Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **534,** 3808–3831 (2024) https://doi.org/10.1093/mnras/stae2243
Advance Access publication 2024 September 27

Improved weak lensing photometric redshift calibration via StratLearn and hierarchical modelling

Maximilian Autenrieth [1]★ Angus H. Wright [2] Roberto Trotta,[3,4,5,6]★ David A. van Dyk [1] David C. Stenning[7] and Benjamin Joachimi[8]

[1]*Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK*
[2]*Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing, 44780 Bochum, Germany*
[3]*SISSA–International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, Italy*
[4]*Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, SW72AZ, London, UK*
[5]*Italian Research Center on High Performance Computing, Big Data and Quantum Computing, Italy*
[6]*INFN–National Institute for Nuclear Physics, Via Valerio 2, 34127 Trieste, Italy*
[7]*Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, B.C., Canada*
[8]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, United Kingdom*

Accepted 2024 September 19. Received 2024 July 10; in original form 2023 December 12

ABSTRACT

Discrepancies between cosmological parameter estimates from cosmic shear surveys and from recent Planck cosmic microwave background measurements challenge the ability of the highly successful ΛCDM model to describe the nature of the Universe. To rule out systematic biases in cosmic shear survey analyses, accurate redshift calibration within tomographic bins is key. In this paper, we improve photo-z calibration via Bayesian hierarchical modeling of full galaxy photo-z conditional densities, by employing *StratLearn*, a recently developed statistical methodology, which accounts for systematic differences in the distribution of the spectroscopic training/source set and the photometric target set. Using realistic simulations that were designed to resemble the KiDS + VIKING-450 data set, we show that *StratLearn*-estimated conditional densities improve the galaxy tomographic bin assignment, and that our *StratLearn*-Bayesian framework leads to nearly unbiased estimates of the target population means. This leads to a factor of ~ 2 improvement upon often used and state-of-the-art photo-z calibration methods. Our approach delivers a maximum bias per tomographic bin of $\Delta\langle z\rangle = 0.0095 \pm 0.0089$, with an average absolute bias of 0.0052 ± 0.0067 across the five tomographic bins.

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

1 INTRODUCTION

Cosmological parameter estimation from the cosmic microwave background (CMB; Planck Collaboration 2020) and from tomographic cosmic shear measurements (e.g. Asgari et al. 2021; Abbott et al. 2022; Sugiyama et al. 2023) lead to discrepancies in the estimated clustering strength of dark matter (see Abdalla et al. 2022 for a recent review on cosmic tensions). Such systematic discrepancies could challenge the highly successful dark energy and cold dark matter paradigm (ΛCDM) in describing the true nature of the Universe. Of course, such a claim needs critical and detailed consideration of the surveys and analysis steps performed by the various collaborations to rule out systematic biases in the different procedures, which might explain said discrepancies.

Extensive explorations of various survey, model, and analysis modifications have been performed in recent cosmological interpretations of the current generation of cosmic shear surveys, the Dark Energy Survey (DES), the Hyper-Suprime Camera (HSC) survey, and the Kilo-Degree Survey (KiDS). These highlight that inaccuracies and statistical uncertainty in the line-of-sight distribution of galaxies as determined by photometric redshifts can limit, and potentially bias, constraints on cosmological parameters (cf. Troxel et al. 2018a, b; Hikage et al. 2019; Joudaki et al. 2020; Asgari et al. 2021; Amon et al. 2022; Secco et al. 2022; Dark Energy Survey and Kilo-Degree Survey Collaboration et al. 2023; Rau et al. 2023).

In cosmic shear tomography, galaxies are assigned to (predefined) tomographic redshift bins (Hu 1999) based on an estimate of their photometric redshift (photo-z). For recent reviews on photometric redshift estimation and its application in large galaxy surveys see Salvato, Ilbert & Hoyle (2019) and Newman & Gruen (2022), respectively. If the estimated population redshift distribution (in a tomographic bin) differs systematically from the (non-observable) true redshift distribution, the parameter estimates from cosmic shear tomography might be systematically biased.

★ E-mail: m.autenrieth19@imperial.ac.uk (MA); rtrotta@sissa.it (RT)

For instance, if the estimated redshift distribution is systematically lower than it is in reality, then observed gravitational distortions are attributed to an overly dense and too highly clustered matter distribution. It is thus essential to obtain accurate redshift distribution estimates. In particular, it is crucial to obtain an unbiased estimate of the first moment (mean) of the true underlying redshift population distribution (per tomographic bin), in order to avoid such systematic biases in the analysis (Amara & Refregier 2008; Reischke 2024). This is because the accuracy of cosmic shear cosmological measurements is highly dependent on the accuracy of the first moment of the binned redshift population distributions, but much less sensitive to the higher-order moments: Reischke (2024), for example, demonstrates that a one-sigma shift in the desired cosmological model parameters (for a *Euclid*-like survey) is induced when the first moment of the redshift distributions is mis-specified at the level of < 1 per cent. Conversely, a similar bias is only introduced for the second-order moment with a ~ 10 per cent mis-specification, and all higher-order moments can be essentially ignored (Reischke 2024, Fig. 2).

Several redshift calibration methods have been investigated to improve cosmic shear tomography. Wright et al. (2019) group these approaches into three categories:

(i) cross-correlation with reference galaxy samples that have precise and accurate redshifts (Schneider et al. 2006; Newman 2008; McQuinn & White 2013; Morrison et al. 2017). This strategy aims to constrain the photometric redshift population distribution by using spatial cross-correlations between the spectroscopic reference sample (with accurate redshift) and the photometric target sample (without accurate redshift). For each tomographic bin, the photometric redshift distribution is reconstructed by cross-correlating spectroscopic samples selected within thin redshift slices with the photometric samples (Gatti et al. 2018, 2022; Rau et al. 2022). More recently, Bayesian hierarchical frameworks have successfully been adopted to improve photometric redshift population estimates (Leistedt, Mortlock & Peiris 2016; Jones & Heavens 2019; Sánchez & Bernstein 2019; Alarcon et al. 2020; Rau, Wilson & Mandelbaum 2020; Gatti et al. 2022; Rau et al. 2022, 2023), allowing the combination of cross-correlation with template fitting and/or empirical approaches (Tanaka et al. 2018; Rau et al. 2022, 2023).

(ii) stacking of individual galaxy redshift distributions, as adopted by Hildebrandt et al. (2012), Hoyle et al. (2018), Tanaka et al. (2018), Hamana et al. (2020), and Malz & Hogg (2022), and lastly;

(iii) direct redshift calibration (Lima et al. 2008; Hildebrandt et al. 2016, 2020; Buchs et al. 2019; Wright et al. 2020). The idea of direct redshift calibration is to reweight the distribution of spectroscopic redshift, obtained only for a small and non-representative subsample of the data (source/training data), to match the distribution of the photometric target data. In recent work, Masters et al. (2015), Buchs et al. (2019), and Wright et al. (2020, hereafter W20) develop direct redshift calibration methods based on self organizing maps (SOM; Kohonen 1982). In W20, their implementation of SOM-based direct calibration is shown to outperform previously proposed methods on comprehensive simulations designed to realistically resemble the KIDS + VIKING-450 data set (Wright et al. 2019, W20; Hildebrandt et al. 2020). Their (and previous) methods obtain a tomographic bin assignment via a Bayesian-Photometric-Redshift estimate (Benitez 2000), further denoted as z_B, calculated for each galaxy. While improving on other direct redshift calibration methods, the SOM method proposed in W20 still leads to potentially concerning bias in some tomographic bins, and mitigates these biases by introducing additional systematic selections to the data. Such selections lead to fewer sources available for science, and therefore constitute a source of increased statistical uncertainty in down-stream cosmological analyses.

In addition, several methods have been employed to assign galaxies to tomographic (typically four or five non-overlapping) bins. The common approach is to group galaxies based on a choice of point estimate of redshift, e.g. from template fitting or machine learning codes. One approach is to employ SOM-based bin assignment by assigning galaxies to tomographic bins with adaptive bin edges based on the SOM cell assignment of the galaxies (Buchs et al. 2019; Alarcon et al. 2020; Myles et al. 2021; Gatti et al. 2022; Secco et al. 2022). Another approach is to assign galaxies to tomographic bins according to a Directional Neighbourhood Fitting (DNF) photo-z estimate (Gatti et al. 2018; Abbott et al. 2022). DNF is a machine learning method that obtains photo-z estimates based on the neighbourhood of galaxies in a multiband flux space (De Vicente, Sánchez & Sevilla-Noarbe 2016; Gatti et al. 2018). More recently, Rau et al. (2023) employ a neural network-based photometric redshift conditional density code (DNNz) to bin galaxies within four tomographic redshift intervals. Rau et al. (2023) identify regions of the data space that are difficult to calibrate and remove some of the galaxies based on differences in the estimates of DNNz and an SED template fitting approach. Others employ photometric redshifts estimated using the Bayesian-Photometric-Redshift code (BPZ; Benitez 2000), which constructs a posterior probability distribution of redshift given a source's observed photometry. This code produces a posterior mode point-estimate of photometric redshift, z_B, which is subsequently used for tomographic binning (Hoyle et al. 2018; Hartley et al. 2020; Hildebrandt et al. 2020; Van Den Busch et al. 2020; Wright et al. 2020; Asgari et al. 2021).

