A&A 647, A124 (2021)
https://doi.org/10.1051/0004-6361/202039018

Astronomy & Astrophysics

KiDS-1000 catalogue: Redshift distributions and their calibration

H. Hildebrandt[1], J. L. van den Busch[1,2], A. H. Wright[1], C. Blake[3], B. Joachimi[4], K. Kuijken[5], T. Tröster[6], M. Asgari[6], M. Bilicki[7], J. T. A. de Jong[8], A. Dvornik[1], T. Erben[2], F. Getman[9], B. Giblin[6], C. Heymans[6,1], A. Kannawadi[10], C.-A. Lin[6], and H.-Y. Shan (陕欢源)[11,12]

[1] Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing, 44780 Bochum, Germany
e-mail: hendrik@astro.ruhr-uni-bochum.de
[2] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[3] Centre for Astrophysics & Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia
[4] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
[5] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
[6] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[7] Center for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland
[8] Kapteyn Astronomical Institute, University of Groningen, 9700 AD Groningen, The Netherlands
[9] INAF – Astronomical Observatory of Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[10] Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA
[11] Shanghai Astronomical Observatory (SHAO), Nandan Road 80, Shanghai 200030, PR China
[12] University of Chinese Academy of Sciences, Beijing 100049, PR China

Received 24 July 2020 / Accepted 22 January 2021

ABSTRACT

We present redshift distribution estimates of galaxies selected from the fourth data release of the Kilo-Degree Survey over an area of ∼1000 deg^2 (KiDS-1000). These redshift distributions represent one of the crucial ingredients for weak gravitational lensing measurements with the KiDS-1000 data. The primary estimate is based on deep spectroscopic reference catalogues that are re-weighted with the help of a self-organising map (SOM) to closely resemble the KiDS-1000 sources, split into five tomographic redshift bins in the photometric redshift range $0.1 < z_B \leq 1.2$. Sources are selected such that they only occupy that volume of nine-dimensional magnitude-space that is also covered by the reference samples ('gold' selection). Residual biases in the mean redshifts determined from this calibration are estimated from mock catalogues to be $\lesssim$0.01 for all five bins with uncertainties of ∼0.01. This primary SOM estimate of the KiDS-1000 redshift distributions is complemented with an independent clustering redshift approach. After validation of the clustering-z on the same mock catalogues and a careful assessment of systematic errors, we find no significant bias of the SOM redshift distributions with respect to the clustering-z measurements. The SOM redshift distributions re-calibrated by the clustering-z represent an alternative calibration of the redshift distributions with only slightly larger uncertainties in the mean redshifts of ∼0.01−0.02 to be used in KiDS-1000 cosmological weak lensing analyses. As this includes the SOM uncertainty, clustering-z are shown to be fully competitive on KiDS-1000 data.

Key words. cosmology: observations – gravitational lensing: weak – galaxies: photometry – surveys

1. Introduction

One of the most important goals of observational astronomy has always been to add a third dimension to the two-dimensional images of the sky. In modern extra-galactic imaging surveys containing tens of millions of galaxies this information is obtained with the technique of photometric redshifts (photo-z; see Salvato et al. 2019, for a recent review). The cosmological redshift leads to a reddening of galaxy spectra that can be detected by observing a galaxy in different photometric passbands. This can yield approximate redshifts at a much higher efficiency and to fainter magnitudes than any spectroscopic technique, albeit at significantly reduced precision.

Measurements of weak gravitational lensing (WL; see e.g., Bartelmann & Schneider 2001) crucially depend on photo-z to estimate the geometric factors that enter the modelling of this effect. The accuracy of the cosmological conclusions drawn from modern WL surveys depends directly on the accuracy of the photo-z used to model the WL observables (Huterer et al. 2006). In this process, it is useful to distinguish two different regimes where multi-band photometric information is typically used (Newman et al. 2015). First, approximate individual redshifts for all galaxies used in a WL measurement are estimated to bin the galaxies along the redshift axis. Secondly, the same photometric information is used to estimate the redshift distributions of the ensembles of galaxies in these so-called tomographic bins. While high precision is desirable for the first task in order to attain a high resolution along the line-of-sight, accuracy of the second task determines the quality of the cosmological estimates from statistical WL measurements in the end.

Recently, most surveys have employed template-based techniques (that assume a physical model) to tackle the first problem and empirical or machine-learning (ML) techniques (using e.g., a spectroscopic calibration sample) for the second problem. These choices follow directly from the requirements for the individual photo-z (low scatter) and the redshift distributions of ensembles of galaxies (low bias).

In this paper, we concentrate on the second problem and how it is solved for the cosmological analysis of the KiDS-1000 data set based on the fourth data release (Kuijken et al. 2019) of the Kilo-Degree Survey (KiDS; de Jong et al. 2013). The two other ongoing stage-III surveys are the Dark Energy Survey (DES; Flaugher et al. 2015) and the Hyper Suprime-Cam Subaru Strategic Program (HSC; Aihara et al. 2018) survey, whose redshift calibrations in their most recent cosmological analyses are described in Hoyle et al. (2018) and Tanaka et al. (2018), respectively.

The main requirement is to get an unbiased estimate of the mean redshift of the galaxies in the different tomographic bins (e.g., Laureijs et al. 2011). The uncertainty on this mean redshift needs to be of the order of $\sigma_{\langle z\rangle} \sim 0.01$ for stage-III surveys to not seriously jeopardise their constraining power (see Appendix A of Hildebrandt et al. 2017). These uncertainties are propagated into the full error budget and the exact requirement depends on which survey is analysed and which degradation with respect to the pure statistical uncertainty is deemed acceptable.

Certainly, higher-order moments of the redshift distributions also play a role. However, as WL is an integrated effect along the line-of-sight, the accuracy in estimating these higher-order moments is less important than the mean redshift and can, under normal conditions, be ignored for stage-III surveys (see e.g., Hoyle et al. 2018); it will become important though for upcoming stage-IV experiments like the ESA/NASA *Euclid* space mission (Laureijs et al. 2011), the *Vera C. Rubin* Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), and the *Nancy Grace Roman* Space Telescope (RST; Spergel et al. 2015). For these future missions, not only do the mean redshifts need to be controlled to $\sigma_{\langle z\rangle} \sim 0.001-0.002$, but the shape of the distribution also needs to be known accurately. At this level of precision, it also becomes relevant that the redshift distributions vary spatially with observing conditions so that even redshift distributions that perfectly describe the average of a survey are no longer sufficient and additional corrections are needed (Heydenreich et al. 2020). Similarly, correlations between the point spread function (PSF) ellipticity and the accuracy of the redshift measurements can no longer be ignored (Asgari et al. 2019).

The calibration of the redshifts is usually achieved with the help of a reference sample, which is often spectroscopic but can occasionally also utilise higher-quality photo-z, like the COSMOS-2015 catalogue (Laigle et al. 2016) based on photometry from more than 30 bands. These reference samples are selected in different ways than the WL source samples and hence are in general neither complete nor representative of those source samples. Different techniques have been developed to overcome this limitation, mostly through re-weighting the reference samples and through culling of the source samples. Lima et al. (2008) describe a re-weighting approach that utilises a k-nearest neighbour search in multi-dimensional magnitude-space to re-weight a spectroscopic sample such that it resembles a target photometric sample with unknown redshifts, which are the WL sources in our case. This approach was tested in Cunha et al. (2009) and later on used for KiDS in Hildebrandt et al. (2017, 2020) as well as for DES (Bonnett et al. 2016; Hoyle et al. 2018), and to some degree also for HSC (Tanaka et al. 2018). The estimated uncertainties with this approach are of the order of $\sigma_{\langle z\rangle} \sim 0.02$, which was sufficient for the first cosmological analyses that used only a fraction of the data from the stage-III surveys. In order to fully exploit the statistical power of the completed surveys, these uncertainties have to be improved by a factor of ∼2.

A similar re-weighting can be achieved by projecting the multi-dimensional magnitude-space into two dimensions with the help of a self-organising map (SOM; Kohonen 1982). This was pioneered in the framework of *Euclid* by Masters et al. (2015) and is now being used by KiDS (Wright et al. 2020a, hereafter W20a) and HSC (Tanaka et al. 2018), and suggested for DES (Buchs et al. 2019), too. These studies show that the SOM can, under certain conditions, reach uncertainties in the mean redshifts of tomographic bins of $\sigma_{\langle z\rangle} \sim 0.01$, or even better. Thus, it represents a very promising technique to calibrate redshifts for WL applications in current and future projects.

A complementary estimate of the source redshift distribution can be obtained through cross-correlation studies (Schneider et al. 2006; Newman 2008; Matthews & Newman 2010; Schmidt et al. 2013; Ménard et al. 2013; McQuinn & White 2013; Morrison et al. 2017; Johnson et al. 2017; Davis et al. 2017; Scottez et al. 2018; Gatti et al. 2018). Here, the colour-information is not used but instead the angular cross-correlation of the positions of a target source sample and a reference sample with known redshifts is employed. The appeal of this technique is that, unlike colour-based methods, the reference and target samples need not share any magnitude-space whatsoever. All galaxies at a given redshift cluster with each other and hence, in principle a bright reference sample that is relatively easy to observe spectroscopically can be used to calibrate the redshift distribution of a faint source sample. Besides spatial overlap on the sky, the most important requirement is that the reference sample covers the whole redshift range that needs to be probed for the target sample. An important nuisance in this method is the presence of galaxy bias: The fact that galaxies are biased tracers of the underlying matter field can influence the measured cross-correlation functions in a systematic fashion. For the purpose of estimating the redshift distribution, the absolute value of the galaxy bias can usually be neglected (its effect is removed through normalisation of the redshift distribution), any redshift evolution of the galaxy bias must be corrected (e.g., Newman 2008; Schmidt et al. 2013).

Clustering-redshift (clustering-z) measurements in the literature differ in the details of the implementation of the measurement itself as well as the galaxy bias correction scheme. All these approaches have one thing in common though: They do not yield a redshift distribution directly but instead some noisy representation of this distribution that needs to be interpreted via a model. This model can either be based on a different calibration approach (like the colour-based techniques discussed above; see e.g., Hoyle et al. 2018) or take the free form of a parametric function (spline, Gaussian process, etc.; see e.g., Johnson et al. 2017). Fitting this model to the clustering-z measurements thereby yields a redshift distribution estimate which can be propagated (along with relevant uncertainties) into a cosmological measurement. Here, we follow the methodology laid out in van den Busch et al. (2020), who test clustering-z measurements on mock catalogues that resemble the KiDS+VIKING-450 data set (Wright et al. 2019).

