.43}_{-0.48}$	$0.31^{+0.22}_{-0.19}$	$-0.36^{+0.19}_{-0.24}$	$0.60^{+0.64}_{-0.58}$	$0.32^{+0.49}_{-2.40}$	157.3	100
	$\tilde{n}_{\rm p}/\mathcal{B}_\alpha$	$0.83^{+0.37}_{-0.93}$	$0.20^{+0.41}_{-0.60}$	$-0.17^{+0.54}_{-0.60}$	$-1.55^{+0.38}_{-0.71}$	$-0.93^{+4.00}_{-1.18}$	93.7	100
	$n_{\rm p}$	$-0.88^{+0.25}_{-0.33}$	$0.38^{+0.20}_{-0.14}$	$-0.46^{+0.16}_{-0.26}$	$0.61^{+0.65}_{-0.56}$	$0.30^{+0.87}_{-0.45}$	169.2	100

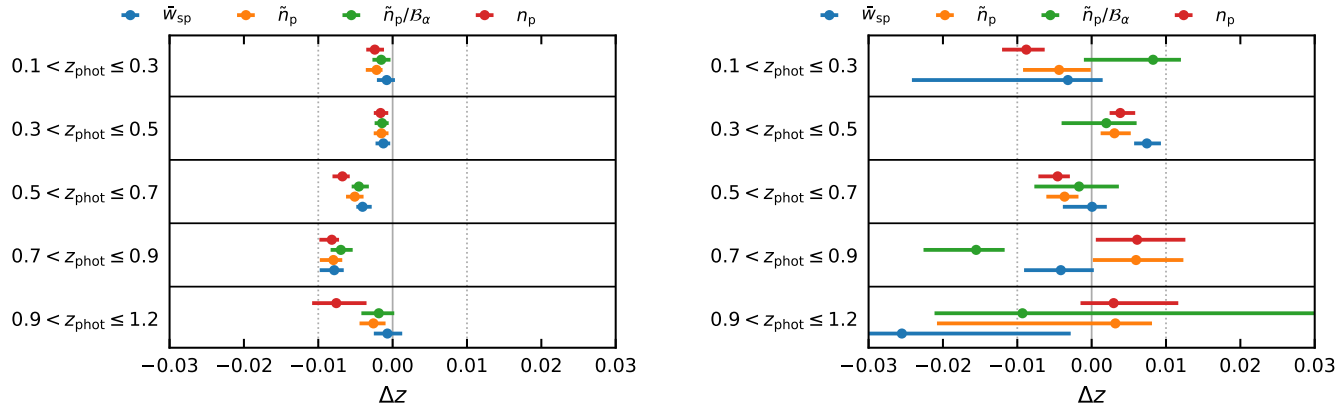

Fig. 12. Visualisation of the shift parameters Δz_i from fitting the true redshift distributions to the clustering redshifts obtained using the idealised (left side) and the realistic (right side) spectroscopic mock samples. The colours indicate different bias correction methods applied: the raw cross-correlation (blue), reference sample bias corrected (orange), additionally the target sample bias corrected using the SBM (green) and the target sample bias corrected using the sample autocorrelation function (red).

bias or a breakdown of the assumption of the cross-correlation formalism (see Sect. 3.1). Albeit, one has to keep in mind that, by design, this shift-fitting approach is mostly sensitive to the overall shape of the redshift distribution rather than individual outliers or noise fluctuations at the tails of the distribution.

In practical applications we do not have access to the true target sample redshift distributions. We therefore repeat the procedure with redshift distributions constructed using the direct calibration method (DIR, Lima et al. 2008; Hildebrandt et al. 2020a) as fit model (presented in Sect. 4.1.2). In short, these DIR redshift distributions are obtained by re-weighting a spectroscopic calibration sample such that it has the same properties in colour and magnitude space as the target sample (the KV450 cosmic shear sample) and are therefore fundamentally different from clustering redshifts. By using photometrically determined redshift estimates as a fit model, our approach can be considered similar to the DES redshift calibration, presented in Gatti et al. (2018). We however approach the problem from a different perspective: Instead of using the clustering redshifts to calibrate the DIR redshift distributions, we utilise the DIR to interpret the clustering redshifts. One can easily construct a variety of more sophisticated and hence more complex models than simply shifting an existing redshift estimate to interpret clustering redshifts. However, fitting these models requires a better understanding of the data covariance, in particular at high redshifts, which are dominated by a small number of deep spectroscopic fields.

4.1.1. Clustering-z accuracy given the true $n(z)$

The best-fit Δz and χ^2-values for using the true redshift distributions as fit model are summarised in Table 4 and visualised in Fig. 12. Almost all shift parameters are smaller than $|\Delta z| < 0.01$, the goal for the KiDS redshift calibration, indicating an insignificant overall bias of the recovered redshift distributions. The clustering redshifts obtained from the idealised reference sample suggest a slight tendency to underestimate the true redshifts, especially in the third and fourth tomographic bins. These shifts are likely related to the finite binning of the clustering redshifts with a constant comoving width of $\Delta\chi \approx 88$ Mpc, which results in a different number of sampling points that cover the peaks of each of the tomographic bins. Furthermore, the results are insensitive to bias corrections given the small evolution of the bias with redshift of the idealised and target samples (see Fig. 11). The clustering-z data points for $\tilde{n}_{\rm p}(z)$ and the best-fit model are shown in Fig. 13. The S/N is good enough to reveal details of the redshift distributions such as the outlier population in the tail of the third tomographic bin.

This is no longer true when using the realistic reference sample (Fig. 14). There is a smaller fraction of high redshift reference galaxies, considerably increasing the uncertainty on the redshift distributions at $z > 0.7$. Still, the overall shape of the clustering redshifts closely matches the true redshift distributions after correcting for the reference sample bias. In the highest tomographic bin, the shift parameter reduces significantly to be

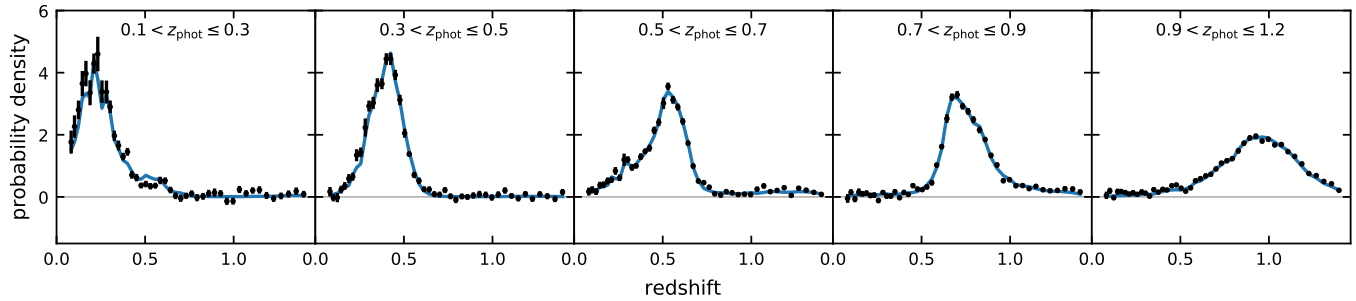

Fig. 13. Reference sample bias corrected clustering redshifts (black data points) fitted with the shifted true redshift distributions (blue, re-binned to the 45 data points) for all tomographic bins of the idealised mocks.