In this paper, we propose a different strategy to improve redshift calibration, based on galaxy (object level) conditional photo-z density estimates. More precisely, we employ a recently proposed statistical method, *StratLearn* (Autenrieth et al. 2024), that allows principled photo-z conditional density estimation under non-representative source/training data. *StratLearn* alleviates (or bypasses) the problem of non-representative source/training data (identified as covariate shift), by subgrouping the data into strata based on estimated propensity scores, a pivotal methodology in causal inference (Rosenbaum & Rubin 1983). In our context, the propensity score is the probability of a galaxy being assigned to the spectroscopic training/source set given the observed covariates (i.e. photometric magnitudes/colours). Autenrieth et al. (2024) demonstrate that fitting conditional density estimators within strata, constructed by partitioning the data based on the estimated propensity scores, improves full conditional photo-z density estimates under non-representative source/training data. Here, we show that the *StratLearn* conditional densities[1] can be used directly to improve the tomographic bin assignment, by assigning each galaxy to the tomographic bin with its highest conditional probability. In a second step, we construct a Bayesian hierarchical framework to model summaries of each galaxy's conditional density (within tomographic bins), leading to nearly unbiased estimates of the mean redshift of each tomographic bin. We evaluate our novel *StratLearn*-Bayes approach on comprehensive simulations

[1] *StratLearn* is a general-purpose statistical method for learning under covariate shift. While Autenrieth et al. (2024) show the effectiveness of conditional density estimation within the *StratLearn* framework, the conditional density estimators themselves are not part of the *StratLearn* methodology, and have been proposed elsewhere (Izbicki & Lee 2016; Izbicki et al. 2017). For conciseness, we loosely refer to the conditional density estimates as '*StratLearn* conditional densities'.

constructed by W20, demonstrating a substantial reduction of bias compared to the previously proposed SOM calibration method.

While the primary sensitivity of cosmic shear is to the first moment of the redshift distribution, other cosmological probes, which also require redshift distribution estimation and calibration, are more sensitive to the accurate recovery of higher-order redshift distribution moments. Reischke (2024) shows that higher-order moments have much more influence on bias in an analysis of photometric galaxy clustering. Additionally, cosmic shear surveys will become increasingly sensitive to higher-order moments with increasing statistical power. As such, it is sensible to consider how we can best estimate the full redshift distribution. While our *StratLearn*-Bayes method is specifically designed to obtain accurate estimates of the first moments of the redshift distributions, we demonstrate how estimated propensity scores can be used in a direct redshift calibration scheme to obtain accurate estimates of the redshift population distribution shapes.

The remainder of the paper is structured as follows. In Section 2.1, we specify notation and we formally introduce the underlying covariate shift scenario, arising through the non-representativeness of the training/source data. In Section 2.2, we summarize the direct redshift calibration method. We then briefly introduce the supervised learning task with a focus on conditional density estimation under the covariate shift scenario. In Section 3, we formally introduce our approach. In Section 3.1, we provide a detailed description of how we estimate conditional densities under covariate shift within *StratLearn*. We then specify how these conditional densities can be used to improve galaxy tomographic bin assignment (Section 3.2). In Section 3.3, we demonstrate how summaries of the estimated conditional densities can be employed in a Bayesian hierarchical framework to accurately estimate the redshift population means (within a tomographic bin). In Section 4, we demonstrate how inverse propensity score weighting (inverse-PS) can be employed to estimate the redshift population shapes for each tomographic bin. In Section 5, we present numerical evaluation of our method. We first introduce the simulation setting in Section 5.1. We then evaluate our new bin assignment with a comparison to previously used z_B bin assignment (Section 5.2). We present our redshift calibration results in Section 5.3, and illustrate the inverse-PS population distribution estimates in Section 5.5. Finally, in Section 6, we conclude with a discussion of our findings, limitations, and implications for future weak lensing survey analyses.

2 ADDRESSING NON-REPRESENTATIVE TRAINING DATA

2.1 Non-representative spectroscopic data and covariate shift

Let z_i be the true spectroscopic redshift of galaxy i, and x_i be the vector of its observed photometric magnitudes/colours (the exact choice of covariates is described in Section 5.1.3). In a cosmic shear analysis, we have access to a relatively small set of galaxies with measured spectroscopic redshift, since obtaining spectroscopy for millions of objects is observationally expensive (over the magnitude range in question). For our purposes, spectroscopically measured redshifts can be considered equal to the true redshift. We denote this spectroscopic set as source (or training) data $D_S = \{(x_S^{(i)}, z_S^{(i)})\}_{i=1}^{n_S}$, with n_S galaxies sampled at random from the joint distribution $p_S(x, z)$. The so-called photo-z estimation problem (Hildebrandt et al. 2010; Freeman, Izbicki & Lee 2017; Izbicki et al. 2017; Dey et al. 2022) is to find a redshift estimate that can be deployed on a much larger set of galaxies, for which only the photometric data x_T are available, but not spectroscopically measured redshifts, z_T. We denote this photometric set as our target data $D_T = \{x_T^{(i)}\}_{i=1}^{n_T}$, with n_T unlabelled samples from the joint distribution $p_T(x, z)$ (with $n_T \gg n_S$). The problem is compounded by the fact that $p_S(x, z) \neq p_T(x, z)$, i.e. the spectroscopic source and photometric target distributions differ systematically due to selection effects in the acquisition of spectroscopy based on characteristics of the photometric magnitudes, leading to $p_S(x) \neq p_T(x)$. We assume, however, that the conditional distributions of redshift z given the magnitudes x are the same in spectroscopic source and photometric target data, i.e. $p_S(z|x) = p_T(z|x)$. The situation, in which $p_S(z|x) = p_T(z|x)$ but $p_S(x) \neq p_T(x)$, is called 'covariate shift' in the statistical learning literature (Moreno-Torres et al. 2012). If such covariate shift is not accounted for, machine learning or other statistical methods that aim to learn the relationship between the covariates and redshift can perform poorly; the training set is not representative of the target/test, meaning that patterns learned from the training set are not generalizable.

The covariate shift assumption has been frequently (sometimes implicitly) made in previous photo-z calibration work, (e.g. Lima et al. 2008; Hildebrandt et al. 2020; Wright et al. 2020). Others (e.g. Hartley et al. 2020 and Newman et al. 2015) argue that redshift failures, the use of quality flags based on galaxy spectral characteristics to address these failures, and selecting data based on these flags may result in the violation of the covariate shift assumption. We provide additional discussion on the matter in Appendix B.

2.2 Direct redshift calibration

Since in the covariate shift scenario $p_S(z|x) = p_T(z|x)$ but $p_S(x) \neq p_T(x)$, it generally follows that the redshift distribution of the spectroscopic set differs from that of the target, $p_S(z) \neq p_T(z)$. Direct redshift calibration methods reweight the spectroscopic redshift sample to match the photometric redshift distribution (Lima et al. 2008).

More precisely, under the covariate shift scenario, it holds that

$$p_T(z, x) = p_T(z|x)p_T(x) \tag{1}$$

$$= p_S(z|x)p_T(x) \tag{2}$$

$$= p_S(z, x)\frac{p_T(x)}{p_S(x)} \tag{3}$$

That is, one can express the joint target distribution by reweighting the joint source distribution. Precisely, $p_T(z, x) = \omega(x)p_S(z, x)$, with weights $\omega(x) = p_T(x)/p_S(x)$. In practice, one can reweight galaxies in the spectroscopic source set [with weights $\omega(x)$] to obtain a representative sample of the joint target distribution. In principle, looking at the marginal sample of z in the weighted joint distribution thus provides us a consistent estimate of the target redshift distribution $p_T(z)$.