This paper is part of a series of KiDS-1000 papers describing the shear catalogue (Giblin et al. 2021), the methodology behind the cosmological analyses (Joachimi et al. 2021), results from cosmic shear (Asgari et al. 2021), a combined-probes analysis using cosmic shear, galaxy-galaxy lensing, and galaxy clustering from KiDS and BOSS data (Heymans et al. 2021), as well as constraints on cosmological models beyond ΛCDM (Tröster et al. 2021). Here, we present the redshift distributions used for the cosmological analyses of KiDS-1000. The structure is as follows. In Sect. 2, we describe the KiDS-1000 data set, the spectroscopic

reference samples, and the mock catalogues that mimic those samples. Section 3 presents results from the SOM method as applied to the KiDS-1000 data and the simulations, and Sect. 4 shows how the clustering-z technique is used to further calibrate the redshift distributions based on the SOM method. The results are discussed and the paper is summarised in Sect. 5, also explaining links to the KiDS-1000 companion papers.

2. Data

2.1. KiDS+VIKING imaging data

The KiDS-1000 catalogues used here are based on the fourth data release of KiDS (DR4; Kuijken et al. 2019), which includes near-infrared (NIR) photometry based on imaging from the fully overlapping VISTA Kilo degree INfrared Galaxy Survey (VIKING; Edge et al. 2013; Venemans et al. 2015). This data set comprises PSF-corrected nine-band $ugriZYJHK_s$ photometry (Kuijken 2008) and BPZ (Bayesian Photometric Redshift; Benítez 2000) photo-z estimates for more than 100 million objects over an area of $\sim$1000 deg^2. This constitutes roughly three quarters of the final KiDS+VIKING data set and more than a doubling of the data volume compared to the third data release of KiDS (KiDS-DR3; de Jong et al. 2017) that was based on $\sim$450 deg^2 and was used for previous KiDS cosmology analyses (Hildebrandt et al. 2020, hereafter H20).

Shapes are measured with the *lens*fit software for $\sim$31 million galaxies covering an effective unmasked area of 777.4 deg^2 with a weighted number density of 8.43 arcmin^{-2} (Giblin et al. 2021). This is the sample used for WL measurements and will be referred to as *sources* in the following. An in-depth description of a very similar sample of roughly half the size called KiDS+VIKING-450 (or KV450) and based on KiDS-DR3 can be found in Wright et al. (2019). There, the properties of the nine-band photo-z are described in detail and quantified by comparisons to deep spectroscopic redshift catalogues that overlap with KiDS. This information still applies to the KiDS-1000 data used here, as the depth and seeing distributions of KiDS-DR3 and KiDS-DR4 are extremely similar (see de Jong et al. 2017; Kuijken et al. 2019).

The photo-z point estimates z_B, corresponding to the peaks of the posterior redshift distributions of individual galaxies, are used to bin the sources into five tomographic redshift bins. In line with H20 the first four bins are spaced by $\Delta z_B = 0.2$ in the range $0.1 < z_B \leq 0.9$ whereas the fifth bin covers the high photo-z range $0.9 < z_B \leq 1.2$. The number densities of the galaxies (according to the definition of Heymans et al. 2012) in the five bins are listed in Table 1 (for an updated $n_{\rm eff}$ estimator that accounts for the impact of the shear responsivity correction, see Appendix C of Joachimi et al. 2021).

2.2. Spectroscopic calibration samples

The different calibration techniques require spec-z reference catalogues with different properties. For the colour-based calibration, it is required that the reference catalogue spans the same hyper-volume in nine-dimensional magnitude-space, whereas for the clustering-z calibration, a spatially overlapping large-area sample with an extended redshift distribution is needed.

2.2.1. Deep spectroscopy for colour-based calibration

The deep spectroscopic sample for the colour-based calibration with the SOM technique (Sect. 3) did not change between DR3 and DR4. It consists of a diverse combination of data from the zCOSMOS (Lilly et al. 2007, 2009), VVDS-Deep (VIMOS VLT Deep Survey; Le Fèvre et al. 2005, 2013, 2015), and DEEP2 (Newman et al. 2013) projects as well as some additional redshifts from the GAMA (Galaxy And Mass Assembly; Driver et al. 2011) deep field G15Deep (Kafle et al. 2018) and the CDFS (*Chandra* Deep Field South; ESO spec-z compilation consisting of spectra from Vanzella et al. 2008; Popesso et al. 2009; Balestra et al. 2010; Le Fèvre et al. 2013). The main properties of the samples are reported in Table 1 of W20a.

All of these fields have been observed in the nine KiDS+VIKING bands to at least KiDS+VIKING depth, in some cases much deeper. The only exception is the COSMOS field that has no VISTA z-band data. However, it has very deep CFHT (Canada France Hawaii Telescope) z-band data (Hildebrandt et al. 2009), which, due to the similarity of the MegaCam@CFHT z-band and the VIRCAM@VISTA z-band, can be used as a substitute. In cases where the imaging data in the deep redshift calibration fields is deeper than in KiDS+VIKING, we added Gaussian noise to arrive at a data set that is representative for KiDS+VIKING. In principle, one could also make use of deeper data in the calibration fields and improve the precision of the calibration for instance as described by Buchs et al. (2019), but we leave such an enhancement of the KiDS+VIKING redshift calibration to future work.

2.2.2. Wide-area spectroscopy for clustering redshifts

In H20, clustering-z (CZ)[1] were estimated with the help of spec-z data from the wide-area surveys GAMA-DR3 (Baldry et al. 2018), SDSS-DR12 (Eisenstein et al. 2011; Alam et al. 2015), 2dFLenS (Blake et al. 2016), and WiggleZ (Drinkwater et al. 2010) and complemented with information about the high-redshift part of the $n(z)$ from zCOSMOS, VVDS-Deep, and DEEP2. The same samples are employed here but with some significant changes, the most important one being approximately a doubling in the size of the overlap area between KiDS+VIKING and SDSS in the Northern Hemisphere as well as between KiDS+VIKING and 2dFLenS in the Southern Hemisphere. This alone significantly increases the signal-to-noise ratio (S/N) of the CZ measurements as described in Sect. 4. Additionally, we have relaxed some of the very conservative masking in previous KiDS CZ analyses.

From the SDSS spec-z compilation we only use sources observed as part of BOSS (Baryon Oscillation Spectroscopic Survey; Dawson et al. 2013) unlike in previous KiDS work where also the SDSS Main Galaxy Sample (MGS; Strauss et al. 2002) and the SDSS Quasar Sample (Schneider et al. 2010) were used. The reason behind this decision is the desire to minimise systematic errors through the correction for evolving galaxy bias, which becomes more complicated when different samples are combined. At low redshift, we have very high S/N from GAMA already and do not need the limited additional information from the SDSS-MGS. While a higher S/N at high redshift would be desirable, the sparsity of the SDSS-QSO sample does not add any significant information and the results are almost indistinguishable whether it is included or not.

The spec-z samples used for the CZ measurements are summarised in Table 2. We note that the areas in the COSMOS and VVDS-Deep fields used for CZ are slightly smaller than those

[1] We note that clustering-z were abbreviated as CC (cross-correlations) in previous KiDS papers. Here, we opt to switch to the new acronym CZ to more specifically refer to clustering-z.

Table 1. Properties of the five tomographic bins and the full source sample.

Bin	Selection	N	$n_{\rm eff}$ [arcmin^{-2}]	σ_ϵ	$N_{\rm gold}$ [arcmin^{-2}]	$n_{\rm eff,gold}$	$\sigma_{\epsilon,\rm gold}$	$n_{\rm eff,gold}/n_{\rm eff}$
1	$0.1 < z_{\rm B} \leq 0.3$	2 814 395	0.90	0.277	1 792 136	0.62	0.270	0.69
2	$0.3 < z_{\rm B} \leq 0.5$	5 612 329	1.62	0.268	3 681 319	1.18	0.258	0.73
3	$0.5 < z_{\rm B} \leq 0.7$	8 184 940	2.28	0.278	6 148 102	1.85	0.273	0.81
4	$0.7 < z_{\rm B} \leq 0.9$	5 797 140	1.53	0.261	4 544 395	1.26	0.254	0.82
5	$0.9 < z_{\rm B} \leq 1.2$	5 394 916	1.37	0.272	5 096 059	1.31	0.270	0.95
1–5	$0.1 < z_{\rm B} \leq 1.2$	27 803 720	7.66	0.272	21 262 011	6.17	0.265	0.80
All	–	31 446 584	8.43	0.273	N/A	N/A	N/A	N/A

Notes. Effective number densities are calculated with Eq. (C.12) of Joachimi et al. (2021), which itself is based on Eq. (1) of Heymans et al. (2012). The columns with the gold label correspond to the selection described in Sect. 3.

Table 2. Spectroscopic redshift samples used for the clustering-z calibration.

Survey	No. of spec-z	Area [deg^2]
zCOSMOS	8422	0.5
DEEP2	8698	0.8
VVDS	4194	0.5
GAMA	114 912	137.4
BOSS	47 332	262.5
2dFLenS	17 231	266.1
WiggleZ	42 328	130.1
Total	321 318	784.8 [b]

Notes. [a]The area quoted for the wide fields is a rough estimate calculated from the number of pointings that go into each cross-correlation measurement and the average unmasked area per pointing. [b]We note that there is significant overlap between GAMA, BOSS, and WiggleZ. Hence, the total area quoted here is not to be understood as an independent area.

used for the colour-based calibration as the former has stricter requirements on the spatial homogeneity of the data.

2.3. MICE mock catalogues

The KiDS+VIKING redshift calibration is validated on simulated mock catalogues based on the MICE simulation (Fosalba et al. 2015a,b; Crocce et al. 2015; Carretero et al. 2015; Hoffmann et al. 2015). The creation and properties of these mock catalogues is covered in detail in van den Busch et al. (2020). The KiDS+VIKING nine-band photometry and the BPZ photo-z are simulated within these mocks, whereas the shape measurement weights are sampled from the real data by assigning each mock galaxy the weight of its nearest neighbour in the KiDS-1000 data in r-band magnitude. This results in a mock source catalogue that closely resembles the data. The most important difference is that MICE only provides mock galaxies out to $z \sim 1.4$. Hence, we cannot test for possible high-z tails with the help of this mock, but we note that the core of the redshift distribution of each tomographic bin is well covered by these mocks.