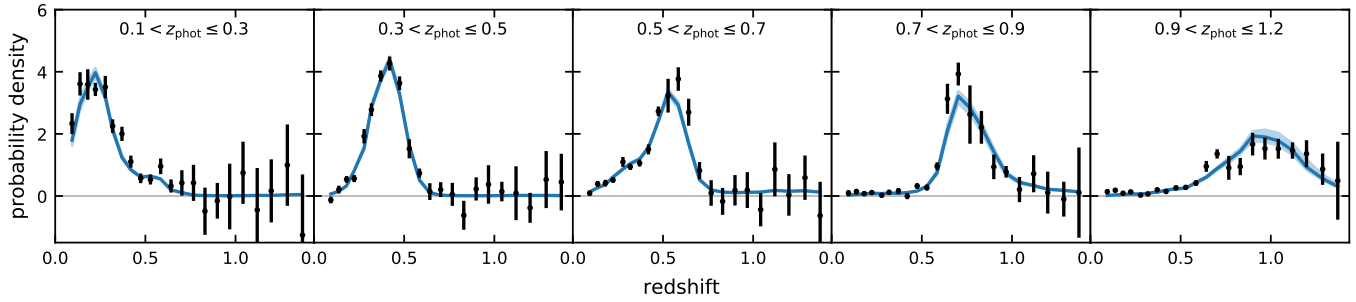

Fig. 14. Reference sample bias corrected clustering redshifts (black data points) fitted with the shifted DIR redshift distributions (blue, re-binned to the 22 data points) for all tomographic bins of the realistic mocks.

within $|\Delta z| < 0.01$ after removal of the reference sample bias via its auto-correlation function. As for the idealised case, addressing the target sample bias does not improve the results any further. To the contrary, the SBM bias correction increases the shifts and their uncertainty in most cases.

4.1.2. Clustering-z accuracy given the DIR $n(z)$

We repeat the analysis, but now use the DIR redshift distributions as fit model. This yields an estimate of the offset between the mean redshift of the best-fit models $\langle z_{\rm model}\rangle_i$ and the mean redshift of the true redshift distributions $\langle z_{\rm true}\rangle_i$. These offsets

$$\Delta z_i = \langle z_{\rm model}\rangle_i - \langle z_{\rm true}\rangle_i \approx \langle z_{\rm CC}\rangle_i - \langle z_{\rm true}\rangle_i \quad (27)$$

(presented in Table 5 and Fig. 15) are then our estimate of the shift of the mean clustering redshifts $\langle z_{\rm CC}\rangle_i$ with respect to the truth. We note that these DIR redshift distributions have been corrected for biases in the mean redshift (see Wright et al. 2020). Even if the models were not corrected and significantly biased instead, this procedure would still serve as a test to detect discrepancies between the model and the clustering redshifts since they would suffer from different systematic effects (see Hildebrandt et al. 2020b for a toy model).

The shift parameters obtained for the DIR and the idealised setup are within $|\Delta z| = 0.01$ for the 1st, 2nd, and 5th tomographic bin. This is not true for the third and fourth bin where the offset is $|\Delta z_{3,4}| \gtrsim 0.02$. We attribute this to the fact that the DIR redshift distributions are broadened compared to the true redshift distributions due to photometric noise (compare the blue lines in Figs. 14 and 16). This broadening reduces the skewness in these two DIR redshift distributions, which results in a bias when these more symmetric DIR $n(z)$ are fit to the correctly skewed clustering-z data points. The true redshift distributions in bins 1, 2, and 5 are less skewed to begin with and hence suffer less from this DIR-specific broadening/symmetrising. The broadening also affects the reduced χ^2-values for the joint fits, which reach up to 12, indicating a very poor fit between DIR and clustering redshift distributions.

The same poor goodness of fit can be observed when fitting the realistic mocks. There the reduced χ^2-values are of order 6, reflecting the significantly larger uncertainty at $z > 0.7$. Most shift parameter values are similar to the results from the idealised setup, but shifted by 0.01. This is probably driven by the strong degradation of the S/N with redshift, with the high signal-to-noise low-z data points driving the fit to the broadened DIR $n(z)$ high. The effect of the bias mitigation is comparable, both in magnitude and direction of shift, to fitting the clustering redshifts with the true redshift distributions.

We therefore conclude that the offsets inferred with the shift-fit method are predominantly dictated by the accuracy of (the shape of) the model redshift-distributions. In cases where the redshift distributions are well matched to the truth, shifts are minimal. Conversely when the distribution shape is discrepant from the truth, the shift method breaks down. The reduced χ^2 may be useful as diagnostic to detect such cases but should not be over-interpreted.

4.2. Scale dependence and magnification

The clustering-redshift formalism assumes linear, deterministic galaxy bias. Our fiducial scale of 100 kpc–1000 kpc is well within the non-linear biasing regime and therefore the relation in Eq. (17) is not guaranteed to hold on such small scales. We additionally measure clustering redshifts using the idealised reference mock sample on the slightly more conservative scales of 500 kpc–1500 kpc. Reducing the inner radius from 500 kpc to 100 kpc yields an approximate gain in S/N of about 50% and we need to test whether changes in the small scale clustering in combination with the improved sensitivity leads to systematic shifts in our analysis. For this comparison we employ clustering redshift estimates, which are

Table 5. Shift fit parameters for different bias correction methods for clustering redshifts obtained using the idealised and realistic spectroscopic mock samples using the DIR redshift distributions as fit model.

Setup	$n(z)$-type	$100 \times \Delta z_1$	$100 \times \Delta z_2$	$100 \times \Delta z_3$	$100 \times \Delta z_4$	$100 \times \Delta z_5$	χ^2	n_{dof}
Idealised	$\bar{w}_{\mathrm{sp}}$	$0.98^{+0.50}_{-0.52}$	$0.73^{+0.43}_{-0.42}$	$3.02^{+0.53}_{-0.56}$	$-2.06^{+0.42}_{-0.43}$	$0.25^{+0.45}_{-0.43}$	2069.5	215
	$\tilde{n}_{\mathrm{p}}$	$1.03^{+0.50}_{-0.52}$	$0.77^{+0.42}_{-0.43}$	$2.83^{+0.53}_{-0.52}$	$-2.21^{+0.42}_{-0.43}$	$-0.09^{+0.45}_{-0.46}$	2371.1	215
	$\tilde{n}_{\mathrm{p}}/\mathcal{B}_\alpha$	$1.18^{+0.53}_{-0.58}$	$0.95^{+0.44}_{-0.42}$	$2.85^{+0.52}_{-0.52}$	$-2.16^{+0.42}_{-0.43}$	$-0.09^{+0.45}_{-0.46}$	2123.4	215
	n_{p}	$1.18^{+0.52}_{-0.55}$	$0.77^{+0.42}_{-0.43}$	$2.49^{+0.53}_{-0.51}$	$-2.37^{+0.42}_{-0.43}$	$-0.66^{+0.47}_{-0.54}$	2587.4	215
Realistic	$\bar{w}_{\mathrm{sp}}$	$3.31^{+0.60}_{-0.88}$	$2.39^{+0.40}_{-0.40}$	$4.56^{+0.51}_{-0.52}$	$0.10^{+0.94}_{-0.49}$	$-1.37^{+0.89}_{-1.61}$	635.9	100
	$\tilde{n}_{\mathrm{p}}$	$2.24^{+0.67}_{-0.72}$	$1.88^{+0.47}_{-0.43}$	$3.37^{+0.50}_{-0.69}$	$0.97^{+0.44}_{-0.40}$	$1.31^{+0.84}_{-0.81}$	638.8	100
	$\tilde{n}_{\mathrm{p}}/\mathcal{B}_\alpha$	$4.46^{+0.97}_{-1.63}$	$2.37^{+0.59}_{-0.40}$	$4.46^{+2.66}_{-0.90}$	$-2.64^{+0.76}_{-0.82}$	$-0.77^{+2.03}_{-1.87}$	203.6	100
	n_{p}	$2.15^{+0.60}_{-0.68}$	$2.04^{+0.42}_{-0.41}$	$3.47^{+0.52}_{-0.61}$	$0.97^{+0.46}_{-0.40}$	$0.70^{+0.71}_{-0.83}$	675.3	100