Accurate estimation of the weights $\omega(x)$ is key for direct redshift calibration methods. Lima et al. (2008) and Hildebrandt et al. (2020) implement a k-nearest-neighbour (kNN) method for weight estimation. W20 demonstrate improvement over the kNN method by computing the weights via an SOM method, a form of unsupervised neural network which can map a high-dimensional covariate space to a lower-dimensional grid.

Unfortunately, weighting methods typically entail high variance, particularly if there is a small number of objects with very large weights. In addition, finding a suitable set of weights ω is not trivial,

but key for direct calibration methods. Noisy, inaccurate weights might lead to potentially severe bias and strongly increased variance.

2.3 Photometric redshift regression

Instead of reweighting, our approach uses the labelled spectroscopic source data as a training set to fit supervised full conditional density models. Our trained models then deliver a non-parametric estimate of the full conditional redshift (photo-z) distribution for each galaxy in the photometric target data (conditional on its observed covariates), $\hat{f}(z|x)$. If source and target data follow the same distribution, conditional density estimators aim to minimize the *generalized* risk under the L^2-loss (generalized in that the underlying loss can be negative), given by:

$$\hat{R}_S(\hat{f}) = \frac{1}{n_S}\sum_{k=1}^{n_S}\int \hat{f}^2\left(z|x_S^{(k)}\right)\mathrm{d}z - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\hat{f}\left(z_S^{(k)}|x_S^{(k)}\right). \quad (4)$$

(see Section A for the derivations of 4 and Izbicki et al. 2017 for further details). To provide intuition for the *generalized* risk in (4), note that, the second term of (4) averages the values of the conditional density estimates at the true spectroscopic redshift (known for the source data); this is optimized if the true redshift is at (or close to) the mode of the conditional density estimate $\hat{f}(z_i|x_i)$, with $\hat{f}(z_i|x_i)$ being very tall and narrow (the Dirac delta distribution at the true redshift value is the optimal limiting case). In contrast, the first term of (4), which integrates the squared conditional density estimates over the redshift range (without information of the true redshift), is minimized for wide and (nearly) uniform conditional density estimates, thus penalizing highly localized predictions. Thus, overall estimates that are very certain (i.e. low variance), but fail to cover the truth lead to a high risk.

In the presence of covariate shift, however, obtaining accurate target estimates requires minimization of the target risk $\hat{R}_T(\hat{f})$, which is obtained by replacing all source samples in (4) with target samples, which typically means $\hat{R}_S(\hat{f}) \neq \hat{R}_T(\hat{f})$. The challenge is to minimize $\hat{R}_T(\hat{f})$ without access to the target true redshift z_T. In the next section, we provide a summary of our approach, called *StratLearn* (Autenrieth et al. 2024), which allows minimization of $\hat{R}_T(\hat{f})$ under the covariate shift scenario.

3 BAYESIAN PHOTOMETRIC REDSHIFT CALIBRATION VIA STRATLEARN

3.1 Photo-z conditional densities within *StratLearn*

StratLearn allows target risk minimization by subgrouping the source and target data into strata based on estimated propensity scores. Within strata, the joint distribution of target data and source data is approximately the same, and target risk can thus be minimized via source risk minimization. In the following, we provide a detailed description of the procedure.

Let S be a binary indicator variable, with $s_i = 1$ indicating the assignment of galaxy i to the spectroscopic source set ($s_i = 0$ indicates assignment to the photometric target set). In the context of this paper, the propensity score is the probability of a galaxy i being in the spectroscopic source data, given its observed covariates (photometry) x_i, i.e.

$$e(x_i) := P(s_i = 1|x_i), \text{ with } 0 < e(x_i) < 1. \quad (5)$$

In practice, we obtain an estimate $\hat{e}(x_i)$ of (5) via binary, probabilistic classification of source and target data using a logistic regression model with all the photometric magnitudes/colours as independent predictor variables (main effects) and the source/target set assignment variable S as the binary dependent variable. We then subgroup (stratify) the source and target data into $k = 5$ strata based on the quintiles of the estimated propensity score distribution $\hat{e}(x)$. The use of five strata is suggested by Autenrieth et al. (2024), based on numerical evidence provided by Cochran (1968) that subgrouping into $k = 5$ strata removes at least 90 per cent of the bias for many continuous distributions.

By Proposition 1 of Autenrieth et al. (2024), within strata,

$$p_{T_j}(z, x) \approx p_{S_j}(z, x), \text{ for } j \in 1, \ldots, k, \quad (6)$$

where S_j indicates conditioning on assignment to the jth source stratum (analogously for target T_j). It directly follows that $\hat{R}_{Tj}(\hat{f}) \approx \hat{R}_{Sj}(\hat{f})$ within strata $j \in 1, \ldots, k$. Thus, we can minimize the target risk $\hat{R}_{Tj}(\hat{f})$ within strata by minimizing the source risk $\hat{R}_{Sj}(\hat{f})$ within strata. See Autenrieth et al. (2024) for details.

Given the strata conditional on the estimated propensity score, we can now fit any supervized model on the spectroscopic source data within each stratum and predict on its respective photometric target stratum. As suggested in Autenrieth et al. (2024), within each strata, we employ a weighted average (convex combination) of two conditional density estimators: *ker-NN* (Izbicki et al. 2017) and *Series* (Izbicki & Lee 2016). The kernel nearest neighbour estimator (ker-NN) computes the conditional density of an object via a kernel smoothed histogram of the redshift of its k nearest neighbours in the respective source stratum. The spectral series estimator (Series) adapts a lower-dimensional subspace of the x-space as the intrinsic dimension of the data, based on data-dependent eigenfunctions of a kernel-based operator (Izbicki & Lee 2016). Details can be found in Izbicki et al. (2017) and Izbicki & Lee (2016). Previous studies (Izbicki et al. 2017; Autenrieth et al. 2024) indicate that each estimator appears to perform better in a different data regime, and combining them leads to a more robust estimator.[2]

We individually optimize the conditional density estimators (ker-NN, Series) by minimizing (4) in each source stratum separately. The final *StratLearn* conditional density estimate is obtained by combining the ker-NN and Series conditional density estimates $\hat{f}_{\text{ker-NN}}(z|x)$ and $\hat{f}_{\text{Series}}(z|x)$ by optimizing

$$\hat{f}(z|x) = (1-\alpha)\hat{f}_{\text{Series}}(z|x) + \alpha\hat{f}_{\text{ker-NN}}(z|x), \quad (7)$$

with $0 \leq \alpha \leq 1$. The parameter α is optimized to minimize the generalized risk

$$\hat{R}_{S2}(\hat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \hat{f}^2\left(z|x_T^{(k)}\right)\mathrm{d}z - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\hat{f}\left(z_S^{(k)}|x_S^{(k)}\right) \quad (8)$$

within each strata. We note that (8) only differs from (4) in that the first term is averaged over the photometric sample, x_T, rather than the spectroscopic sample, x_S, which does not require any target redshift z_T. Finally, (7) provides a galaxy-by-galaxy full conditional density redshift estimate $\hat{f}(z_i|x_i)$.[3] Some illustrative examples of the resulting galaxy conditional density estimates are shown in Fig. 1. Section 5.1.3 provides additional details on computation and parameter optimization of the conditional density estimators.

[2] We refer the interested reader to the ensemble learning literature for further background (Wolpert 1992; Van der Laan, Polley & Hubbard 2007; Naimi & Balzer 2018).

[3] Note that, for better readability in (4) and (8), we use superscripts (k) to enumerate objects, elsewhere we use subscripts i.