In a similar way, the spec-z calibration samples are simulated by applying the original (or in some cases slightly modified) selection criteria to the mock photometry and implementing realistic magnitude- and redshift-dependent spectroscopic success rates. For details, we refer the reader to van den Busch et al. (2020). We also create an idealised reference sample by taking every 10th KiDS mock source. This somewhat unrealistic case can be used to test the CZ methodology and explore the unavoidable systematic error floor inherent to our CZ implementation, agnostic to the complexities of reference sample construction.

The mock catalogues for the deep spectroscopic fields are identical to the ones used in W20a. Hence, the mock results for the SOM calibration from W20a also apply to the data set presented here. These results will be discussed in Sect. 3.

The mock catalogues for the CZ measurement are simply expanded in area compared to the ones in van den Busch et al. (2020) to account for the larger area of the KiDS-1000 source sample compared to KV450. In fact, for the analysis presented here we create mock catalogues for all samples (WL sources, deep spec-z surveys, wide spec-z surveys) over an area of 744.4 deg^2 split into 1024 pointings of 0.727 deg^2 each. In particular for the deep fields, having such a large number of realisations makes it possible to estimate covariance matrices from the simulations that can be used to combine the results from the different surveys on the real data. One notable difference to the mock catalogues presented in van den Busch et al. (2020) is the fact that we use a pure BOSS sample instead of a combined SDSS sample also including the Main Galaxy and QSO samples, mirroring the approach taken on the KiDS-1000 data (see Sect. 2.2.2).

3. Colour-based redshift calibration with a self-organising map

Photometric redshifts rely on the fact that galaxy colours strongly correlate with redshift. The same information is exploited in the calibration of redshift distributions for WL applications with the help of a deep spec-z reference sample. In essence, this is quite similar to the well-known category of ML photo-z, with the important difference that we want to apply this to a target ensemble of galaxies with unknown redshifts rather than to individual galaxies. Additionally, the goals of colour-based $n(z)$ calibration are somewhat different from the goals of most ML photo-z codes, with the former being optimised towards low bias in the mean redshift and the latter often towards low scatter and low outlier rates.

3.1. Method

The inherent differences between a spec-z calibration sample and a typical WL source sample can – under certain circumstances – be overcome by re-weighting. This re-weighting of the

calibration sample is supposed to make the distributions of relevant quantities as similar as possible between the two samples. Once this is achieved, it is assumed that the weighted distribution of spec-z in the calibration sample should be a good estimate of the unknown distribution of redshifts of the target source sample. It is clear that this works better the more spec-z are available, the more complementary information (e.g., number of photometric bands) is used to establish the weighting, and the closer the selection of the spec-z sample resembles the source sample to start with (Gruen & Brimioulle 2017).

Lima et al. (2008) suggest an approach that estimates the density of both samples in high-dimensional magnitude-space via a k-nearest-neighbour (kNN) method. The ratio of the densities in each point in magnitude-space is then used as a weight for the spec-z in that place. Essentially, spec-z that are underrepresented compared to the unknown target sample are up-weighted and spec-z that are over-represented are down-weighted. It can be shown on simulations (Lima et al. 2008; Wright et al. 2020a) that this approach yields good results if the magnitude-space is sufficiently high-dimensional, the photometry has high S/N, and the magnitude-space of the target sample is fully covered with spec-z calibrators.

Whether the first two requirements are sufficiently met can realistically only be investigated with simulations. The third requirement, however, can partly be assessed with the data themselves by checking the overlap of the target and calibration samples. Being a high-dimensional problem, such checks need to make use of some dimensionality-reduction technique. Masters et al. (2015) argued that SOMs are well suited for this purpose.

W20a show that the SOM method can be used to actually carry out the estimation of the redshift distribution, $n(z)$, without further need for the kNN method. Instead of estimating the densities of target and calibration sample at the location of each calibration source, the densities are estimated in each cell of the SOM. Moreover, the SOM gives the user a simple tool to cull from the target sample sources that are not represented by the spec-z calibration sample, that is sources that lie in cells that are not filled with at least one reference object. Each tomographic redshift bin is calibrated individually with a calibration sample that is limited to the same photo-z (z_B) range. The following additional criterion is established to select good SOM cells:

$$|\langle z^\mathrm{s}_\mathrm{spec}\rangle_i - \langle z^\mathrm{p}_\mathrm{B}\rangle_i| < \max\left[5 \times \mathrm{nMAD}\left(\langle z^\mathrm{s}_\mathrm{spec}\rangle - \langle z^\mathrm{s}_\mathrm{B}\rangle\right), 0.4\right], \quad (1)$$

where the superscripts s and p refer to the spectroscopic calibration and the photometric target samples, respectively, the angular brackets indicate an unweighted average, the index i refers to a single SOM cell, and the normalised median absolute deviation[2] on the right-hand side is taken over the full SOM. This criterion rejects cells that show suspiciously large deviations between the mean spectroscopic redshift of all calibration objects in a cell and the mean photo-z of all target objects in that cell. See W20a and Wright et al. (2020b) for more details.

In this way, W20a define a KiDS 'gold' sample with smaller number density and more robust $n(z)$ estimates. The KiDS-1000 WL analyses all make use of this gold selection to benefit from the robustness of the $n(z)$ estimates.

Wright et al. (2020b) analyse the gold sample for KV450 with the corresponding SOM-based $n(z)$, finding very good agreement in their cosmological parameter estimates with previously published results based on the full samples and an $n(z)$ estimated with the kNN method (H20). The SOM analysis presented here follows the methods presented in W20a. Given that the spec-z calibration sample is identical in both studies, the only difference is the larger (by a factor of ∼2) target catalogue with slightly updated absolute photometric calibration and updated *lens*fit weights (Giblin et al. 2021). As the SOM analysis was not limited by the (already large) size of the target sample in W20a, this should only result in very minor changes to the $n(z)$.

3.2. Results from MICE mocks

W20a use the MICE mock catalogues described in Sect. 2.3 to estimate residual biases in their SOM-estimated redshift distributions. After optimising the SOM setup with a series of tests they also introduce additional clustering of the SOM cells[3]. By combining multiple cells into a cluster, an optimal compromise between fidelity and shot-noise is found. The redshift distributions estimated with the SOM technique on the mock catalogues are displayed in Fig. 1.

W20a report values for the bias of the mean redshift in the five tomographic bins used for the gold cosmic shear analysis of Wright et al. (2020b) in the form

$$\Delta\langle z\rangle^\mathrm{SOM}_j = \langle z\rangle^\mathrm{SOM}_j - \langle z\rangle^\mathrm{true}_j, \quad (2)$$

where the averages are taken per tomographic bin $j \in \{1, 2, 3, 4, 5\}$. Since the mocks did not change in the meantime and the KiDS-1000 data closely resemble the KV450 data, the same biases apply to the KiDS-1000 calibration presented here. We note that the improvement to the *lens*fit weight recalibration methodology between KV450 and KiDS-1000, as discussed in Sect. 2.2 of Giblin et al. (2021), is not propagated into the mock catalogues as it does not significantly change the mean properties of the KiDS-1000 tomographic bins compared to the KV450 bins. Values for the mean biases and their uncertainties as estimated from 100 simulated lines-of-sight are reported in the second column of Table 3. Those can be compared to the biases estimated for the mean redshifts of the full samples of H20 with the kNN method (last line of Table 3 of W20a), which are significantly larger and range from 0.047 in the first bin to −0.013 in the fifth bin.

As the uncertainties quoted in Table 3 are estimated from 100 realisations along different lines-of-sight for the mock spec-z calibration sample, these uncertainties include contributions from photometric noise, shot-noise due to the limited sample size, spectroscopic selection effects and incompleteness, and sample variance due to large-scale structure. The latter effect leads to a correlation of the uncertainties, which is also estimated from these 100 realisations. We report the correlations in Fig. 2. Neighbouring tomographic bins are correlated by up to 36%, while more widely separated bins are only weakly correlated or also weakly anti-correlated.

The uncertainties and their correlations are taken into account in the cosmological analyses with the KiDS-1000 data (Asgari et al. 2021; Heymans et al. 2021; Tröster et al. 2021). In order to account for inherent imperfections in the simulation we conservatively enlarge all these uncertainties by a factor of two in the fiducial analyses.

[2] Normalised in such a way that it equals the standard deviation for a Gaussian distribution.

[3] This is not to be confused with the physical clustering of galaxies and just describes the merging of SOM cells with similar properties.

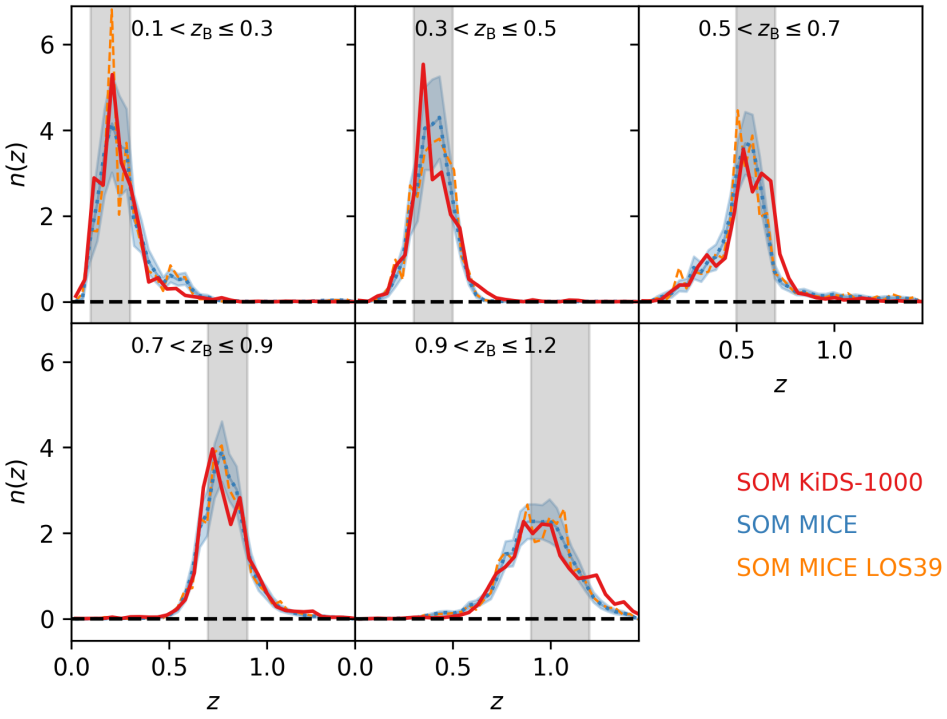

Fig. 1. Redshift distributions for the five tomographic redshift bins used in the KiDS-1000 cosmological analyses estimated with the SOM method of Wright et al. (2019). The grey vertical bands indicate the photo-z cuts defining the bins. Solid red lines show the estimate from the KiDS-1000 data whereas the dotted blue lines and their confidence intervals represent the average and standard deviation of all lines-of-sight of the MICE mocks. The dashed orange lines show one representative (in terms of its mean redshifts) line-of-sight (number 39 in our list) that is used in Sect. 4.2.