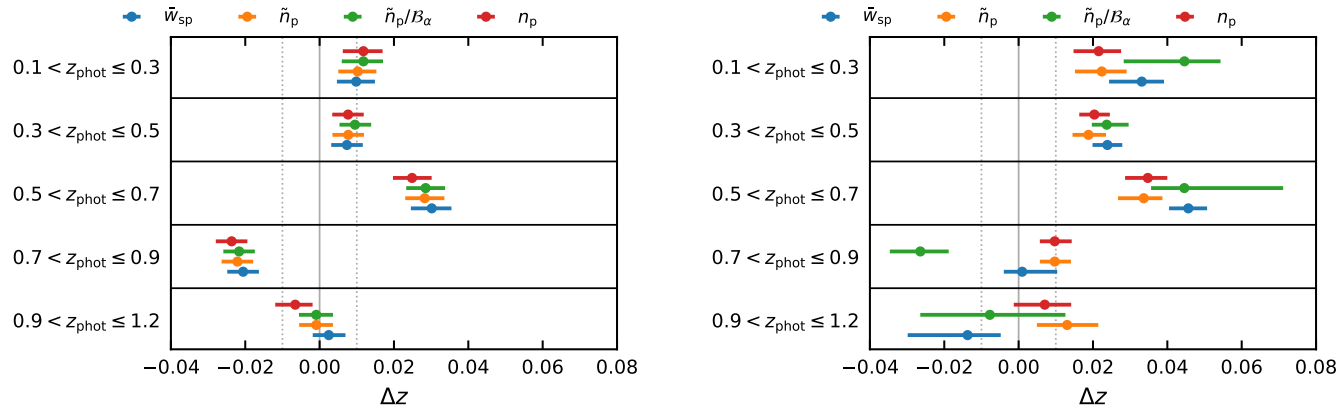

Fig. 15. Visualisation of the shift parameters Δz_i from fitting the DIR redshift distributions to the clustering redshifts obtained using the idealised (left side) and the realistic (right side) spectroscopic mock samples. The colours indicate different bias correction methods applied: the raw cross-correlation (blue), reference sample bias corrected (orange), additionally the target sample bias corrected using the SBM (green) and the target sample bias corrected using the sample autocorrelation function (red).

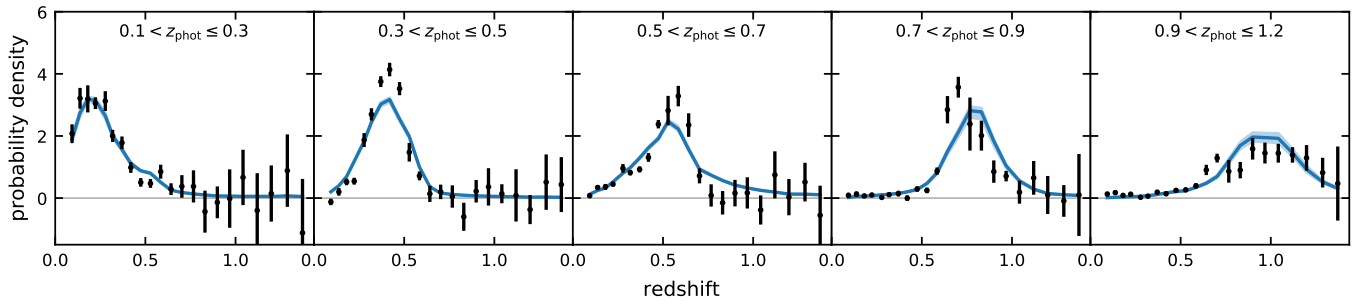

Fig. 16. Reference sample bias corrected clustering redshifts (black data points) fitted with the shifted DIR redshift distributions (blue, re-binned to the 22 data points) for all tomographic bins of the realistic mocks.

corrected for the reference sample bias and fit with the true redshift distributions, since these exhibit the smallest statistical uncertainties

We find that the best-fit parameters (Table 6) are in good agreement between both scales for the low redshift tomographic bins. For the two highest tomographic bins we detect a small positive shift when using scales of 500 kpc–1500 kpc. We note that these shifts are still smaller than the redshift binning of the cross-correlation measurements. The uncertainty of the Δz_i is roughly a factor two larger when using scales of 500 kpc–1500 kpc, which is due to the smaller clustering signal on large scales and the smaller logarithmic extent of the interval 500 kpc–1500 kpc compared to 100 kpc–1000 kpc. These larger uncertainties result in differences of the Δz_i between the two scales that are always insignificant at $\lesssim 2\sigma$. In summary, we find no strong indication that measuring on sub-Mpc scales biases the KiDS clustering redshifts significantly. For future surveys, however, this test should be repeated, as redshift calibration requirements will be more restrictive, and reference/target sample selections will change.

The cross-correlation between two galaxy samples arises not only from gravitational clustering, but also from correlations introduced by background sources that are lensed by foreground structures. This has two effects, first it changes the

Table 6. Shift fit parameters for two different correlation scales determined from the idealised mock data, once with lensing enabled and once disabled.