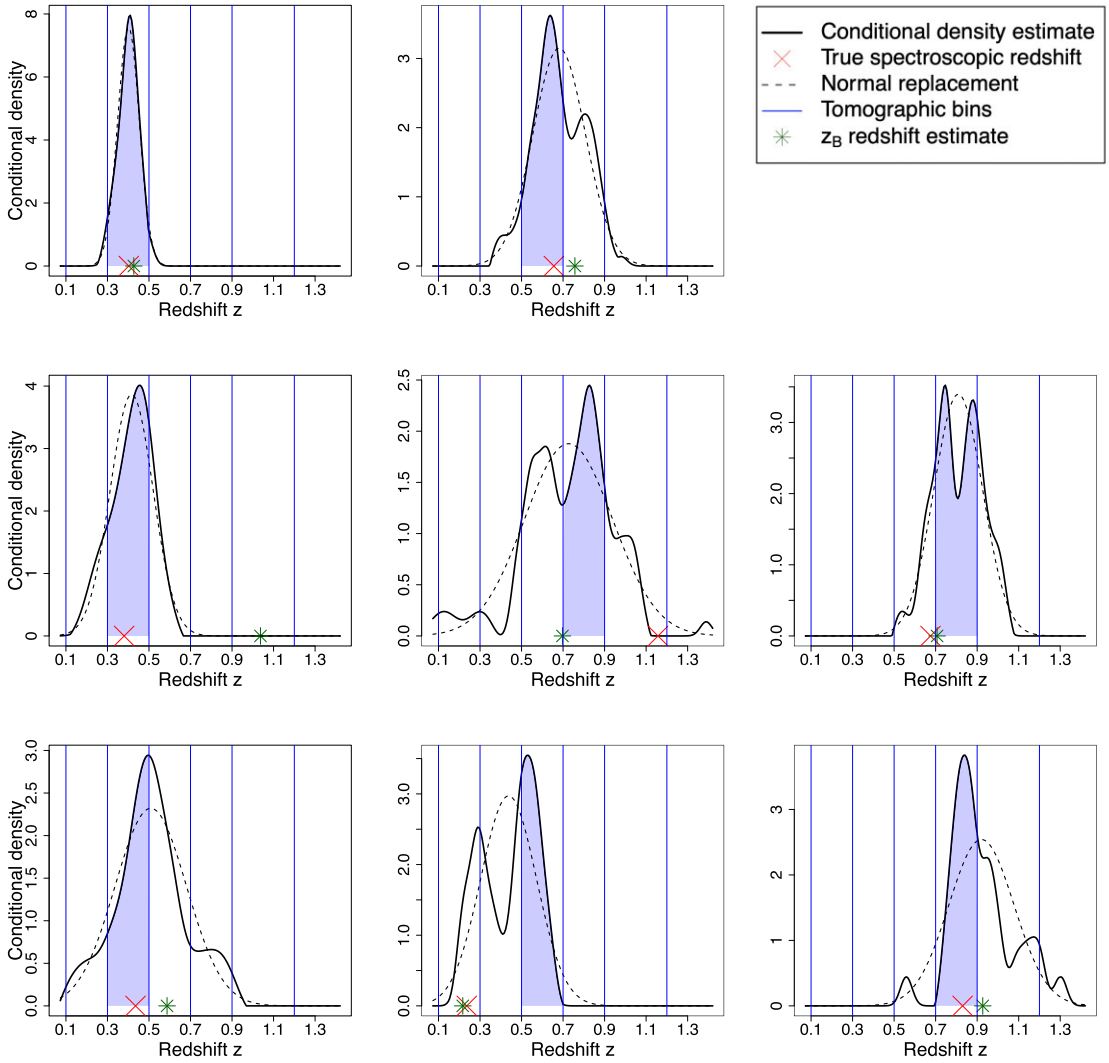


Figure 1. Examples of the conditional density estimates $\hat{f}(z_i|x_i)$, for galaxies in the photometric target samples, illustrated on the tomographic bin grid. The *StratLearn* assigned bin (the one containing the highest conditional probability) is shaded in blue. The true spectroscopic redshift is shown by the red cross. The z_B estimate is shown by the green star. A fraction of 30 per cent of the conditional density estimates appear to be roughly bell-shaped like in the top left example, but many conditional densities can be skewed and multimodal. Normal distributions with the same means and variances as the conditional density examples $\hat{f}(z_i|x_i)$ are added as dashed lines (as discussed in Section 3.3.).

3.2 Tomographic bin assignment

For cosmic shear analysis, the photometric galaxies are assigned to groups along the line-of-sight called tomographic bins. These bins are constructed with the best available proxy for the true line-of-sight distance of the galaxies. For wide-field photometric surveys, bins are typically constructed based on a redshift estimate determined from broad-band photometry (called the photometric redshift estimate or photo-z estimate).

The KiDS survey employs photometric redshifts estimated using the BPZ code (Benitez 2000). More precisely, based on the z_B point-estimate (i.e. the posterior mode of the BPZ photometric redshift posterior probability distributions), W20 assign the photometric galaxies to five non-overlapping top-hat photometric redshift bins: (0.1, 0.3], (0.3, 0.5], (0.5, 0.7], (0.7, 0.9], (0.9, 1.2]; we denote these ranges as bins 1 through 5, respectively. Galaxies with z_B estimates outside of the five bin ranges ($z_B \leq 0.1$ and $z_B > 1.2$) are discarded.

A central step of our proposed photo-z calibration method is the computation of galaxy-by-galaxy full conditional density redshift estimates $\hat{f}(z_i|x_i)$. Instead of relying on z_B, we can therefore use the full conditional density estimates $\hat{f}(z_i|x_i)$ to provide an alternative tomographic bin assignment for each galaxy i. A natural choice is to assign each galaxy to the bin which contains the highest conditional probability: let $b(i)$ be the bin assignment of galaxy i, then

$$b(i) = \text{argmax}_m \int_{B(m)} \hat{f}\left(z_i|x_i\right) \mathrm{d}z, \quad m = 1, \ldots, 5, l, r, \tag{9}$$

with $B(i)$, $i = 1, \ldots, 5$ specifying the five tomographic bin redshift ranges. $B(l) := \{z|z \leq 0.1\}$ and $B(r) := \{z|z > 1.2\}$ specify two end bins for galaxies outside of the bin ranges; galaxies assigned to the end bins are not used in the analysis. Fig. 1 shows examples of conditional density estimates, plotted on the tomographic redshift bin grid. The assigned bins $b(i)$ with highest conditional probabilities are shaded.

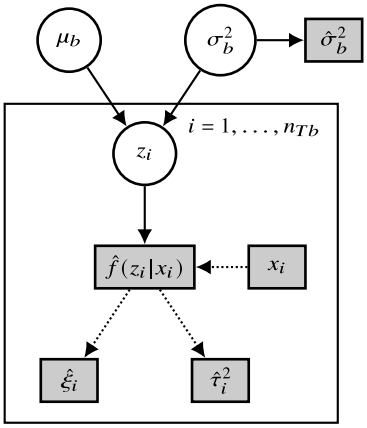


Figure 2. Graphical representation of our Gaussian hierarchical Bayesian model for the estimation of the redshift population mean in each tomographic bin, based on (summaries) of the photo-z conditional density estimates $\hat{f}(z_i|x_i)$. Observed quantities are illustrated in shaded squares. Unobserved parameters are illustrated via circles. Dashed arrows illustrate deterministic relations, and distributional relations are illustrated via solid arrows. We note that $\hat{\xi}_i$ and $\hat{\tau}_i^2$ are summary statistics derived from the conditional density estimates $\hat{f}(z_i|x_i)$.

3.3 Bayesian hierarchical modelling of conditional densities

In this section, we detail our Bayesian hierarchical framework for accurate estimation of the redshift population mean within each tomographic bin, given the object-level (galaxy) conditional density estimates. Employing a hierarchical Bayesian framework allows us to model the conditional density estimates in a statistically principled framework, with optimal shrinkage on the object-level photo-z estimates, allowing more precise population mean estimates. Fig. 2 provides an overview of our hierarchical Bayesian framework, with details described hereafter.

On the object (galaxy) level, $\hat{f}(z_i|x_i)$ is an estimate of the conditional density $p(z_i|x_i)$. Via Bayes theorem, the conditional density $p(z_i|x_i)$ can be expressed as

$$p(z_i|x_i) \propto p(x_i|z_i)p(z_i). \tag{10}$$

The estimation of the conditional densities $\hat{f}(z_i|x_i)$ is performed before and outside of the hierarchical Bayesian model fit (as described in Section 3.1) and without incorporation of prior information on the redshift distributions. By assuming a flat prior on z_i (e.g. a wide uniform prior that covers the expected photometric redshift range),[4] we have $p(z_i) \propto 1$. Then, (10) simplifies to

$$p(z_i|x_i) \propto p(x_i|z_i). \tag{11}$$

On the population level, recall that we aim to accurately estimate the population mean μ_b of z_i ($i = 1, \ldots, n_{Tb}$, with n_{Tb} being the number of galaxies within tomographic bin b). Accurate estimation of μ_b is crucial to avoid systematic biases in the downstream cosmic shear analysis (Amara & Refregier 2008; Reischke 2024). Thus, we model the redshift population within each bin with a normal distribution – a convenient choice that facilitates the introduction of a hierarchical Bayesian framework and a reasonable simplification given that we are primarily interested in the population mean. Specifically, at the redshift population level within bin b, we model

$$\text{Population Level:} \quad z_i|\mu_b, \sigma_b \overset{\text{indep.}}{\sim} N(\mu_b, \sigma_b^2), \tag{12}$$

with σ_b^2 being the redshift population variance.