Table 3. Redshift calibration for the five tomographic bins used in the KiDS-1000 cosmology analyses.

Bin	$\Delta\langle z\rangle^{\rm SOM}$ MICE	$\delta z^{\rm CZ}$ ± stat. ± syst. MICE	$\delta z^{\rm CZ}$ ± stat. ± syst. KiDS	$\delta z^{\rm CZ}$ ± comb. KiDS
1	0.000 ± 0.011	0.001 ± 0.002 ± 0.004	−0.001 ± 0.004 ± 0.004	−0.001 ± 0.012
2	0.002 ± 0.011	−0.002 ± 0.002 ± 0.004	0.004 ± 0.003 ± 0.005	0.004 ± 0.013
3	0.013 ± 0.012	0.004 ± 0.003 ± 0.010	0.011 ± 0.004 ± 0.016	0.011 ± 0.020
4	0.011 ± 0.009	0.015 ± 0.001 ± 0.024	−0.008 ± 0.006 ± 0.007	−0.008 ± 0.013
5	−0.006 ± 0.010	0.003 ± 0.002 ± 0.004	0.003 ± 0.007 ± 0.003	0.003 ± 0.013

Notes. Bias in the mean redshift (Col. 2) as estimated with the SOM method from the MICE mocks (W20a). The uncertainties have been multiplied by a factor of two to account for residual differences between mocks and data. Columns 3 and 4 report the best-fit values for the $\delta z^{\rm CZ}$ parameters (defined in Eq. (9)) on the MICE mocks and the KiDS-1000 data, respectively. The values from Col. 4 are based on fits to the SOM redshift distributions, which carry their own uncertainty (Col. 2). In Col. 5 we report the same shifts as in Col. 4 but combine all sources of uncertainty.

3.3. Results from KiDS-1000 data

We update the SOM analysis of W20a by populating the SOM with the new KiDS-1000 catalogues instead of the KV450 catalogues that were used in that paper. This leads to slightly different redshift distributions, which are also displayed in Fig. 1, and different effective number densities ($n_{\rm eff}$) as well as ellipticity dispersions σ_ϵ reported in Table 1. By applying the gold selection, roughly 20% of the effective source density is removed, which slightly increases the statistical noise (shape noise). Partly counteracting the decrease in $n_{\rm eff}$, however, is a small reduction in σ_ϵ, which sets the noise level per source of WL measurements.

Comparing the number densities of the gold selection for KV450 and KiDS-1000 (see Table 1 here and Table 2 of W20a) reveals some notable differences. In particular, the first and second tomographic bins show significantly lower representation fractions on the KiDS-1000 data. We attribute this to subtle differences in the absolute photometric calibration between KV450 and KiDS-DR4 (see Wright et al. 2019; Kuijken et al. 2019) combined with our assumed number of hierarchical clusters in the SOM. W20a demonstrate that, while the choice of cluster number (see their Fig. B1, panel b) can introduce swings of >20% in representation fraction, the reconstructed redshift distributions remain entirely unbiased (panel c). As a result of this conclusion, we chose not to re-optimise the number of hierarchical clusters used for DR4 even after our slight change in calibration and retraining of the fiducial SOM.

4. Calibration with clustering redshifts

In the following, we describe the complementary clustering redshift technique that yields an independent estimate of the mean redshifts of the galaxies in the tomographic bins.

4.1. Method

4.1.1. Measurement

The clustering redshift methodology used for KiDS-1000 closely follows the approach described in van den Busch et al. (2020). This approach implements the technique suggested by Schmidt et al. (2013) using small-scale clustering in a single broad radial bin with an additional radial weighting. By pre-selecting galaxy samples that are already relatively narrowly

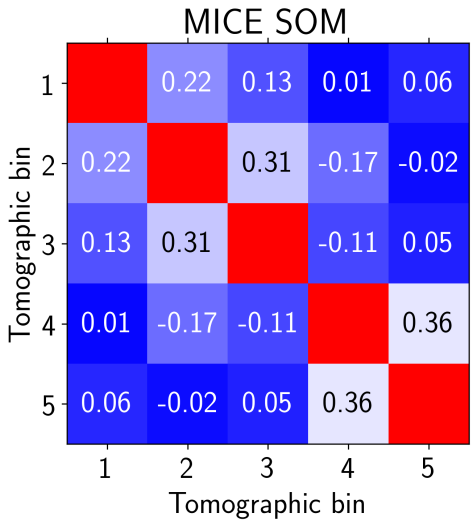

Fig. 2. Correlation matrix of the uncertainties of the $\Delta\langle z\rangle_i^{\rm SOM}$ from the SOM analysis of the MICE mocks reported in Col. 2 of Table 3.

located in redshift (i.e. the tomographic redshift bins), any effects of evolving galaxy bias are minimised to start with. The evolution of the galaxy bias of the reference sample is mitigated by estimating the angular auto-correlation function of this sample in the same radial- and redshift bins as the cross-correlation measurement. Any residual effect of the bias evolution of the source sample itself can in principle be mitigated via an internal consistency check (also called self-consistency bias mitigation or SBM; see Sect. 3.3 of van den Busch et al. 2020). This check is based on comparing the results from a broad target sample (e.g., all tomographic bins combined) with the weighted sum of the narrow samples. As each of the narrow samples gets normalised individually, the weighted sum is not exactly equal to the measurement on the broad sample. Differences can be interpreted as being due to evolving galaxy bias and approximated by a parametric model (Davis et al. 2018).

However, the galaxy bias of typical WL source samples evolves only very slightly over the redshift baseline of the core of a single tomographic bin. This limits the importance and usefulness of this approach. Only at high S/N of the CZ measurements can this additional complexity in the model be constrained by the data. We distinguish the following estimates of the redshift distribution:

$w_{\rm CZ}(z)$ raw CZ measurements (3)

$\tilde{n}_{\rm CZ}(z)$ CZ after correction for reference bias (4)

$n_{\rm CZ}(z)$ fully corrected CZ (reference bias and SBM), (5)

(see van den Busch et al. 2020, for the performance of the different options on mock catalogues). We note that the bias of outlier populations with a redshift very different from the core of the $n(z)$, and potentially also a linear bias value that is very different, cannot be reliably corrected with this method. As such, this method is only useful for narrow, unimodal redshift distributions.

In comparison to previous KiDS analyses, we have implemented a number of changes to the CZ methodology. Unlike Hildebrandt et al. (2017, 2020) we use an updated version of the angular cross-correlation code `the-wizz` (Morrison et al. 2017) called `yet-another-wizz` or `yaw`. We refer the reader to van den Busch et al. (2020) for a detailed description of the features of `yaw`. The main advantage of this new version is that it avoids the inherent sky pixelisation of `the-wizz`, which is inherited from the library `STOMP` (Scranton et al. 2002). This improvement yields more realistic uncertainties, especially for small angular scales that are often probed at high redshift for a given comoving scale.

We further experiment with different radial scales. van den Busch et al. (2020) used comoving scales of $100\,{\rm kpc} < r < 1\,{\rm Mpc}$ for their measurements throughout. Here we also explore the performance of the CZ method with additional scales of $30\,{\rm kpc} < r < 300\,{\rm kpc}$, $50\,{\rm kpc} < r < 500\,{\rm kpc}$, and $500\,{\rm kpc} < r < 1.5\,{\rm Mpc}$[4]. Especially, the smaller scales yield very high S/N, at the price of potentially more complex bias evolution. With the mock catalogues, the impact of this can be tested. We note that the limited CZ analysis of Hildebrandt et al. (2017) also measured over scales of $30\,{\rm kpc} < r < 300\,{\rm kpc}$ and reached a usable S/N from less than $2\,{\rm deg}^2$ of area covered by deep pencil-beam surveys.

The redshift binning is less critical as our approach should be able to correct for all biases regardless of this binning. Here we choose 45 redshift bins of constant radial comoving length in the redshift range $0 < z < 3$. We use the same binning for the data and the mocks but can essentially only use the lower half of the redshift range for the mocks as the MICE galaxy population only extends to $z = 1.4$.

4.1.2. Covariance

The other update compared to van den Busch et al. (2020) and the CZ analysis in H20 concerns the covariance matrix of the CZ measurements. Due to the limited size of the reference samples all previous CZ analyses with KiDS estimated the covariance from a bootstrap or jackknife re-sampling over all ($\sim$1 deg^2) pointings that went into the measurement. We will call this approach of estimating the covariance via bootstrap (A) in the following and use it by default.

This implementation of re-sampling neglects any differences in spectroscopic coverage between pointings. In effect, the subsamples, which the bootstrap samples are constructed from, can have different statistical weights, especially at high redshift. In general, this leads to an underestimation of the uncertainty of the CZ measurements. As shown by the mock analysis of van den Busch et al. (2020), which also uses approach (A) and in principle also suffers from the same deficiency, this can still yield sufficiently accurate results. In the following we try to estimate the additional uncertainty due to this effect and propagate it into our results.