Scale (kpc)	Lensing	$100 \times \Delta z_1$	$100 \times \Delta z_2$	$100 \times \Delta z_3$	$100 \times \Delta z_4$	$100 \times \Delta z_5$	χ^2	$n_{\rm dof}$
100−1000	off	$-0.40^{+0.12}_{-0.12}$	$-0.18^{+0.06}_{-0.09}$	$-0.39^{+0.10}_{-0.10}$	$-0.60^{+0.13}_{-0.19}$	$0.14^{+0.23}_{-0.25}$	279.4	215
	on	$-0.22^{+0.08}_{-0.14}$	$-0.15^{+0.10}_{-0.10}$	$-0.51^{+0.12}_{-0.12}$	$-0.80^{+0.12}_{-0.18}$	$-0.26^{+0.16}_{-0.19}$	261.5	215
500−1500	off	$-0.58^{+0.17}_{-0.23}$	$0.10^{+0.12}_{-0.20}$	$-0.28^{+0.14}_{-0.10}$	$-0.02^{+0.21}_{-0.20}$	$0.89^{+0.40}_{-0.30}$	214.3	215
	on	$-0.24^{+0.29}_{-0.15}$	$0.14^{+0.14}_{-0.17}$	$-0.44^{+0.21}_{-0.12}$	$-0.20^{+0.32}_{-0.28}$	$0.29^{+0.24}_{-0.88}$	234.5	215

Notes. The fits are based on $\tilde{n}_{\rm p}(z)$ using the true redshift distributions as model.

effective survey area through a change in solid angle and secondly it increases the sample depths through magnification (e.g. Morrison et al. 2012; Choi et al. 2016). This additional correlation may become dominant at the tails of the clustering redshifts where the overlap between the reference and target samples is low (Gatti et al. 2018). Since the shift-fitting is mostly insensitive to changes in the tails of the distributions, we expect that magnification has little impact on our results.

To test this assertion we repeat the clustering redshift measurements and the model fitting on a version of our MICE2 mock catalogues in which we have switched off all lensing magnification effects. The corrections are measured using the true galaxy positions and the galaxy samples are selected from galaxy colours and photometric redshifts that are based on magnitudes with no flux magnification applied. The shift parameters (Table 6) from both samples agree within their respective uncertainties. The systematic shifts induced by lensing magnification in particular is not a concern for KiDS clustering redshifts since they are significantly smaller than the KiDS redshift calibration goal of $|\Delta z| < 0.01$.

5. Discussion

In the following we discuss the results from Sect. 4 and highlight some of the challenges that need to be overcome for clustering-z to become a fully complementary tool for redshift estimation.

The results we find in Sect. 4.1 are very encouraging. We are able to constrain the redshifts bias of the KV450 like mock galaxies within $|\Delta z_i| \leq 0.006$ when applying the shift-fit with the true redshift distributions. These figures may even improve with upcoming data releases which allow us to utilise more overlap of KiDS with SDSS and 2dFLenS. However, the aforementioned figures are sensitive to systematic features in the redshift distributions that serve as a model for the shift-fit. When we fit the clustering redshift data points with the DIR redshift distributions, we see shifts of up to $|\Delta z_i| \lesssim 0.04$. These shifts are enhanced by the broadening of the DIR $n(z)$, which is induced by photometric noise. This broadening in combination with the redshift-dependent S/N of the clustering-z based on the realistic mock data yields a significant bias. In future KiDS analyses we will utilise SOM redshift distributions (Wright et al. 2020) for our shift fit analyses. These redshift distributions are more robust than their DIR counterparts, demonstrating reduced photometric broadening and overall bias. Using the SOM redshift distributions as a fit model, we expect, will therefore improve the modelling of clustering redshift estimates.

We find qualitatively different behaviour between our recovered Δz_i estimates when calibrating with idealised and realistic mock reference samples. We hypothesise that this is driven by the two following effects. First, the bias evolution of the idealised reference sample is small since its amplitude varies only by approximately 25% over the full redshift baseline. Secondly, the bias evolution has very little impact on the redshift distribution if it is sufficiently narrow. If the bias changes by 25% over the full redshift range this also means that it only changes by ∼5% over each of the tomographic bins, an effect that is lost in the noise. Even if there is an outlier population of galaxies e.g. at high redshift, correcting for the evolving galaxy bias may well change the true mean redshift significantly, but the shift-fitting is mostly unaffected, since the model is not flexible enough to account for high-z outliers. This explains why the idealised mock setup is stable against changes of the sample bias. The realistic mock, however, has more complex clustering properties since it utilises a mixture of different reference samples. Thus, the spectroscopic bias correction has a significant impact, as can be seen by comparing the blue and orange data points in the right-hand panel of Fig. 12.

We therefore conclude that it is sufficient, at the sensitivity of KV450, to solely correct the reference sample bias in clustering-z estimates. This assessment, however, is dependent on the clustering properties of the target sample. As a result, this conclusion will need to be revisited in future KiDS-like analyses that utilise different source-sample selections (such as those that may be induced by additional colour-based selections; Wright et al. 2020). The same applies to stage-IV surveys. The challenging redshift calibration requirements of these programs will likely demand a careful treatment of the target sample bias evolution.

In the future these results can be improved upon by optimising the combination of the results from different reference samples. The KiDS footprint and the calibration fields overlap with a rich set of spectroscopic surveys. This allows us to calibrate redshifts to $z = 1$ and beyond, but adds the additional challenge of combining independent measurements into an unbiased redshift estimate. We solve this problem by employing a bootstrap resampling combination method (see Appendix B). The main issue with this approach is that it violates a basic assumption of bootstrap resampling: the individual spatial regions (defined by KiDS pointings and the deep spectroscopic pointings) are not statistically equivalent. Each of the spectroscopic reference samples has a different density, redshift distribution, and clustering properties. Thus, the number of surveys (and hence also pointings) that contribute to the combined cross-correlation amplitude vary for the 22 redshift bins (in case of the realistic mock setup). This is problematic especially if the clustering of the reference galaxies varies over a short redshift interval, such as in the transition region between the wide surveys (which dominate the $z \lesssim 0.7$ regime) and the deep spectroscopic surveys (which dominate the high redshifts). Figure 14 shows a systematic feature at exactly this redshift that biases the resulting clustering redshifts. Fortunately, this is not reflected in the shift parameters since the fit is only sensitive to the overall shape of the redshift distributions due to the very restricted model we are fitting.

Stage-IV surveys, such as *Euclid* or the *Vera C. Rubin* Observatory Legacy Survey of Space and Time (LSST), will be in a similar situation as KiDS since they likewise overlap with a number of (mostly, but not necessarily) spectroscopic reference samples that have distinct properties. Therefore, additional efforts are vital in order to make clustering redshifts a competitive and complementary method to meet the strict requirements of these projects. This can be achieved by either exploring other combination methods or by optimising the reference samples such that they bridge the gap between low-redshift, wide area surveys and high-redshift surveys with low area coverage.

Another fundamental issue for the clustering-z method is correcting for the galaxy bias evolution of the target sample. The bias evolution estimates we obtain from the SBM do not agree with the results from directly fitting the bias model to the auto-correlation terms. By design the SBM picks up any other redshift-dependent systematic error that skews the full source sample in a different way than the weighted sum of the individually normalised tomographic bin measurements. On the other hand, recovering the bias evolution requires a sufficiently accurate bias model. Furthermore, individual tomographic bins can have a slightly different bias evolution due to their selection criteria. Hence, a perfect agreement between both methods is probably too much to expect.

Regardless of the underlying cause for the disagreement, the left-hand panel of Fig. 12 shows that this disagreement in the galaxy bias evolution estimates for the target sample has a vanishing impact on the recovered redshift bias parameters Δz_i. This can be explained by the fact that the bias of the KiDS mock sample evolves very little with redshift. Even a somewhat inaccurate correction does not have a strong influence on the end result. The issue becomes more evident when applied to the realistic mock clustering redshifts (right-hand panel of Fig. 12). Due to the low reference sample galaxy densities at high redshifts, the S/N drops significantly at $z > 0.7$. The highest tomographic redshift bin, which contributes essential information to determine the bias evolution via the SBM, is not well constrained by the clustering redshifts. Consequently, the uncertainty of the bias corrected redshifts and the shift parameters Δz_i is greatly increased, dominating the total error budget.