Thus, we formulate the joint posterior distribution $p(z_1, \ldots, z_{n_{Tb}}, \mu_b, \sigma_b|\mathbf{X}_{\mathbf{n_{Tb}}})$, with $\mathbf{X}_{\mathbf{n_{Tb}}} := \{x_i\}_{i=1}^{n_{Tb}}$, via

$$\begin{aligned} &p(z_1, \ldots, z_{n_{Tb}}, \mu_b, \sigma_b|\mathbf{X}_{\mathbf{n_{Tb}}}) \\ &\quad \propto p(\mathbf{X}_{\mathbf{n_{Tb}}}|z_1, \ldots, z_{n_{Tb}}, \mu_b, \sigma_b)p(z_1, \ldots, z_{n_{Tb}}|\mu_b, \sigma_b)p(\mu_b, \sigma_b) \end{aligned} \tag{13}$$

$$= p(\mu_b, \sigma_b) \prod_i p(x_i|z_i)p(z_i|\mu_b, \sigma_b) \tag{14}$$

$$\propto p(\mu_b, \sigma_b) \prod_i p(z_i|x_i)p(z_i|\mu_b, \sigma_b), \tag{15}$$

where (13) to (14) holds due to the independence in (12) and the conditional independence $x_i \perp (\{z_j\}_{j\neq i}, \mu_b, \sigma_b)|z_i$. That is, given the redshift z_i for an object i, the distribution of its photometry x_i does not depend on other observed redshifts, nor on the parameters describing the population of redshift. (14) to (15) follows from (11). Since we are not targeting the object-level redshifts z_i themselves, but rather an accurate estimate of the population-level mean, μ_b, we integrate over the individual galaxies' redshifts to obtain the marginal posterior distribution

$$p(\mu_b, \sigma_b|\mathbf{X}_{\mathbf{n_{Tb}}}) \propto p(\mu_b, \sigma_b) \prod_i \int p(z_i|x_i)p(z_i|\mu_b, \sigma_b)\, \mathrm{d}z_i. \tag{16}$$

3.3.1 *Replacing $p(z_i|x_i)$ with a Gaussian approximation*

Although we could substitute the estimates, $\hat{f}(z_i|x_i)$, of $p(z_i|x_i)$ directly into (16), the required integrals would be computationally expensive. Instead, we simplify the problem by modelling each $p(z_i|x_i)$ with a normal distribution:

$$z_i|x_i \overset{\text{indep.}}{\sim} N(\hat{\xi}_i, \hat{\tau}_i^2), \tag{17}$$

where the estimate of the object-level mean $\hat{\xi}_i$ is simply the mean of the conditional density estimate, $\hat{f}(z_i|x_i)$, while the object-level Gaussian variance $\hat{\tau}_i^2$ is obtained by computing the variance of $\hat{f}(z_i|x_i)$. More precisely, by treating $\hat{f}(z_i|x_i)$ as a histogram evaluated on K bins, we have

$$\hat{\tau}_i^2 = \frac{1}{\sum_k \hat{f}_{m_k}(z_i|x_i)} \sum_k \hat{f}_{m_k}(z_i|x_i)(m_k - \hat{\xi}_i)^2, \tag{18}$$

where each $k = 1, \ldots, K$ specifies a histogram bin with location m_k, and $\hat{f}_{m_k}$ is the value of the conditional density (histogram) at location m_k.[5] The summary statistics $\hat{\xi}_i$ and $\hat{\tau}_i^2$ are observed quantities summarized by $\hat{\mathbf{X}}_{\mathbf{n_{Tb}}} := \{\hat{\xi}_i, \hat{\tau}_i\}_{i=1}^{n_{Tb}}$.

There are two reasons behind our replacement of the conditional density estimates $\hat{f}(z_i|x_i)$ by normal distributions. First, by modelling both the population- and object-level distributions as Gaussians, the Bayesian posterior distribution for the population mean can be calculated analytically. This allows us to scale our model to the large photometric data set at hand. Second, and more importantly, modelling the conditional densities as Gaussians leads to

[4] We note that in future work a more informative prior on the object level redshift distributions could principally be included via our hierarchical Bayesian model in (14).

[5] We note that by bins k (with locations $m(k)$), we refer to the density (histogram) bins and not the tomographic redshift bins.

almost unbiased estimates of the population means in all tomographic bins, as demonstrated in Section 5.3. In our simulation studies, we also investigate an alternative hierarchical model that uses the conditional densities $\hat{f}(z_i|x_i)$ directly (without Normal replacement). In this case, we obtained the posterior distributions via MCMC sampling (using 5 per cent of the target data due to computational limitations). Given the results from the subset, the Normal–Normal model led to better estimates of the population means than using the conditional density estimates directly. We refer to Appendix C4 for further details.

With this approximation, (16) can be written as

$$p(\mu_b, \sigma_b|\hat{\mathbf{X}}_{\mathbf{nTb}}) \propto p(\mu_b, \sigma_b) \prod_i \int N\left(z_i|\hat{\xi}_i, \hat{\tau}_i^2\right) \times N\left(z_i|\mu_b, \sigma_b^2\right) \mathrm{d}z_i, \tag{19}$$

where $N(t|\theta, \phi^2)$ is the probability density function (pdf) of a normal distribution with mean θ and variance ϕ^2, evaluated at t.

The integral in (19) can be solved analytically (see Appendix C for theoretical justifications) to obtain the (joint) marginal posterior density

$$p(\mu_b, \sigma_b|\hat{\mathbf{X}}_{\mathbf{nTb}}) \propto p(\mu_b, \sigma_b) \prod_i N\left(\hat{\xi}_i|\mu_b, \hat{\tau}_i^2 + \sigma_b^2\right). \tag{20}$$

Writing $p(\mu_b, \sigma_b) = p(\mu_b|\sigma_b)p(\sigma_b)$ and adopting a uniform conditional prior density $p(\mu_b|\sigma_b) \propto 1$, yields the conditional posterior distribution of μ_b given σ_b:

$$\mu_b|\sigma_b, \hat{\mathbf{X}}_{\mathbf{nTb}} \sim N(\tilde{\mu}_b, V_{\mu_b}), \tag{21}$$

with

$$\tilde{\mu}_b = \frac{\sum_i \frac{1}{\hat{\tau}_i^2+\sigma_b^2}\hat{\xi}_i}{\sum_i \frac{1}{\hat{\tau}_i^2+\sigma_b^2}} \quad \text{and } V_{\mu_b}^{-1} = \sum_i \frac{1}{\hat{\tau}_i^2 + \sigma_b^2}, \tag{22}$$

with $V_{\mu_b}^{-1}$ being the total precision.

Since we are not interested in the posterior uncertainty of σ_b, we choose an empirical Bayesian approach by setting σ_b to a fixed value estimated from the data, i.e. by choosing $p(\sigma_b) = \delta(\sigma_b - \hat{\sigma}_b)$. An obvious choice for the estimate $\hat{\sigma}_b$ is the MAP of the marginal posterior $p(\sigma_b|\hat{\mathbf{X}}_{\mathbf{nTb}})$ (shown in the Appendix 36). However, in our simulations, we found that the MAP estimate strongly and consistently underestimates σ_b. For this reason, we do not advocate the MAP estimate of σ_b and instead choose a different estimation strategy, as detailed below.

Finally, given an estimate of σ_b, an estimate of μ_b can be obtained analytically via (22), as $\tilde{\mu}_b$, the MAP estimate of μ_b.

3.3.2 *Population variance estimation via stacking of conditional densities*

Given the poor performance of the MAP estimate for σ_b, we instead estimate the population variance σ_b^2 via a 'stacked estimate' of the marginal redshift population distribution $p_b(z)$ of galaxies within tomographic bin b. More precisely, we obtain an estimate $\hat{p}_b^{\text{stack}}(z)$ of $p_b(z)$ by averaging (stacking) the conditional densities within bin, that is,

$$\hat{p}_b^{\text{stack}}(z) = \frac{1}{n_{Tb}} \sum_j \hat{f}(z_j|x_j). \tag{23}$$

with x_j, $j = 1, \ldots, n_{Tb}$, being the photometric magnitudes of the observed galaxies within tomographic bin b. While quite intuitive, the form of (23) is justified more formally in Section C3.