Instead of treating all measurements from the different reference surveys equally, we can also split the analysis and first analyse the different surveys independently. While this was not really possible with previous KiDS data releases, even after splitting the KiDS-1000 data volume still leaves >100 pointings for each of the wide-area reference surveys i to empirically estimate the CZ data $n_i(z)$ and a corresponding covariance matrix C_i via bootstrap re-sampling. The measurements of all wide-area surveys are then combined with precision weighting (or inverse covariance weighting) assuming uncorrelated Gaussian uncertainties

[4] We note that a cosmological model needs to be assumed to convert angular scales into comoving distances. Here, we assume a *Planck*-2015 cosmology (Planck Collaboration XIII 2016), but this choice has negligible influence on our results as long as the same scales are used consistently for all correlation function measurements of a given tomographic bin.

$$n(z) = \mathsf{C} \sum_i \mathsf{C}_i^{-1} \boldsymbol{n}_i, \tag{6}$$

where C is the combined, precision-weighted covariance estimated as

$$\mathsf{C} = \left(\sum_i \mathsf{C}_i^{-1} \right)^{-1} \tag{7}$$

and $\boldsymbol{n}_i$ is the redshift distribution vector of the ith bin. This approach will be called (B) in the following and it can be applied to any of the three estimates for CZ described in Eqs. (3)–(5). In this way, only subsamples with comparable statistical properties enter each of the bootstrap estimates, making those more reliable. However, as mentioned above, this method also assumes that there is no correlation between the measurements from different reference samples, which is not true due to the overlap of some of these samples (GAMA, BOSS, WiggleZ). Again, we test the impact of violating this assumption on the mocks, which replicate the overlap of the reference surveys in the KiDS-1000 data. Due to the small data volume this approach (B) is not feasible on the deep pencil-beam fields.

We complement these two empirical estimates of the CZ covariance with a simulation-based approach that we will call (C) in the following. Instead of applying any bootstrap resampling to the data, we leverage the MICE mock catalogues described in Sect. 2.3 to estimate a covariance matrix. For each of the seven reference samples quoted in Table 2 – regardless of whether it is a wide-area or deep pencil-beam sample – we measure the CZ on 1024 pointings of the MICE mocks. This is sufficient to estimate a low-noise covariance matrix via bootstrap resampling for each individual reference sample, which can then be scaled to the actual area quoted in Table 2. Measurements and corresponding covariance matrices from the different reference samples are then combined again with precision weighting.

All three covariances show consistently that the CZ measurements are not strongly correlated between different redshifts, as expected from uncorrelated large-scale structure along the line-of-sight. See an example correlation matrix of the idealised reference sample cross-correlated with the target tomographic bins in Fig. 3. We always suppress noise in the covariance by setting those covariance elements to zero that correspond to different redshift bins. However, at a given redshift there is some correlation between the measurements for different tomographic bins (most pronounced for neighbouring bins) as these measurements are based on the same reference objects.

The different approaches to estimate the covariance are affected by different levels of noise, with approach (A) on the deep pencil-beam surveys being the noisiest and approach (C) generally being the least noisy. Depending on the noise level we decide whether to use or ignore the off-diagonal elements that correlate measurements in different tomographic bins at the same redshift. For example, with our fiducial approach (A) on the deep fields the estimates of these covariance elements are too noisy and need to be ignored to allow for an inversion of the matrix.

4.1.3. Model fitting

Clustering-z measurements are noisy representations of an underlying redshift probability distribution. Noise fluctuations can lead to negative clustering amplitudes that cannot be readily converted into a probability density. Hence, one needs a model to interpret these noisy data points. In general, we minimise

$$\chi^2 = \left[n_{\rm CZ}(z) - m_\theta(z) \right]^{\rm T} \mathsf{C}^{-1} \left[n_{\rm CZ}(z) - m_\theta(z) \right], \tag{8}$$

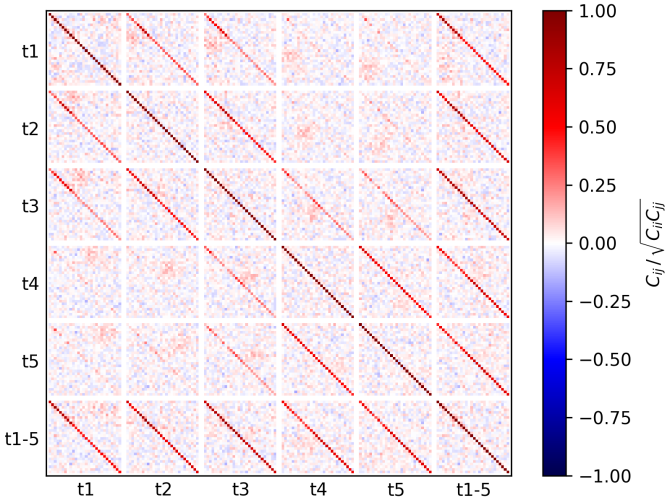

Fig. 3. Correlation matrix of CZ measurements from the MICE mocks using an idealised reference sample with high number density. There are six blocks in a line, each 30 pixels wide corresponding to 30 redshift bins in the range $0 < z < 1.4$ (with the first and last redshift bin containing no galaxies due to the redshift limits of MICE and shown white here). The first five blocks correspond to the five tomographic bins and the sixth block to the combined sample. The latter one is obviously correlated with all other samples as it shares target galaxies with the other bins.

where $m_\theta(z)$ is some model of the clustering-redshift distribution $n_{\rm CZ}(z)$ with parameters θ.

In H20 and van den Busch et al. (2020) we used redshift distributions from the kNN re-weighting technique (Lima et al. 2008, dubbed DIR in previous KiDS papers) as a model, which was shifted by an offset $\delta z^{\rm CZ}$ to yield a best-fit to the CZ data. Here, we switch to the SOM-estimated redshift distributions from Fig. 1 as a model, that is

$$m_\theta(z) = A\, n_{\rm SOM}(z + \delta z^{\rm CZ}), \tag{9}$$

where $\theta = (A, \delta z^{\rm CZ})$ are the fit parameters to be minimised. We typically only report $\delta z^{\rm CZ}$ as the value of the amplitude A is unimportant after normalisation of the best-fit model. We note that we distinguish between discrete differences in mean redshifts as $\Delta\langle z\rangle^{\rm SOM}$ in Eq. (2) and continuous fitting parameters such as $\delta z^{\rm CZ}$ in Eq. (9) by the use of capital Δ and small δ, respectively.

We expect the SOM redshift distributions to be less biased than the DIR-estimated ones (Wright et al. 2019) so that the results should come closer to the idealised case discussed in van den Busch et al. (2020), where the true redshift distributions were used on the MICE simulations to discover any residual biases in the CZ method. We also report such results from some tests with the true redshift distribution for MICE below.

The motivation for using the SOM $n(z)$ and fitting a shift is the fact that cosmic shear measurements are mostly sensitive to the mean redshift of the source sample. A bias in the mean redshift due to a coherent offset of the core of the redshift distribution is readily captured in the best-fit value of this shift parameter. However, it should be noted that a bias in the mean due to outliers cannot be captured by this simple model.

Another general problem with this approach is that the shape of the DIR or SOM $n(z)$ is not perfectly accurate, meaning their higher-order moments differ from the true redshift distribution. While the DIR $n(z)$ are typically too broad, the opposite is true for the SOM $n(z)$ that are typically slightly too narrow. These

properties are revealed by the mock analysis of Wright et al. (2019). If the S/N of the CZ measurements changes significantly with redshift, such a bias in the shape of the model can lead to a bias in the mean redshift as estimated from the best-fit shift parameter $\delta z^{\rm CZ}$. This can be easily understood by imagining some CZ measurement for a tomographic bin, whose S/N is high on the low-z side of the peak and low on the high-z side. The fit of any model will be driven by the high S/N data points at low-z and influenced little by the low S/N data points at redshifts higher than the peak. If the model is too broad this will bias the inferred mean redshift high, if the model is too narrow the inferred mean redshift will be biased low. See Appendix A for a toy model illustrating this effect.

In order to avoid such problems, one could add more parameters to the model that would account for this behaviour, for example by parametrically modifying the width. We investigate such more complex models and apply those to the KV450 data in Stölzner et al. (2020). Instead, we opt to not combine the CZ measurements from the wide-area and deep pencil-beam surveys, as was done before, because such a combination would exactly yield a strongly varying S/N over the peaks for the fourth and fifth tomographic bins (see Fig. 16 from van den Busch et al. 2020). We therefore use the wide-area surveys exclusively for the first three tomographic bins and the deep pencil-beam surveys for the fourth and fifth bin. This yields relatively symmetric S/N over the redshift range of the peak of each of these bins, and hence fortifies our results against this particular systematic effect. It also allows us to pick different scales over which we evaluate the correlation functions. We use scales of $100\,{\rm kpc} < r < 1\,{\rm Mpc}$ with a full bias correction ($n_{\rm CZ}(z)$, Eq. (5)), for tomographic bins 1−3, and $30\,{\rm kpc} < r < 300\,{\rm kpc}$ with a correction for the bias of the reference sample only ($\tilde{n}_{\rm CZ}(z)$, Eq. (4)), for tomographic bins 4 and 5. These choices are justified by the S/N in the different bins. However, the effect of these choices is captured in our systematic error budget as described in the following.

The model chosen here is quite inflexible. Thus, it cannot be expected to give a good fit (e.g., in terms of reduced χ^2) to the complex CZ data that are affected by residual galaxy bias, variable observing conditions, spectroscopic selection effects, etc., all of which are not modelled. Furthermore, the model itself, being based on the SOM method (Sect. 3), is noisy, which is not accounted for in our fit. Most importantly, its shape can be slightly different for systematic reasons or due to sample variance. This can be tested on the realistic mocks by using the true redshift distributions as a model. This yields a very good χ^2 arguing that a mismatch in the shape is the most important aspect driving the χ^2 high in realistic situations (see also van den Busch et al. 2020, for a discussion of the effect of the shape of the model). Instead of using only the (possibly unreliable) uncertainties of the best-fit parameters when fitting the SOM $n(z)$ to the CZ data we opt to explore systematic errors by also estimating results for alternative choices of measurement scales, covariance determination, and galaxy bias removal.

Our fiducial approach replicates the methodology of van den Busch et al. (2020), with a purely empirical covariance matrix estimated from bootstrap re-sampling, that is approach (A). For the deep pencil-beam surveys (bins 4 and 5) we also conduct alternative measurements with the simulation-based covariance (C) and a corresponding precision-weighted combination of the different deep fields, as well as measurements at slightly larger scales, $50\,{\rm kpc} < r < 500\,{\rm kpc}$. For the wide-area surveys we consider the covariance alternatives (B) and (C), which allow for a combination of the results from the different surveys via precision-weighting, and alternative measurement scales of $500\,{\rm kpc} < r < 1.5\,{\rm Mpc}$. We also try all different alternatives for bias removal listed in Eqs. (3)–(5) on the wide fields (bins 1−3) but limit ourselves to the methods referred to in Eqs. (3) and (4) for the deep fields (bins 4 and 5). We take the weighted scatter between these alternatives as an estimate of the systematic error inherent to our fiducial choices.