The only solution to this problem seems to be a more comprehensive spectroscopic calibration sample at high redshift that can reliably probe the core of the redshift distributions of the highest-redshift tomographic bins used in cosmic shear studies.

6. Summary and outlook

In this paper we detail the creation of mock catalogues based on the MICE2 simulation that closely resemble the KiDS-VIKING-450 (KV450) dataset and its overlapping spectroscopic calibration samples. We use this mock data to replicate the clustering redshift estimates for KV450 and estimate their accuracy in direct comparison to the true redshifts in the simulation. The main result is that clustering-z with KV450-like quality can reliably calibrate residual biases in the redshift distribution of typical galaxy samples used in cosmic shear measurements if the shape of the redshift distribution is a priori well known. After correcting for evolving galaxy bias of a realistic spectroscopic reference sample via a measurement of its auto-correlation function, the clustering-z recover the mean redshifts of all five tomographic bins at better than $|\Delta z_i| < 0.006$. Without this correction, the highest-z tomographic bin shows a bias of $\Delta z_5 \approx 0.026$, underlining the importance of the bias modelling of the spectroscopic sample.

Further correcting for the evolving galaxy bias of the target sample, constrained by comparing a weighted sum of the $n(z)$ of all five tomographic bins (individually normalised) to the full source sample, does not lead to a further reduction in the biases. This indicates that the very mild bias evolution of the KiDS source galaxies does not need to be corrected at this level of precision.

Next we used redshift distributions estimated from multi-colour photometry by re-weighting a deep spectroscopic calibration sample (determined using the direct calibration method of the form presented by Lima et al. 2008) to constrain the high redshift tails of the clustering redshifts and to interpret them as probability distributions. Even with an idealised reference sample for the cross-correlation measurements, these noisy DIR redshift distributions are not fully able to model the clustering redshifts, due to a systematic shape mismatch between both distributions. This is exacerbated when using a more realistic reference sample, yielding biases of up to $|\Delta z_i| \approx 0.04$. The difference, seen between our results when using the true and the DIR redshift distributions as our fit model, demonstrates an important conclusion for clustering redshift calibration. When performing the shift-fitting with a model that is the same shape as the true redshift distribution, the resulting best-fit solution is a good representation of the truth, regardless of the model bias itself. Therefore, these biases may be reduced by fitting redshift distributions estimated via the less noisy and less biased SOM redshift distributions of Wright et al. (2020), or alternatively by increasing the amount of information extracted from the clustering redshifts. A possible approach would be to use fit-models that are more flexible than a fixed redshift distribution with a single free parameter. Such models must be constrained to positive amplitudes, as for example the Gaussian mixtures that we employed in Hildebrandt et al. (2020a). However, fitting more complex models requires a better understanding of the covariances of clustering redshifts, especially when derived from a combination of different reference samples. In the long run, exploiting the full potential of synergies between clustering- and photometry-based redshift estimates (e.g. Sánchez & Bernstein 2019; Alarcon et al. 2020) seems to be the most promising strategy to meet the stringent redshift requirements of upcoming stage IV survey missions.