An estimate for the redshift population variance σ_b^2 can then be obtained by calculating the variance of $\hat{p}_b^{\text{stack}}(z)$, via

$$\hat{\sigma}_b^2 = \frac{1}{\sum_k \hat{p}_{b,m(k)}^{\text{stack}}(z)} \sum_k \hat{p}_{b,m(k)}^{\text{stack}}(z) \left(m(k) - \hat{\mu}_b^{\text{stack}}\right)^2, \tag{24}$$

where $k = 1, \ldots, K$ specifies the (density/histogram) bin with location m_k, $\hat{p}_{b,m(k)}^{\text{stack}}$ is the value of the stacked density (histogram) $\hat{p}_b^{\text{stack}}$ of bin b at location $m(k)$, and $\hat{\mu_b}^{\text{stack}}$ is the mean of $\hat{p}_b^{\text{stack}}(z)$. An estimate of the population standard deviation σ_b is then obtained by simply taking the square-root of (24).

We compare two versions of (24). First, we compute (24) via stacking over $\hat{f}(z_i|x_i)$, the galaxy conditional density estimates obtained via *StratLearn*. We denote this method as option *StratLearn*-Bayes (A). Second, we substitute the conditional density estimates $\hat{f}(z_i|x_i)$ by their normal replacements described in (17). We denote this option as *StratLearn*-Bayes (B).

4 ESTIMATING THE POPULATION DISTRIBUTION VIA INVERSE-PROPENSITY SCORE WEIGHTING

As we demonstrate below, our hierarchical Bayesian framework delivers highly accurate and precise estimates of the redshift means within each tomographic bin, the quantity of main interest. Its use of Gaussian distributions for the redshift populations, however, precludes realistic distribution shapes. Here, we propose a different approach for estimation of the redshift population distributions, where we use propensity scores for direct redshift calibration, thereby yielding an estimate of the full tomographic redshift distribution.

As described in Section 2.2, direct redshift calibration methods depend on the estimation of weights $\omega(x) = p_T(x)/p_S(x)$, used to reweight a spetroscopic sample (with known true redshifts) to obtain an estimate of the redshift distribution of the photometric sample (per tomographic bin). The weights $\omega(x)$ can also be expressed via

$$\omega(x) = \frac{p_T(x)}{p_S(x)} = \frac{p(s=1)}{p(s=0)} \frac{p(s=0|x)}{p(s=1|x)} \propto \left(\frac{1}{p(s=1|x)} - 1\right). \tag{25}$$

We can thus obtain an estimate of the weights $\omega(x)$ by employing the inverse of the propensity score (inverse-PS), via the right-hand side of (25). To estimate the tomographic binned redshift distributions, we first obtain a *StratLearn* conditional density estimate $\hat{f}(z_i|x_i)$ as described in Section 3.1 for each galaxy in both the photometric and the spectroscopic set. Based on these estimates, each galaxy (in both sets) is assigned to its respective tomographic bin, following the *StratLearn* binning strategy described in Section 3.2.[6]

For each tomographic bin, following (3), we obtain an estimate of the binned joint target distribution $p_{Tb}(z, x)$ via

$$p_{Tb}(z, x) = \omega_b(x) p_{Sb}(z, x), \tag{26}$$

[6]The binning of the spectroscopic set (as previously performed by W20) is needed when performing direct redshift calibration of the binned photometric set, since the photometric bin assignment is based on (summaries) of the conditional density estimates (as described in Section 3.2). The conditional density estimates implicitly incorporate information of source redshift (through the fitting process described in Section 3.1). To prevent unmeasured confounding (information encoded in the spectroscopic redshift, but not in the magnitudes/colors) the same selection function is applied for source and target data.

with $p_{sb}(z, x)$ being the binned joined source distribution of tomographic bin b, and with weights $\omega_b(x)$ computed via inverse-PS following (25) for each bin b. We estimate the propensity scores (employing logistic regression as detailed in Section 5.1) based on the covariates of the spectroscopic source galaxies and photometric target galaxies in the respective tomographic bin. In practice, we employ the relation in (26) by reweighting galaxies in the binned spectroscopic source data using the respective estimated inverse-PS weights (obtained via 25). We then obtain an estimate of the photometric redshift distribution $\hat{p}_b(z)$ (for each tomographic bin b) by looking at the marginal sample of z in the weighted joint distribution. This method is numerically demonstrated in Section 5.5.

5 NUMERICAL DEMONSTRATIONS

5.1 Simulation study

We explore the performance of our framework using the comprehensive set of realistic simulations introduced in W20. The simulations aim to mimic the KiDS + VIKING-450 data set, presented in Wright et al. (2019) and Hildebrandt et al. (2020), starting from the MICE2 simulation (Carretero et al. 2015; Crocce et al. 2015; Fosalba et al. 2015; Hoffmann et al. 2015) and based on a framework provided in Van Den Busch et al. (2020). In the following, we provide a summary of the simulated data employed in our study (see W20 for a full description of the construction and validation of the simulations).

5.1.1 *Photometric survey*

The simulations are designed to mimic the wide-field, multiband photometric data set of KiDS + VIKING-450. The KiDS + VIKING-450 data set consists of imaging in nine photometric bands ($ugriZYJHK_s$): the four optical bands ($ugri$) are observed as part of the KiDS survey (Kuijken et al. 2019) using the VLT Survey Telescope (VST; Capaccioli, Mancini & Sedmak 2005) located at the European Southern Observatory's Cerro Paranal observatory in Chile, and the five near-infrared filters ($ZYJHK_s$) are observed as part of the VIKING survey (Edge et al. 2013) using the Visible and Infrared Survey Telescope for Astronomy (VISTA; Dalton et al. 2006; Emerson, McPherson & Sutherland 2006, also located at Cerro Paranal). The first 450 square degrees of joint imaging from the two surveys forms the KiDS + VIKING-450 cosmic shear survey (referred to simply as 'KiDS' hereafter). The simulated photometric data D_T^u (where the superscript u refers to 'unweighted'; below, we describe a pre-processing step to produce a shear-measurement weighted photometric sample as employed in the downstream scientific analysis) consists of $\sim 21 \times 10^6$ galaxies, for each of which, a simulated measurement of its position, lensing convergence, morphological information, and model magnitudes in the $ugriZYJHK_s$-bands is provided. Magnitudes include photometric noise, realistic to KiDS survey data (W20, Section 5.1). A large proportion of galaxies ($\sim$17 per cent) have a flux error greater than or equal to the flux measurement in at least one band, and are therefore flagged as 'non-detections' in the KiDS photometric processing pipeline. Fig. D1 in the Appendix illustrates the full pattern of such cases. The flux measurement of such non-detections was removed prior to our analysis and only placeholder/indicator values were available to indicate these non-detection cases. We thus treat these cases as 'missing data' (details on processing of these cases appear at the end of this section). While the spectroscopic redshift z is unavailable for galaxies in the photometric set, a Bayesian redshift estimate z_B is also provided (see Section 3.2), which was used in previous works to assign galaxies to the tomographic bins.

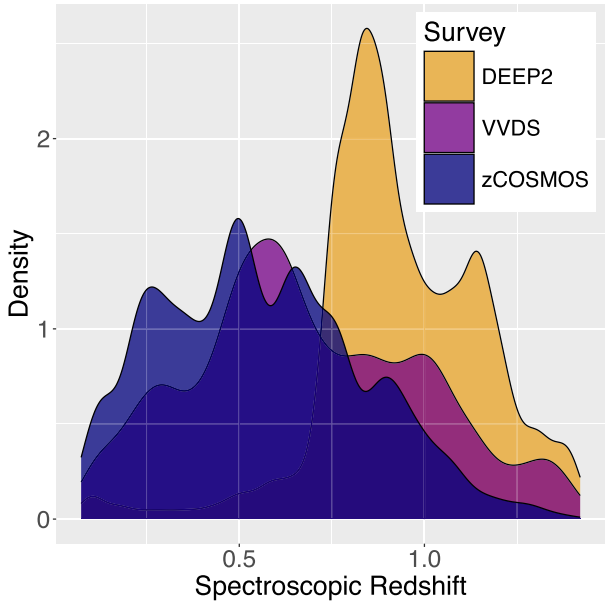


Figure 3. Spectroscopic redshift distributions of the three spectroscopic surveys used as source data.

Finally, a cosmic shear-measurement weight $\hat{w}_i$ is provided for each galaxy, which relates to the quality of its shear measurement, and which filters through to the cosmological analysis for cosmological parameter estimation. As a pre-processing step, we resample the data D_{Ti}^u proportionally to the cosmic shear-measurement weights $\hat{w}_i$. In this way, we obtain the target sample D_T, with $|D_T| = 12.48 \times 10^6$ galaxies. Fig. D2 in the appendix demonstrates that there is a negligible difference in targeting the shear-measurement resampled distribution (obtained via the above pre-processing step) and targeting the shear-measurement weighted distribution (as done in W20). Our results on the resampled data thus hold without loss of generality.