This is far from a perfect estimate of the systematic uncertainty and should not be considered highly precise. Rather it should give a rough idea of possible systematic problems, which is still preferable over quoting a purely statistical uncertainty here. It is clear that this area needs further attention in the future when statistical errors shrink further.

4.2. Results from MICE mocks

First, we repeat the analysis of van den Busch et al. (2020) with the idealised reference sample described in Sect. 2.3. Using the true redshift distributions as a model yields very small shifts $\delta z_i^{\rm CZ} \lesssim 0.005$ for all tomographic bins i. This can be regarded as the systematic error floor of our current implementation. We cannot expect the clustering-z with more realistic reference samples and less idealised models to perform any better than this. It should be noted that under these idealised conditions with the very small uncertainties achieved with this dense reference sample, the goodness-of-fit is poor, with values of $\chi^2/{\rm d.o.f.} \gtrsim 3$ (unlike the case with the more realistic reference samples, where a fit with the true $n(z)$ yields a $\chi^2/{\rm d.o.f.} \sim 1$, as reported above). We attribute this to the inherent systematic limitations, even with an idealised setup, a general tendency for our errors to be underestimated, and the simplicity of our model. Hence the decision to ignore the goodness-of-fit in the following and estimate the full error budget by exploring alternative analysis choices.

Moving to the fiducial setup, with the first three tomographic bins being calibrated by CZ measurements on the wide-area surveys and the upper two tomographic bins being calibrated by the deep pencil-beam surveys, these numbers vary only very slightly, when the SOM $n(z)$ from one representative line-of-sight (in terms of the mean redshifts of the five tomographic bins) are used as a model. The best-fit solutions and their respective best-fit parameters $\delta z_i^{\rm CZ}$ are reported in Fig. 4 and Table 3 (Col. 3). Only bin 4 shows a somewhat larger bias of $\delta z_4^{\rm CZ} \sim 0.015$, indicating that the CZ prefers a slightly lower mean redshift than the SOM estimate (in agreement with the value of $\Delta\langle z\rangle^{\rm SOM}$ in that bin). Fitting uncertainties are of the order $\sigma(\delta z_i^{\rm CZ}) \lesssim 0.003$, but should not be taken at face value due to the limitations mentioned above.

As described, we explore some alternative scenarios to estimate robust systematic uncertainties for these shifts. The standard deviation between these scenarios ranges from $\sigma_{\rm syst.}(\delta z_i^{\rm CZ}) = 0.004$ for bins $i \in (1, 2, 5)$ to $\sigma_{\rm syst.}(\delta z_4^{\rm CZ}) = 0.024$ for bin 4. This indicates that all shifts quoted above are consistent with zero. We report the fitting errors and these systematic error estimates in Table 3 (Col. 3).

As we are using the SOM redshift distributions from a single line-of-sight (see Fig. 1 for the differences in 100 lines-of-sight), it is clear that there is some residual sample variance that is not fully accounted for in either of the uncertainties quoted in Col. 3 of Table 3. We have, for now, ignored this effect. We do, however, propagate an estimate of the calibration uncertainty due to sample variance in our final CZ results for KiDS-1000 in Sect. 4.3.

Figure 4 highlights some of the problems encountered with the interpretation of CZ measurements. The uncertainty estimates for the upper two bins are quite noisy due to the small

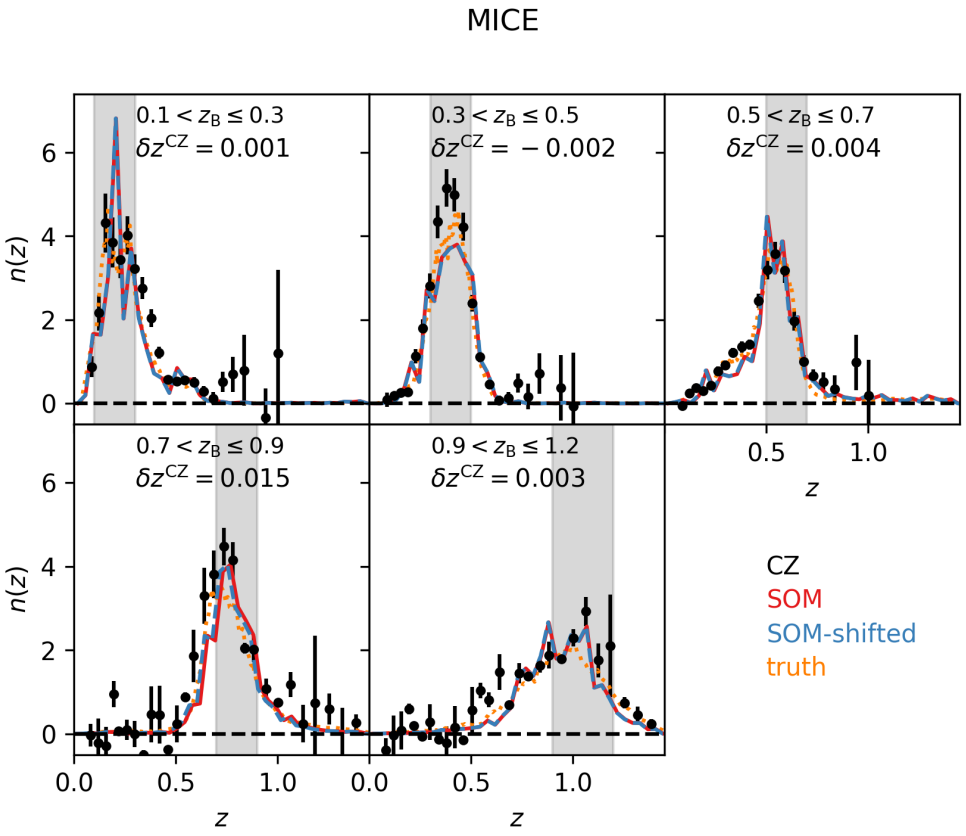

Fig. 4. Clustering-z measurements on the MICE mocks with the fiducial setup, i.e. using the wide fields and scales of 100 kpc $< r <$ 1 Mpc for the first three tomographic bins (*top row*) and the deep fields and scales of 30 kpc $< r <$ 300 kpc for the upper two tomographic bins (*bottom row*). The original SOM redshift distributions from a representative line-of-sight are shown in solid red and the best-fit model is shown in dashed blue. The true redshift distributions are shown in dotted orange for comparison.

number of deep fields that contribute to the bootstrap resampling. Some of these data points clearly influence the fit but the δz^{CZ} results suggest that this problem does not lead to an overall large bias. Moreover, the shape of the SOM redshift distributions is somewhat different than the shape suggested by the CZ data points. This mismatch will depend on the line-of-sight chosen for the SOM and highlights the limitations of our modelling. We take the pragmatic stance that, as long as the results for δz^{CZ} indicate almost unbiased measurements, these limitations are unimportant for the conclusions drawn in this work.

4.3. Results from KiDS-1000 data

Having verified the methodology from Sect. 4.1 with the mock catalogues in Sect. 4.2, we finally apply the clustering-z technique to the KiDS-1000 data. Results for the fiducial setup are reported in Fig. 5 and Table 3. The best-fit shift parameters δz^{CZ} are of the same order as in the simulated analysis, which increases our confidence in the realism of our mock catalogues. There are some subtle differences, such as the bias and systematic scatter in the third bin being slightly larger on the data than on the simulations, with the opposite behaviour in the fourth bin, but the details certainly depend on the line-of-sight chosen for the mocks. Overall the agreement is quite good. We note that there is some mismatch in the shape of the $n(z)$ between the SOM and CZ data for some of the bins. We attribute this partly to sample variance as the SOM $n(z)$ is based on a few lines-of-sight (the deep fields) whereas the clustering-z are estimated from hundreds of square degrees.

We propagate the uncertainty of the mean redshifts of the SOM $n(z)$ into this estimate as we are essentially using a noisy model (the noise being a combination of statistical shot noise, cosmological sample variance, and some other contributions; see Wright et al. 2020a). We conservatively multiply this SOM uncertainty by a factor of two (Col. 2 of Table 3) to account for limitations in our MICE mocks, in particular the $z < 1.4$ redshift limit. Then we add this inflated error and the other errors quoted in Col. 4 of Table 3 in quadrature to arrive at the combined uncertainty quoted in the last column.

The magnitude of the uncertainties in the clustering-z measurements is very comparable to the ones from the SOM (compare Cols. 2 and 4 of Table 3). This means that with the KiDS-1000 data set we reach full complementarity between these different approaches of calibrating the $n(z)$.

The uncertainties in the different bins are correlated, with the covariance matrix calculated as the sum of the covariances of the SOM uncertainties, the covariance of the fit parameters δz^{CZ}, and the covariance of the different alternatives explored in the systematic error estimation. This combined correlation matrix is shown in Fig. 6.

5. Discussion and summary

The primary calibration method to estimate redshift distributions for the KiDS-1000 cosmology analyses is based on the SOM method. This method projects the high-dimensional magnitude-space into two dimensions so that one can easily identify KiDS galaxies that are not represented by a reference sample. With the same spectroscopic calibration surveys used as a reference, this calibration is almost identical to the one presented in W20a, with the minor exception of updated *lens*fit shape measurement weights and photometric calibration in the current analysis. The accuracy estimates and systematic error discussion of W20a also hold for KiDS-1000 due to the similarity to the KiDS+VIKING-450 data set (Wright et al. 2019). We expect the gold samples defined by the SOM method to be