Acknowledgements. We acknowledge support from the European Research Council under grant numbers 770935 (J. vd B., H. H., A. H. W.). H. H. is also supported by a Heisenberg Grant (Hi1495/5-1) of the Deutsche Forschungsgemeinschaft. 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. C. B. M. acknowledges support from the DIRAC Institute in the Department of Astronomy at the University of Washington. The DIRAC Institute is supported through generous gifts from the Charles and Lisa Simonyi Fund for Arts and Sciences, and the Washington Research Foundation. C. H. acknowledges support from the European Research Council under Grant number 647112, and support from the Max Planck Society and the Alexander von Humboldt Foundation in the framework of the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. K. K. acknowledges support by the Alexander von Humboldt Foundation. We are grateful to the zCOSMOS team to give us early access to additional deep spectroscopic redshifts that were not available in the public domain. This work is based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 100.A-0613, 102.A-0047, 179.A-2004, 177.A-3016, 177.A-3017, 177.A-3018, 298.A-5015, and on data products produced by 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. The GAMA website is http://www.gama-survey.org/. The MICE simulations have been developed at the MareNostrum supercomputer (BSC-CNS) thanks to grants AECT-2006-2-0011 through AECT-2015-1-0013. This work has made use of CosmoHub (Carretero et al. 2017). CosmoHub has been developed by the Port d'Informació Científica (PIC), maintained through a collaboration of the Institut de Física d'Altes Energies (IFAE) and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), and was partially funded by the "Plan Estatal de Investigación Científica y Técnica y de Innovación" program of the Spanish government. Author contributions: all authors contributed to the development and writing of this paper. The authorship list is given in three groups: the lead authors (J. L. vd B., H. H., A. H. W., C. B. M.), 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., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
Alarcon, A., Sánchez, C., Bernstein, G. M., & Gaztañaga, E. 2020, MNRAS, 498, 2614
Bartelmann, M., & Schneider, P. 2001, Phys. Rep., 340, 291
Benítez, N. 2000, ApJ, 536, 571
Blake, C., Amon, A., Childress, M., et al. 2016, MNRAS, 462, 4240
Bonnett, C., Troxel, M. A., Hartley, W., et al. 2016, Phys. Rev. D, 94, 042005
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Buchs, R., Davis, C., Gruen, D., et al. 2019, MNRAS, 489, 820
Capak, P. L. 2004, PhD Thesis, UNIVERSITY OF HAWAI'I
Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., & Fosalba, P. 2015, MNRAS, 447, 646
Carretero, J., Tallada, P., Casals, J., et al. 2017, PoS, EPS-HEP2017, 488
Choi, A., Heymans, C., Blake, C., et al. 2016, MNRAS, 463, 3737
Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., & Carretero, J. 2015, MNRAS, 453, 1513
Davis, M., & Peebles, P. J. E. 1983, ApJ, 267, 465
Davis, C., Rozo, E., Roodman, A., et al. 2018, MNRAS, 477, 2196
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
de Jong, J. T. A., Kleijn, G. A. V., Erben, T., et al. 2017, A&A, 604, A134
DeRose, J., Wechsler, R. H., Becker, M. R., et al. 2019, ArXiv e-prints [arXiv:1901.02401]
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
Edge, A., Sutherland, W., Kuijken, K., et al. 2013, Messenger, 154, 32
Fenech Conti, I., Herbonnet, R., Hoekstra, H., et al. 2017, MNRAS, 467, 1627
Flaugher, B., Diehl, H. T., Honscheid, K., et al. 2015, AJ, 150, 150
Fosalba, P., Crocce, M., Gaztañaga, E., & Castander, F. J. 2015a, MNRAS, 448, 2987
Fosalba, P., Gaztañaga, E., Castander, F. J., & Crocce, M. 2015b, MNRAS, 447, 1319
Gatti, M., Vielzeuf, P., Davis, C., et al. 2018, MNRAS, 477, 1664
Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012, MNRAS, 427, 146
Hikage, C., Oguri, M., Hamana, T., et al. 2019, PASJ, 71, 43
Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454
Hildebrandt, H., Köhlinger, F., van den Busch, J. L., et al. 2020a, A&A, 633, A69
Hildebrandt, H., van den Busch, J. L., Wright, A. H., et al. 2020b, A&A, submitted [arXiv:2007.15635]
Hoffmann, K., Bel, J., Gaztañaga, E., et al. 2015, MNRAS, 447, 1724
Hoyle, B., Gruen, D., Bernstein, G. M., et al. 2018, MNRAS, 478, 592
Johnson, A., Blake, C., Amon, A., et al. 2017, MNRAS, 465, 4118
Joudaki, S., Hildebrandt, H., Traykova, D., et al. 2020, A&A, 638, L1
Kannawadi, A., Hoekstra, H., Miller, L., et al. 2019, A&A, 624, A92
Kuijken, K. 2008, A&A, 482, 1053
Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, MNRAS, 454, 3500
Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845
Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, A&A, 559, A14
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
Mandelbaum, R. 2018, ARA&A, 56, 393
Maraston, C., Pforr, J., Henriques, B. M., et al. 2013, MNRAS, 435, 2764
Masters, D. C., Capak, P. L., Stern, D., et al. 2016, in American Astronomical Society Meeting Abstracts #227, Am. Astron. Soc. Meeting Abstr., 227, 139.14
Matthews, D. J., & Newman, J. A. 2010, ApJ, 721, 456
McQuinn, M., & White, M. 2013, MNRAS, 433, 2857
Ménard, B., Scranton, R., Schmidt, S., et al. 2013, ArXiv e-prints [arXiv:1303.4722]
Miller, L., Kitching, T. D., Heymans, C., Heavens, A. F., & van Waerbeke, L. 2007, MNRAS, 382, 315
Miller, L., Heymans, C., Kitching, T. D., et al. 2013, MNRAS, 429, 2858
Morrison, C. B., Scranton, R., Ménard, B., et al. 2012, MNRAS, 426, 2489
Morrison, C. B., Hildebrandt, H., Schmidt, S. J., et al. 2017, MNRAS, 467, 3576
Newman, J. A. 2008, ApJ, 684, 88
Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5
Raichoor, A., Mei, S., Erben, T., et al. 2014, in SF2A-2014: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, eds. J. Ballet, F. Martins, F. Bournaud, R. Monier, C. Reylé, 359
Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nat. Astron., 3, 212
Sánchez, C., & Bernstein, G. M. 2019, MNRAS, 483, 2801
Schmidt, S. J., Ménard, B., Scranton, R., Morrison, C., & McBride, C. K. 2013, MNRAS, 431, 3307
Schneider, D. P., Richards, G. T., Hall, P. B., et al. 2010, AJ, 139, 2360
Scottez, V., Benoit-Lévy, A., Coupon, J., Ilbert, O., & Mellier, Y. 2018, MNRAS, 474, 3921
Scranton, R., Johnston, D., Dodelson, S., et al. 2002, ApJ, 579, 48
Sérsic, J. L. 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41
Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, PASJ, 70, S9
Troxel, M. A., MacCrann, N., Zuntz, J., et al. 2018, Phys. Rev. D, 98, 043528
Venemans, B. P., Verdoes Kleijn, G. A., Mwebaze, J., et al. 2015, MNRAS, 453, 2259
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Wright, A. H., Hildebrandt, H., Kuijken, K., et al. 2019, A&A, 632, A34
Wright, A. H., Hildebrandt, H., van den Busch, J. L., & Heymans, C. 2020, A&A, 637, A100
York, D. G., Adelman, J., Anderson, J. E. Jr., et al. 2000, AJ, 120, 1579

Appendix A: Spectroscopic Mock Sample Selection

Here we present the spectroscopic selection functions and their modifications for selecting these samples on MICE2 as described in Sect. 2.2: SDSS in presented in Table A.1, 2dFLenS in Table A.2, WiggleZ in Table A.3 and DEEP2 in Table A.4. Furthermore we show the remaining magnitude-, colour- and colour-colour plots comparing the spectroscopic data to the mock samples of VVDS-02h (Fig. A.1) and zCOSMOS (Fig. A.2).

Table A.1. Summary of the selection functions applied in MICE2 compared to the literature selection functions, for the SDSS Main galaxy sample, the BOSS CMASS and LOWZ samples, and the SDSS QSO sample.

Sub-sample	SDSS selection	MICE2 object selection	Comments
Main	$r_{\rm pet} < 17.77$	$r < 17.7$	
LOWZ	$16.0 < r < 19.6$	$16.0 < r < \mathbf{20.0}$	
	$\lvert c_\perp \rvert < 0.2$	$\lvert c_\perp \rvert < 0.2$	
	$r < 13.5 + c_\parallel/0.3$	$r < \mathbf{13.35} + c_\parallel/0.3$	
CMASS	$17.5 < i < 19.9$	$17.5 < i < \mathbf{20.1}$	
	$d_\perp > 0.55$	$d_\perp > 0.55$	
	$i < 19.86 + 1.6(d_\perp - 0.8)$	$i < \mathbf{19.98} + 1.6(d_\perp - \mathbf{0.7})$	
	$r - i < 2.0$	$r - i < 2.0$	
QSO	—	`flag_central` == 1	Substitute selection to compensate that MICE2 does not contain quasars.
	—	$\log_{10}(M_{\rm halo}) > 13.3$	
	—	$\log_{10}(M_\star) > 11.2$	

Notes. All selections here invoke "and" logic: rule$_1$&rule$_2$& etc. A long dash (—) indicates a selection which cannot be applied to MICE2. Deliberate adjustments that yield a better match in the simulated and real redshift distributions are highlighted in bold-face.

Table A.2. Summary of the selection functions applied in MICE2 compared to the literature selection functions for the 2dFLenS sample.

Sub-sample	2dFLenS selection	MICE2 object selection	Comments
low-z, Cut I	$16.0 < r < 19.2$	$\mathbf{16.5} < r < 19.2$	
	$r < 13.1 + c_\parallel/0.3$	$r < 13.1 + c_\parallel/\mathbf{0.32}$	
	$\lvert c_\perp \rvert < 0.2$	$\lvert c_\perp \rvert < 0.2$	
low-z, Cut II	$16.0 < r < 19.5$	$\mathbf{16.5} < r < 19.5$	
	$\lvert c_\perp \rvert > 0.45 - (g - r)/6$	$\lvert c_\perp \rvert > 0.45 - (g - r)/6$	
	$g - r > 1.3 + 0.25(r - i)$	$g - r > 1.3 + 0.25(r - i)$	
low-z, Cut III	$16.0 < r < 19.6$	$\mathbf{16.5} < r < 19.6$	
	$r < 13.5 + c_\parallel/0.3$	$r < 13.5 + c_\parallel/\mathbf{0.32}$	
	$\lvert c_\perp \rvert < 0.2$	$\lvert c_\perp \rvert < 0.2$	
mid-z	$17.5 < i < 19.9$	$17.5 < i < 19.9$	
	$r - i < 2.0$	$r - i < 2.0$	
	$d_\perp > 0.55$	$d_\perp > 0.55$	
	$i < 19.86 + 1.6(d_\perp - 0.8)$	$i < 19.86 + 1.6(d_\perp - \mathbf{0.9})$	
high-z	$r - W1 < 2(r - i)$	$r - \mathbf{K_s} > \mathbf{1.9}(r - i)$	Used $r - K_s$ as substitute for missing $r - W1$ in MICE2.
	$r - i > 0.98$	$r - i > 0.98$	
	$i - Z > 0.6$	$i - Z > 0.6$	
	$19.9 < i < 21.8$	$19.9 < i < 21.8$	
	$z < 19.95$	$z < \mathbf{19.9}$	

Notes. All selections here invoke "and" logic: rule$_1$&rule$_2$& etc.Deliberate adjustments that yield a better match in the simulated and real redshift distributions are highlighted in bold-face.