5.1.2 *Spectroscopic survey*

A much smaller spectroscopic source data set D_S, with $|D_S| = 21,537$, galaxies is provided, composed of simulated data mimicking three surveys: zCOSMOS (9930 galaxies), DEEP2 (6919 galaxies), and VVDS (4688 galaxies), altogether spanning a redshift range of $0.07 \leq z \leq 1.43$. The spectroscopic redshift distributions of the three surveys are illustrated in Fig. 3. The spectroscopic source set is not a representative sample of the photometric target distribution (selection effects are described in W20). Fig. 4 illustrates the density of the spectroscopic source redshift distribution (red), and the (true) redshift distribution of the photometric simulated target data (blue; not available in practice). While both distributions cover the same redshift range, the difference in densities is immediately apparent, an effect of the underlying covariate shift. For each galaxy in the spectroscopic set, the same set of covariates as provided for galaxies in the photometric set is available. In addition, an accurate spectroscopic measurement of the true redshift z is available for each galaxy, with measurement error that is negligible for our purposes.

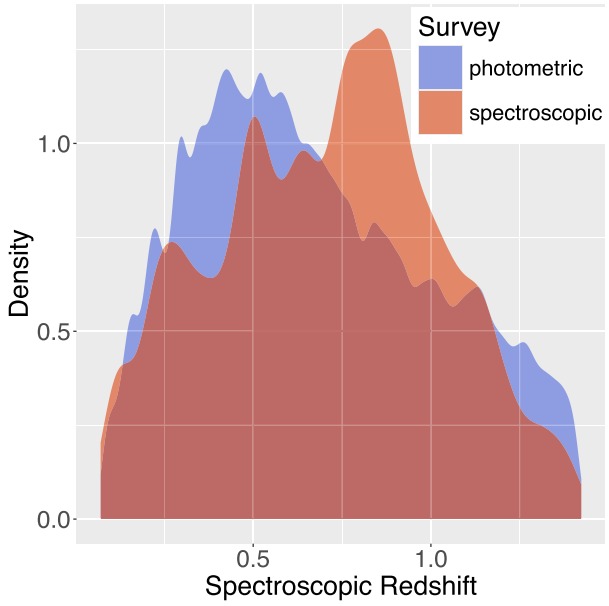


Figure 4. Spectroscopic (true) redshift distributions of the photometric simulated target data D_T (not available in practice) compared with the spectroscopic source data.

To account for sampling variance, 100 independent spectroscopic catalogues are provided, each with above described specifications. These correspond to 100 independent fields (lines-of-sights, abbreviated as LoS). The fields of the three spectroscopic surveys are independent of each other across the 100 LoS.

5.1.3 Choice of covariates and handling of missing data

To obtain the conditional density estimates for all objects in the photometric target data, we choose as covariates the r-band magnitude and the 8 colours: $(u-g, g-r, r-i, i-Z, Z-Y, Y-J, J-H, H-Ks)$, a set-up previously adopted (e.g. W20; Izbicki et al. 2017; Autenrieth et al. 2024). Using colours instead of magnitudes does not worsen the 'missing data' pattern, as illustrated in Fig. D1. As a pre-processing step, all covariates are scaled to have mean zero and standard deviation one. In the *StratLearn* framework, the missing data pattern has to be taken into account in the propensity score estimation step, and in the computation of the conditional density estimators within strata. For propensity score estimation, we use mean imputation of the 9 covariates to fill the missing values; we also add 9 binary indicator variables as dependent variables (main effects) to the logistic regression propensity score model, which describe the missingness of each covariate. Fig. D3 in the appendix illustrates the distributions of the estimated propensity scores for source and target data. The support of the target propensity score distribution is well covered by the support of the source propensity score distribution, demonstrating the availability of source galaxies that match the covariate space of the target galaxies.

The computation of the conditional density estimators (*ker-NN* and *Series*) requires the calculation of Euclidean distances between the covariate vectors of each galaxy. In the missing data cases, we compute the pairwise distances of two galaxieas using only the covariate values with measurements (no missingness) for both galaxies. The large size of the photometric target set causes additional computational challenges: for prediction of the conditional densities on the target data, distance matrices between photometric target set and spectroscopic source set are required, which is not computationally feasible for the entire target set at once. We thus process the prediction on the photometric target set in batches of 60 000 target samples. Table 1 shows the strata composition of spectroscopic source and photometric target data for one random batch, illustrating that there is enough spectroscopic source/training data in each stratum to fit the conditional density estimators within strata separately. While there is a slight discrepancy between the average redshift in source (0.71) and target data (0.68) overall, most strata have well-balanced redshift means between source and target, an indicator of reduced covariate shift after the propensity score stratification (Autenrieth et al. 2024). Other batches demonstrate a similar pattern. To reduce the computational burden, for each LoS, we use a fixed set of hyperparameters for prediction of the conditional density estimators on all target batches. For each LoS, we obtained the fixed hyperparameter set by optimizing (4) and (8), separately for each stratum, using one initial strata composition, with a randomly selected target batch.[7] Using batches of photometric target data has the advantage that distance matrices can be stored in memory, and predictions can be processed in parallel on several batches.[8]

Table 1. Composition of the five *StratLearn* strata. The number of galaxies and the average spectroscopic (true) redshift is presented in each source and target stratum. (Composition of one random batch of 60 000 photometric samples is shown for illustration).

Stratum	Set	#galaxies	Mean z
1	Source	6091	0.74
	Target	10 217	0.74
2	Source	5036	0.77
	Target	11 271	0.74
3	Source	4351	0.72
	Target	11 957	0.72
4	Source	3668	0.65
	Target	12 639	0.66
5	Source	2391	0.58
	Target	13 916	0.57
All	Source	21 537	0.71
	Target	60 000	0.68

5.2 Improved bin assignment accuracy

In this section, we evaluate the accuracy of the new tomographic bin assignment, obtained via *StratLearn*-based conditional density estimates, as described in Section 3.2, and illustrated in Fig. 1.

In Table 2, we compare our bin assignment with the standard practice of using z_B for the assignment, across five different classification performance metrics, demonstrating improvement in all

[7]Optimization of (4) was performed by splitting the source data within each strata in a training and validation set (one half each). The parameters which led to the best predictive performance on the source validation sets (in each stratum) were then selected for each conditional density estimator (ker-NN and Series) separately. The final optimization in (8) was then performed on the same source strata validation sets, using the optimized ker-NN and Series source validation set predictions. The hyperparameters for the five strata and for all 100 LoS are illustrated in the Appendix, Figs D4, D5, and D6.

[8]We performed all computations on a CPU cluster employing up to $\sim$ 150 CPU, simultaneously.

Table 2. Tomographic bin assignment performance evaluated over 100 LoS, comparing the *StratLearn* bin assignment (following Section 3.2) and the z_B bin assignment. The average (sd) of the performance metrics computed for each of the 100 LoS is reported for *StratLearn*. Using z_B, the bin assignment is consistently the same for all 100 LoS. For all metrics, higher values indicate better performance.

Performance metric	*StratLearn* mean (sd)	z_B mean (sd)
Accuracy	0.622 (0.003)	0.526 (−)
Balanced accuracy	0.718 (0.003)	0.706 (−)
Sensitivity	0.502 (0.006)	0.493 (−)
Specificity	0.934 (0.001)	0.918 (−)
Cohen's Kappa	0.439 (0.006)	0.415 (−)

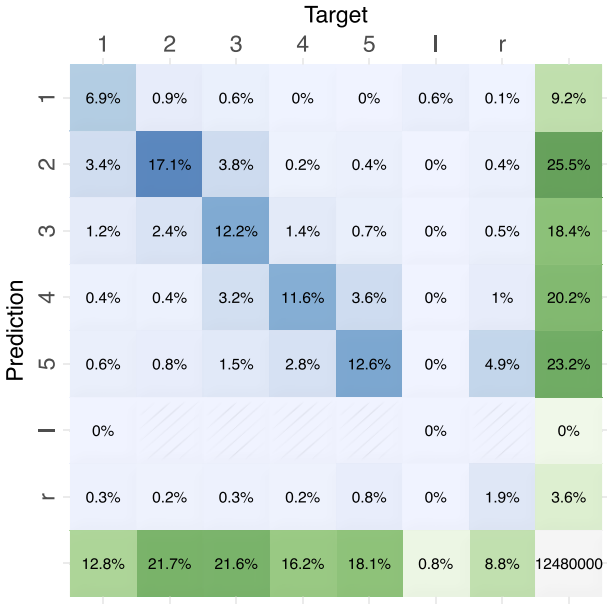
of them. On average across the 100 LoS, the *StratLearn* binning assigns the photometric target galaxies to the correct tomographic bin in 62.2 per cent of the cases (considering the five tomographic bins, and both end bins separately). This is a substantial improvement over the z_B binning, with an accuracy of 52.5 per cent. *StratLearn* improves both the sensitivity (true positive rate) and the specificity (true negative rate) of tomographic bin assignment compared to z_B, thus also leading to an improvement of the balanced accuracy and Cohen's kappa, which take into account the imbalance of class (bin) proportions.[9] The standard deviations of all performance measures across the 100 LoS is relatively low (Table 2), which demonstrates that the improvement is consistent throughout the 100 LoS (the z_B assignment is the same for all 100 LoS).