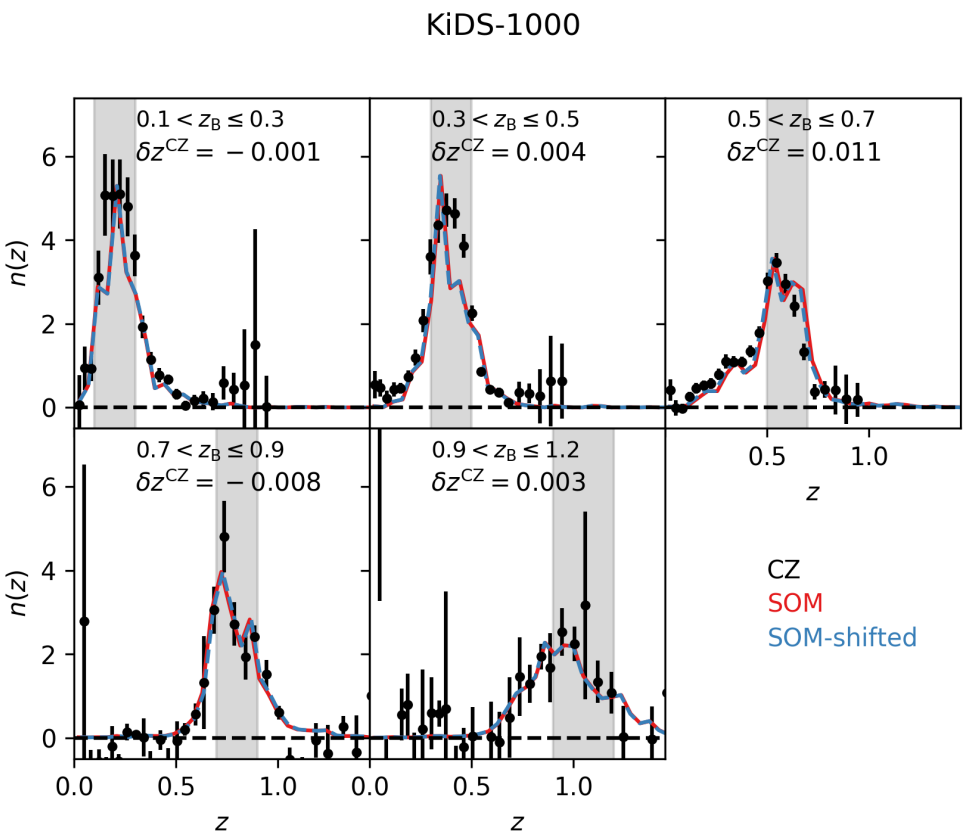

Fig. 5. Same as Fig. 4 but for the KiDS-1000 data.

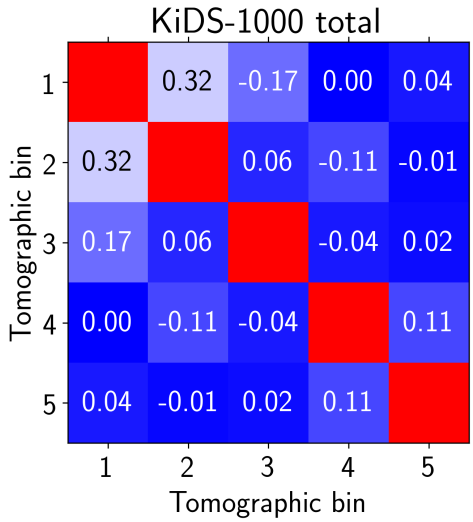

Fig. 6. Correlation matrix of the combined uncertainties of the δz_i^{CZ} from the CZ analysis of the KiDS-1000 data reported in Col. 5 of Table 3.

more robustly represented by their corresponding redshift distributions than the full samples used in H20 were represented by their DIR-estimated $n(z)$. It should be noted that Wright et al. (2020b) showed that cosmological conclusions, in particular the tension w.r.t. Planck, are not strongly affected by this switch in the redshift calibration, while the use of SOM calibration in the KiDS-1000 analyses reduces the redshift calibration systematic uncertainties (compared to the DIR). This represents an important step for systematic error control to keep pace with the growing statistical power of WL surveys.

In previous KiDS analyses, we neglected the correlation of the uncertainties in the mean redshifts of the tomographic bins. Here we report these correlations for the SOM method, as estimated from the covariance of 100 lines-of-sight of the MICE mock catalogues. These correlations will be taken into account in accompanying KiDS cosmological measurements (Asgari et al. 2021; Heymans et al. 2021; Tröster et al. 2021) as described in Joachimi et al. (2021). The SOM $n(z)$ are further validated in Giblin et al. (2021) together with the calibration of the multiplicative shape measurement bias by performing a shear-ratio test (Jain & Taylor 2003; Heymans et al. 2012; Kitching et al. 2015; Schneider 2016) similar to previous KiDS analyses. The $n(z)$ pass this test despite the greater statistical power of KiDS-1000, lending further credence to the stability of the SOM redshift calibration presented here.

Clustering redshifts (CZ) are used as a validation technique for the SOM $n(z)$ in this paper. With unprecedented overlap with spectroscopic surveys over hundreds of square degrees containing more than 300 000 spectroscopic reference objects, we estimate precise CZ for the KiDS-1000 tomographic redshift bins. The wide-area spec-z reference samples GAMA, BOSS, 2dFLenS, and WiggleZ are used to estimate CZ for the first three tomographic bins with a photo-z range of $0.1 < z_B \leq 0.7$, whereas the deep pencil-beam surveys zCOSMOS, DEEP2, and VVDS are used for the two high-redshift bins ($0.7 < z_B \leq 1.2$). This yields a homogeneous S/N of cross-correlation amplitudes as a function of redshift, which is important for the unbiased interpretation of the results.

The same analysis is replicated on mock catalogues based on the MICE simulation to identify and estimate systematic uncertainties (with the caveat that mock galaxies are only available for $z < 1.4$). Using the SOM $n(z)$ as a model we fit for residual biases in these primary estimates of the KiDS-1000 redshifts. We find

no significant bias in any of the five tomographic bins, neither in the simulated analysis nor on the real KiDS data. The combined uncertainties that are associated with this validation method are at most a factor ∼2 larger than the ones estimated for the SOM. As these numbers include the SOM uncertainty (the SOM $n(z)$ are used as a model after all), the CZ method is shown to be fully competitive here.

The most important systematic errors to account for in a CZ analysis are the evolution of the galaxy bias, the non-trivial combination of surveys with different redshift range, density, spatial overlap, and – connected to this – model bias from the interpretation of the results with an imperfect model. We mitigate all of these effects and estimate residual systematic uncertainties by analysing a variety of alternative choices for basic analysis parameters (radial measurement scales, covariance estimate, data selection, bias model). These systematic uncertainties are fully propagated into the final CZ results, which constitute an alternative estimate of the KiDS-1000 $n(z)$ to be used in upcoming cosmological measurements.

The work presented here indicates a clear way forward to reach the stringent requirements of stage-IV WL surveys like *Euclid* (Laureijs et al. 2011), LSST (Ivezić et al. 2019), and RST (Spergel et al. 2015). Given sufficient deep, multi-band photometry, the SOM method allows for a robust gold selection whose accuracy is ultimately only limited by shot noise and competing requirements on the number density of the gold samples. Spectroscopic campaigns like the C3R2 (Masters et al. 2017, 2019; Euclid Collaboration 2020) will push the envelope and allow for increasingly inclusive gold selections with the SOM at further-reduced uncertainties in the redshift distributions.

An interesting addition to these purely spectroscopic approaches to colour-based calibration is offered by the inclusion of high-quality photo-z, not as the single calibration source but as a complement to the spectroscopic calibration data already present in the SOM. Multi-wavelength campaigns like the ones in COSMOS (Ilbert et al. 2009, 2013; Laigle et al. 2016) can yield exquisite redshift estimates with close to spectroscopic quality but without the drawback of incompleteness. Even better precision can be obtained from intermediate- or narrow-band surveys such as PAUS (Padilla et al. 2019; Eriksen et al. 2019) and J-PAS (Benitez et al. 2014), at least at brighter magnitudes. A smart combination of these surveys with the more traditional spectroscopic reference samples in a colour-based calibration like the SOM will mitigate the individual weaknesses of these catalogues and leverage their complementary advantages.

The future of the CZ technique looks similarly bright. Most of the limiting systematic effects seem to be understood by now and mitigation techniques have been established. The interpretation with a suitable model and subsequent estimation of realistic uncertainties is currently the biggest methodological problem to overcome. On the data side, the redshift range covered by wide-area surveys is still not sufficient to leverage the full potential of CZ. Currently, only the cores of the redshift distributions of typical weak lensing source samples can be calibrated with CZ. But with the advent of new spectroscopic facilities like DESI[5] (DESI Collaboration 2016), 4MOST[6] (Richard et al. 2019), WEAVE[7] (Dalton 2016), and PFS[8] (Takada et al. 2014) this situation will improve and the crucial calibration of high-redshift tails will become possible at high precision.

All of these data-related efforts need to be accompanied by improved mock catalogues and better theoretical understanding. In terms of mocks, larger volumes, higher redshifts, even more realistic galaxy colours, and a realistic integration of galaxy colours and shapes is needed. On the theoretical side, the standard practice in the analysis of weak lensing surveys regards the work presented here as calibration steps that are carried out before the main cosmological inference. In the future, this clear distinction could be broken up, with parts or all of this calibration being integrated into the inference pipeline itself (Bernstein 2009). This is more obvious for CZ, which represents "just another two-point function" to model and fit, but such an integration can also be imagined for the colour-based calibration approach. While systematic error control is an issue in such integrated approaches, the optimal use of information in the data for example through Bayesian-hierarchical modelling (Sánchez & Bernstein 2019; Alarcon et al. 2020) makes this idea extremely attractive for established methods that have left the exploratory stage.

The work presented here means that the KiDS-1000 cosmological analyses based on these weak lensing source samples will not be limited in their statistical power by the uncertainties in the redshift distributions. The constant progress and the developments sketched above make it seem realistic to meet the extremely tight requirements on the redshift calibration for stage-IV surveys a few years from now, a situation that seemed almost inconceivable not too long ago.