Table A.3. Summary of the selection functions applied in MICE2 compared to the literature selection functions for the WiggleZ sample.

Selection type	WiggleZ selection	MICE2 object selection	Comments
Exclusion	$g < 22.5$	$g < 22.5$	
	$i < 21.5$	$i < 21.5$	
	$r - i < g - r - 0.1$	$r - i < g - r - 0.1$	
	$r - i < 0.4$	$r - i < 0.4$	
	$g - r > 0.6$	$g - r > 0.6$	
	$r - z < 0.7(g - r)$	$r - z < 0.7(g - r)$	
Inclusion	NUV < 22.8	—	UV selection is mimicked by weighted sampling to match the $n(z)$'s.
	$20.0 < r < 22.5$	$20.0 < r < 22.5$	
	FUV − NUV > 1 or no FUV	—	
	$-0.5 <$ NUV $- r < 2.0$	—	
	$S/N_{\rm NUV} > 3.0$	—	
	Match within 2.5″	—	

Notes. All selections here invoke "and" logic: rule$_1$&rule$_2$& etc. Deliberate adjustments that yield a better match in the simulated and real redshift distributions are highlighted in bold-face. A long dash (—) indicates a selection which cannot be applied to MICE2.

Table A.4. Summary of the selection functions applied in MICE2 compared to the literature selection functions for the DEEP2 sample.

Selection type	DEEP2 selection	MICE2 object selection	Comments
Magnitude	$18.5 < R < 24.0$	$18.5 < R < 24.0$	
Colour	$B - R < 2.45\,(R - I) - 0.2976$	$B - R < \mathbf{2.0}\,(R - I) - \mathbf{0.4}$	Compensate noiseless model magnitudes and template differences, see Fig. 7.
	$R - I > 1.1$	$R - I > 1.1$	
	$B - R < 0.5$	$B - R < \mathbf{0.2}$	

Notes. The magnitude and colour selections are linked by "and" logic, whereas the individual colour cuts invoke "or" logic. Deliberate adjustments that yield a better match in the simulated and real redshift distributions are highlighted in bold-face.

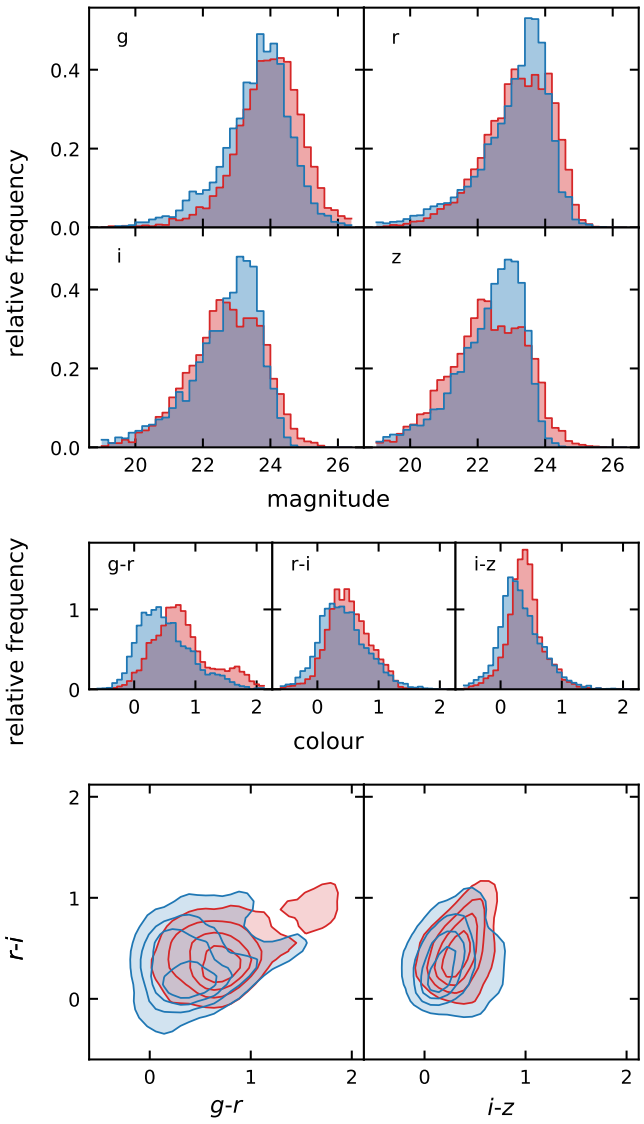

Fig. A.1. Comparison of colour and magnitude distributions for the simulated (blue) and observed (red) VVDS-02h dataset.

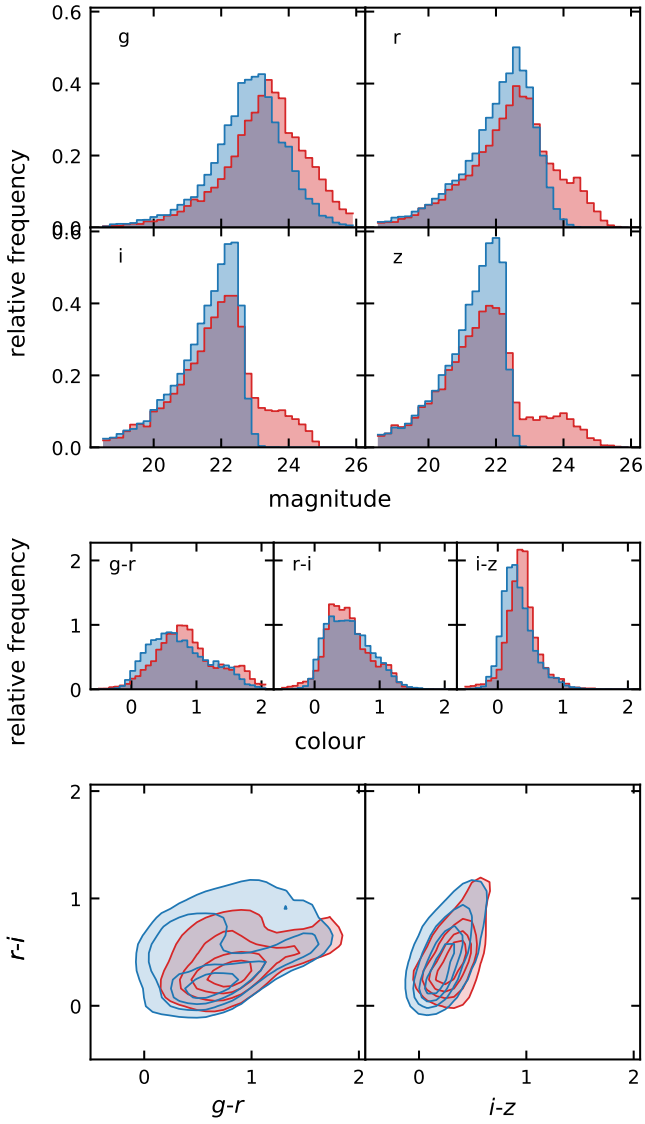

Fig. A.2. Same as Fig. A.1 but for the zCOSMOS dataset. The missing tails in the *i*- and *Z*-band originate from the zCOSMOS galaxies with $z > 1.4$ which do not exist in MICE2.