Acknowledgements. We are grateful to the anonymous referee for some suggestions that improved the paper. We are indebted to the staff at ESO-Garching and ESO-Paranal for managing the observations at VST and VISTA that yielded the data presented here. Based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017, 177.A-3018, 179.A-2004, 298.A-5015, and on data products produced by the KiDS consortium. The 2dFLenS survey is based on data acquired through the Australian Astronomical Observatory, under program A/2014B/008. It would not have been possible without the dedicated work of the staff of the AAO in the development and support of the 2dF-AAOmega system, and the running of the AAT. GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. 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/. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the US Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org. We are grateful to the zCOSMOS team to give us early access to additional deep spec-z that were not available in the public domain. zCOSMOS is based on observations made with ESO Telescopes at the La Silla or Paranal Observatories under programme ID 175.A-0839. This research has made use of the zCOSMOS database, operated at CeSAM/LAM, Marseille, France. Funding for the DEEP2 Galaxy Redshift Survey has been provided by NSF grants AST-95-09298, AST-0071048, AST-0507428, and AST-0507483 as well as NASA LTSA grant NNG04GC89G. This research uses data from the VIMOS VLT Deep Survey, obtained from the VVDS database operated by Cesam, Laboratoire d'Astrophysique de Marseille, France. We acknowledge support from European Research Council grants 770935 (HH, JLvdB, AHW, AD) and 647112 (CH, MA, BG, CL, TT), the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 797794 (TT), as well as the Deutsche Forschungsgemeinschaft (HH, Heisenberg grant Hi 1495/5-1). MB is supported by the Polish Ministry of Science and Higher Education through grant DIR/WK/2018/12, and by the Polish National Science Center through grant no. 2018/30/E/ST9/00698. CH acknowledges 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. KK acknowledges support by the Alexander von Humboldt Foundation. HYS acknowledges the support from NSFC of China under grant 11973070, the Shanghai Committee of Science and Technology grant No.19ZR1466600 and Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-7013. JTAdJ is supported by the Netherlands Organisation for Scientific Research (NWO) through grant 621.016.402. Author

[5] Dark Energy Spectroscopic Instrument; www.desi.lbl.gov
[6] www.4most.eu
[7] www.ing.iac.es/weave/
[8] Subaru Prime Focus Spectrograph; https://pfs.ipmu.jp

Contributions: All authors contributed to the development and writing of this paper. The authorship list is given in three groups: the lead authors (HH, JLvdB, AHW), 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

Aihara, H., Armstrong, R., Bickerton, S., et al. 2018, PASJ, 70, S8
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
Asgari, M., Heymans, C., Hildebrandt, H., et al. 2019, A&A, 624, A134
Asgari, M., Lin, C.-A., Joachimi, B., et al. 2021, A&A, 645, A104
Baldry, I. K., Liske, J., Brown, M. J. I., et al. 2018, MNRAS, 474, 3875
Balestra, I., Mainieri, V., Popesso, P., et al. 2010, A&A, 512, A12
Bartelmann, M., & Schneider, P. 2001, Phys. Rep., 340, 291
Benítez, N. 2000, ApJ, 536, 571
Benitez, N., Dupke, R., Moles, M., et al. 2014, ArXiv e-prints [arXiv:1403.5237]
Bernstein, G. M. 2009, ApJ, 695, 652
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
Buchs, R., Davis, C., Gruen, D., et al. 2019, MNRAS, 489, 820
Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., & Fosalba, P. 2015, MNRAS, 447, 646
Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., & Carretero, J. 2015, MNRAS, 453, 1513
Cunha, C. E., Lima, M., Oyaizu, H., Frieman, J., & Lin, H. 2009, MNRAS, 396, 2379
Dalton, G. 2016, in Multi-Object Spectroscopy in the Next Decade: Big Questions, Large Surveys, and Wide Fields, eds. I. Skillen, M. Balcells, & S. Trager, ASP Conf. Ser., 507, 97
Davis, C., Gatti, M., Vielzeuf, P., et al. 2017, ArXiv e-prints [arXiv:1710.02517]
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., Kuijken, K., Applegate, D., et al. 2013, The Messenger, 154, 44
de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., et al. 2017, A&A, 604, A134
DESI Collaboration (Aghamousa, A., et al.) 2016, ArXiv e-prints [arXiv:1611.00036]
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, The Messenger, 154, 32
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Eriksen, M., Alarcon, A., Gaztanaga, E., et al. 2019, MNRAS, 484, 4200
Euclid Collaboration (Guglielmo, V., et al.) 2020, A&A, 642, A192
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
Giblin, B., Heymans, C., Asgari, M., et al. 2021, A&A, 645, A105
Gruen, D., & Brimioulle, F. 2017, MNRAS, 468, 769
Heydenreich, S., Schneider, P., Hildebrandt, H., et al. 2020, A&A, 634, A104
Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012, MNRAS, 427, 146
Heymans, C., Tröster, T., Asgari, M., et al. 2021, A&A, 646, A140
Hildebrandt, H., Pielorz, J., Erben, T., et al. 2009, A&A, 498, 725
Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454
Hildebrandt, H., Köhlinger, F., van den Busch, J. L., et al. 2020, A&A, 633, A69
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
Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, MNRAS, 366, 101
Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236
Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jain, B., & Taylor, A. 2003, Phys. Rev. Lett., 91, 141302
Joachimi, B., Lin, C. A., Asgari, M., et al. 2021, A&A, 646, A129
Johnson, A., Blake, C., Amon, A., et al. 2017, MNRAS, 465, 4118
Kafle, P. R., Robotham, A. S. G., Driver, S. P., et al. 2018, MNRAS, 479, 3746
Kitching, T. D., Viola, M., Hildebrandt, H., et al. 2015, ArXiv e-prints [arXiv:1512.03627]
Kohonen, T. 1982, Biol. Cybern., 43, 59
Kuijken, K. 2008, A&A, 482, 1053
Kuijken, K., Heymans, C., Dvornik, A., et al. 2019, A&A, 625, A2
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
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
Le Fèvre, O., Tasca, L. A. M., Cassata, P., et al. 2015, A&A, 576, A79
Lilly, S. J., Le Fèvre, O., Renzini, A., et al. 2007, ApJS, 172, 70
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
Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53
Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2017, ApJ, 841, 111
Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2019, ApJ, 877, 81
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]
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
Newman, J. A., Abate, A., Abdalla, F. B., et al. 2015, Astropart. Phys., 63, 81
Padilla, C., Castander, F. J., Alarcón, A., et al. 2019, AJ, 157, 246
Planck Collaboration XIII. 2016, A&A, 594, A13
Popesso, P., Dickinson, M., Nonino, M., et al. 2009, A&A, 494, 443
Richard, J., Kneib, J. P., Blake, C., et al. 2019, The Messenger, 175, 50
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, P. 2016, A&A, 592, L6
Schneider, M., Knox, L., Zhan, H., & Connolly, A. 2006, ApJ, 651, 14
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
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, ArXiv e-prints [arXiv:1503.03757]
Stölzner, B., Joachimi, B., Korn, A., Hildebrandt, H., & Wright, A. H. 2020, A&A, submitted [arXiv:2012.07707]
Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810
Takada, M., Ellis, R. S., Chiba, M., et al. 2014, PASJ, 66, R1
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, PASJ, 70, S9
Tröster, T., Asgari, M., Blake, C., et al. 2021, A&A, in press, https://doi.org/10.1051/0004-6361/202039805
van den Busch, J. L., Hildebrandt, H., Wright, A. H., et al. 2020, A&A, 642, A200
Vanzella, E., Cristiani, S., Dickinson, M., et al. 2008, A&A, 478, 83
Venemans, B. P., Verdoes Kleijn, G. A., Mwebaze, J., et al. 2015, MNRAS, 453, 2259
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. 2020a, A&A, 637, A100
Wright, A. H., Hildebrandt, H., van den Busch, J. L., et al. 2020b, A&A, 640, L14

Appendix A: Toy model for data with variable S/N

Here we illustrate with a simple toy model how variable S/N can bias a model fit. This situation is quite common in clustering-z measurements that typically exhibit a large number of reference galaxies at low redshift and a small number at high redshift due to the difficulties of measuring redshifts for high-z galaxies.

In Fig. A.1 we show a simulated data set that is based on a normal distribution with the S/N decreasing as a function of x. Fitting a model with a shift parameter and a free amplitude to different noise realisations yields the coloured lines. If the model has the correct width (i.e. standard deviation STD = 1) the model fits (blue lines) are on average unbiased as shown in the bottom panel. If the model is too narrow (STD < 1) the model fits (teal and orange) are biased low whereas if the model is too broad (STD > 1) the model fits (magenta and green) are biased high.

This observation led to our decision to analyse the clustering-z of the wide and deep fields separately. The number of reference galaxies in the two sets is just too different so that a sharp drop in S/N is observed at the transition redshift ($z \sim 0.8$). Analysing the wide and deep fields together and using the SOM (DIR) $n(z)$, whose widths are typically to small (large), would result in a similar model bias as shown in Fig. A.1.

One alternative would be to randomly subsample the Wide data to homogenise the S/N. We leave this idea to future work.

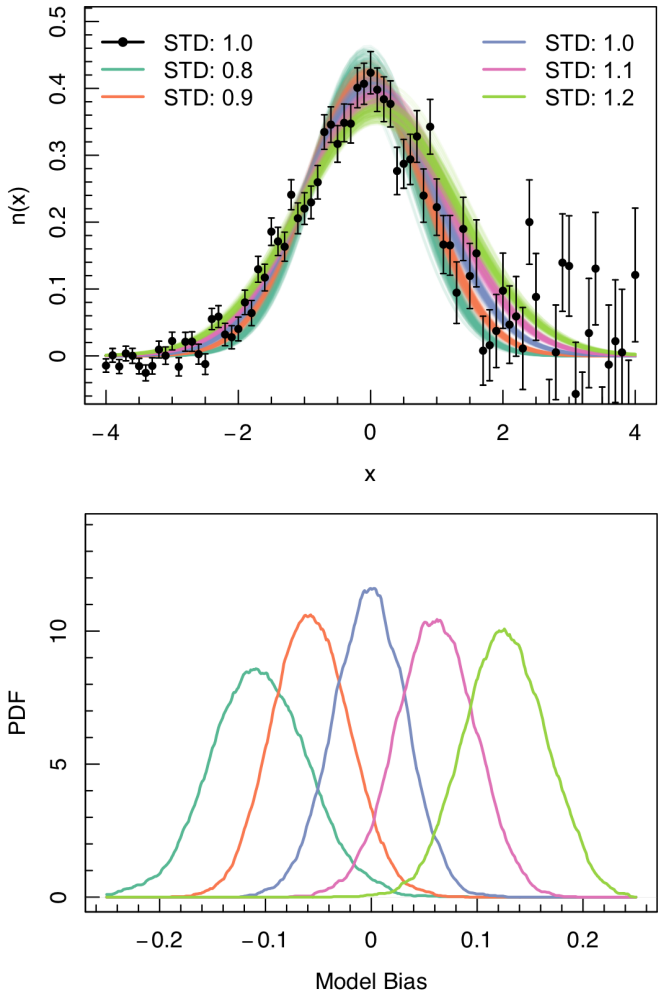

Fig. A.1. Toy model to illustrate the effect of variable S/N on model fits. *Top*: black data points correspond to one noise realisation with decreasing S/N. The blue lines correspond to fits with a model of perfect width whereas the teal and orange lines correspond to models that are to narrow and the magenta and green lines correspond to models that are too wide. *Bottom*: if the model has the correct width the mean of the best fit is on average unbiased (blue) whereas it is on average biased low if the model is too narrow (teal and orange) and biased high if the model is too broad (magenta and green).