Appendix B: Clustering-redshift pipeline

The `yet_another_wizz` package is an end-to-end pipeline for clustering redshift estimation. It not only allows the user to compute clustering redshifts, but also takes care of the bias mitigation. In the following we summarise our pipeline, step-by-step, from the input data to the final, bias corrected clustering redshifts.

1. *Spatial regions.* We estimate the uncertainties and covariance of our clustering redshifts using bootstrap re-sampling. Therefore, we split the input data catalogues and the spectroscopic randoms into spatial regions. For the wide area spectroscopic fields the most convenient choice in KV450 was to divide the data based on the KiDS VST pointings. For the mock data used here we mimic the pointings by creating a 20×22 grid with each cell covering $0.7\,\mathrm{deg}^2$, the mean, unmasked area of a KiDS pointing.
   Due to their small area the deep spectroscopic fields were treated as one region each, no matter how many VST pointings were required to cover the spectroscopic footprints. As a result, there were four spatial regions originating from the deep fields, since the two DEEP2 fields (02h/23h) were considered separate entities. The mock catalogues replicate this behaviour but instead of two smaller DEEP2 mock catalogues we use one larger contiguous catalogue.
2. *Random generation.* We generate a uniform mock-KV450 random catalogue for each spatial region independently, effectively scaling the random density based on the mean density of each pointing. We adopt the same strategy that we use for processing the KiDS data, where this local density estimation is designed to mitigate observational density variations in the data. Since we measure correlations for each spatial region independently (see next paragraph), we are not concerned that this strategy destroys large-scale correlation modes. Finally, we clone the photometric and weight distributions for the random catalogues by randomly sampling from their distributions in the mock data catalogue. Similarly, we generate spectroscopic random catalogues, cloning the spectroscopic redshift.
3. *Cross-correlation.* We measure the cross-correlation between the KV450 mock data and each of the mock spectroscopic samples within an annulus of 100 kpc–1000 kpc[10]. Limiting ourselves to such small scales is necessary due to the small size of the deep spectroscopic fields. The redshift resolution of our measurement is given by the binning of the spectroscopic data: for the idealised mock setup we use 45 comoving bins and the realistic mock setup 22 comoving bins between $0.01 \leq z_\mathrm{spec} < 1.42$. We measure the cross-correlations for the full sample and also for each tomographic bin, weighting the galaxy pairs by the KiDS *lensfit*-weight.
4. *Reference sample bias.* We determine the reference sample bias evolution for each cross-correlation measurement by measuring the sample auto-correlation function using the same constant comoving binning and the same physical scales. With this proxy for the bias evolution we correct the cross-correlation measurements according to Eq. (20).
5. *Combining measurements.* So far we have independent estimates for the KiDS redshift distribution for each of the spatial regions defined above that overlap with the footprints of the spectroscopic mock samples (see Fig. 3). The contribution of each measurement to the combined redshift distribution varies with redshift, depending on the reference sample redshift distribution. We apply a spatial bootstrapping approach to merge the cross- and auto-correlation measurements into a single redshift distribution estimate $\tilde{n}_\mathrm{p}(z)$. First, we create create a pool of pair counts DD and DR from each spatial region for each redshift bin z_j. Then we sum the pair counts from all N_reg regions and re-compute the correlation estimator (from Eq. (25))
$$\bar{w}(z_j) = \frac{\sum_n^{N_\mathrm{reg}} \mathrm{DD}_n(z_j)}{\sum_n^{N_\mathrm{reg}} \mathrm{DR}_n(z_j)} - 1 \tag{B.1}$$
   for the cross- and auto-correlation functions and calculate $\tilde{n}_\mathrm{p}(z_j)$ from Eq. (20).
   This combination method violates a basic assumption of classical bootstrapping, since the sub-samples (spatial regions) are not statistically equivalent. Each of our spectroscopic samples has a different density, redshift distribution and biasing which can potentially bias the combined clustering redshifts.
6. *Target sample bias.* Since we can assume that the bias evolution of the KiDS mock galaxies is the same everywhere, except for sample variance, we apply the correction after combining the measurements. We use the self-consistency bias mitigation described in Sect. 3.3 and assume that the sample bias is approximately given by Eq. (21). We constrain the model parameter α using Eq. (23) by comparing the redshift distribution estimate of step 5 for the full KiDS mock sample and the weighted sum of the tomographic bins. The weight of a bin is given by the sum of the *lensfit*-weights of all galaxies in that bin, divided by the sum of the weights of all galaxies between $0.1 < z_\mathrm{phot} \leq 1.2$.
   We also explore one alternative way of correcting the target sample bias. Analogous to step 4 we compute the sample auto-correlation function using the true redshifts of the mock galaxies and combine the measurements with the bootstrap method from step 5. According to Eq. (17), this should give the most accurate clustering redshifts and we use it to validate the bias fitting approach. Certainly, such an approach is not possible on real data.
7. *Covariance Estimation.* To estimate uncertainties and a covariance matrix for the clustering redshifts we apply bootstrap resampling based on the spatial regions. We implement this in the same way as the survey combination in step 5, but instead of summing all regions together, we randomly draw with replacement from the pool of spatial regions to generate samples. We propagate the bias mitigation to these samples which allows us to compute uncertainties from the standard error and a covariance matrix.

There are two key differences between `yet_another_wizz` and `the-wizz` which we used previously in Hildebrandt et al. (2020a). `yet_another_wizz` is no longer built around the pixelation library `STOMP`. Furthermore, we removed one of the most distinct features of `the-wizz`, which is its ability to create look-up-tables with galaxy indices for each pairs. This allows to compute clustering redshifts for arbitrary sub-samples of the photometric galaxies at any later time by querying the look-up table before summing the pair counts and computing the correlation estimator. Creating the look-up table is very time-consuming and our datasets are not big enough to harvest the full potential of this approach. Instead, `yet_another_wizz` stores just the pair counts for each reference source, giving the user the freedom to change the redshift binning of the correlation measurements at a later stage.

[10] We note that Hildebrandt et al. (2017) used significantly smaller scales of 30 kpc $\leq r <$ 300 kpc to enable a clustering-z measurement from the deep fields alone.