Figs 5(a) and (b) show the confusion matrices of tomographic bin assignment using *StratLearn* and z_B (averaged over the 100 LoS). The confusion matrices demonstrate that *StratLearn* improves the bin assignment across all five tomographic bins (top five diagonal values), and most substantially in the second bin (with $z \in (0.3, 0.5]$), the one with the largest fraction of galaxies (21.7 per cent). In this bin, *StratLearn* improves over the z_B bin assignment by more than 55 per cent.

The heatmap in Fig. 6 provides a visual comparison of the confusion matrices in Figs 5(a) and (b). Green squares in Fig. 6 correspond to an improvement of *StratLearn* over z_B, while pink squares correspond to better performance of the z_B assignment. The heatmap is computed by subtracting the diagonal values of Fig. 5(b), the bin assignment accuracy of z_B, from the diagonal values of Fig. 5(a), the bin assignment accuracy of *StratLearn*; on the off-diagonal, the sign is reversed, so that green (positive) values denote *StratLearn* improvement everywhere. The improvement of *StratLearn* is particularly strong in the top five diagonal squares, the five tomographic bins, which are of highest interest for the scientific analysis.

Fig. D7 in the appendix illustrates the changes in bin assignment of *StratLearn* versus z_B, showing a moderate to strong disagreement of *StratLearn* and z_B in most of the bins. The reassignment of galaxies, and especially the improved bin assignment accuracy of *StratLearn*, might thus substantially improve cosmological results – this will be subject of a future, dedicated study.

[9]The balanced accuracy is defined as (specificity + sensitivity)/2. The Cohen's Kappa measures the relative performance of the classifier with the performance of a random guess (based on the class frequency). Both metrics take on values between 0 and 1 (with 1 being a perfect classifier).



(a) StratLearn tomographic bin assignment.

(b) z_B tomographic bin assignment.

Figure 5. (a) Confusion matrix of the *StratLearn* tomographic bin assignment (averaged over 100 LoS). (b) Confusion matrix of the z_B tomographic bin assignment. The labels 'l' and 'r' refer to the left and right end bins, respectively (for galaxies outside the tomographic bin ranges).

5.3 Improved population mean estimates accuracy

The main purpose of this study is to obtain accurate estimates of the (true) redshift population means within tomographic bins. The foremost criteria to evaluate redshift calibration methods (Newman & Gruen 2022) is the mean discrepancy

$$\mathbb{E}[\hat{\mu}_b - \mu_b^{\text{true}}] \simeq \frac{1}{L}\sum_{l=1}^{L}(\hat{\mu}_{b,l} - \mu_{b,l}^{\text{true}}), \quad (= \widehat{\text{bias}}_b) \tag{27}$$

where $L = 100$ is the number of LoS, $\hat{\mu}_{b,l}$ the estimated mean redshift and $\mu^{\rm true}_{b,l}$ the true redshift mean for LoS l for galaxies assigned to tomographic bin b. We note that (27) is not a bias in a strict statistical sense, since the true redshift mean (within tomographic bin) varies across lines of sights. However, being consistent with the notation of previous studies, we will loosely refer to (27) as 'bias'. In addition to (27), we are interested in the standard deviation, SD, of the mean differences across the 100 LoS:

$$\mathrm{SD}\left(\hat{\mu}_b - \mu^{\rm true}_b\right) = \sqrt{\frac{\sum_{l=1}^{L}\left(\mu^{\rm diff}_{b,l} - \widehat{\mathrm{bias}}_b\right)^2}{L-1}} \tag{28}$$

with $\mu^{\rm diff}_{b,l} = \hat{\mu}_{b,l} - \mu^{\rm true}_{b,l}$, for $l = 1, \ldots, L$ and $b = 1, \ldots, 5$.

Table 3 presents the bias results obtained for our novel *StratLearn*-Bayes method, with a comparison to the SOM direct redshift calibration method introduced by W20. While there is a variety of cutting-edge redshift calibration methods in the literature (e.g. Rau et al. 2020; Wright et al. 2020; Myles et al. 2021; Malz & Hogg 2022; Rau et al. 2023, among others), the SOM method is an obvious choice for comparison, since it has been shown to outperform other direct redshift calibration methods (e.g. using k-nearest-neighbour methods, *k*NN; Hildebrandt et al. 2016, 2020) on the realistic and comprehensive simulations (mimicking the KiDS + VIKING-450 data) considered in this work (W20), thus making it most comparable. Based on the *StratLearn* binning, our method options *StratLearn*-Bayes (A) and (B) lead to an average absolute bias of 0.0053 and 0.0052 across the five tomographic bins, an improvement of ~ 40 per cent w.r.t the SOM method with z_B binning, which leads to an average absolute bias of 0.0085. We further note that the SOM (with z_B binning) method requires systematic quality cuts, which reduce the data size for the scientific analysis (we return to this point below).

We also apply the SOM calibration method using the new *StratLearn* binning, applying quality cuts as described in W20, which leads to an increase of bias (0.105 absolute average bias) compared to SOM with the z_B binning (0.0085 absolute average bias). Using the *StratLearn*-Bayes model on the z_B bin assignment also leads to an increase in bias to 0.0141 and 0.0131 (on absolute average across the five bins). Such a reduction in performance could in fact be expected: the *StratLearn*-Bayes model is based on the modelling of the *StratLearn* object level (galaxy) conditional density estimates, but by applying a different binning (e.g. via z_B) additional (external) errors are introduced in the assignment of galaxies per tomographic bin. The *StratLearn*-Bayes framework is not designed for correction of such external errors (biases), which lead to systematic shifts of the population mean estimates. For instance, if the z_B galaxy bin assignment is correlated with the variance of *StratLearn* galaxy conditional density estimates, then the (tomographic bin) population mean estimate in (22) can be systematically shifted. We thus advise against the combination of *StratLearn*-Bayes based on z_B binning, and advocate for the use of *StratLearn*-Bayes via the *StratLearn*-based binning, which leads to the best performance.

Table 4 shows the standard deviation (SD) population scatters from the 100 LoS. *StratLearn*-Bayes with *StratLearn* binning leads to slightly increased standard deviations of 0.0066 [option (A)] and 0.0067 [option (B)] on average across the five tomographic bins, compared to the SOM method based on z_B binning with an average of 0.0051. The results in Table 4 indicate that the standard deviation results are related to the binning strategy, rather than to the calibration method. Using SOM calibration on the *StratLearn* binning (with gold quality cuts) leads to comparable increase in SD of 0.0066 on average across the five bins. On the other hand, using the *StratLearn*-Bayes model applied on the z_B binning leads to a decrease in SD to an average of 0.0048 and 0.0047, even lower than applying SOM on the z_B binning.

In general, the similarity in results of the *StratLearn*-Bayes options (A) and (B) demonstrate robustness with respect to the computation of the population variance (last paragraph of Section 3.3). Both methods outperform the most comparable calibration method (SOM on z_B binning). Given a slight improvement of bias reduction, we propose the application of *StratLearn*-Bayes (B) as our best method.

Finally, our proposed method, *StratLearn*-Bayes (B), leads to a maximum bias within tomographic bins of $\Delta\langle z\rangle = 0.0095 \pm 0.0089$ (in bin 1). In contrast, using SOM based on z_B binning, leads to maximum biases of 0.0135 ± 0.0052 and 0.0147 ± 0.0040 (in bin 3 and 4). In addition, SOM based on z_B binning requires systematic quality cuts (gold selection), which are not necessary for our methodology.

The improved accuracy that we see with our proposed method brings the biases down to $\Delta\langle z\rangle < 0.01$ in all bins. This threshold has been chosen in previous work as delineating 'negligible' and 'non-negligible' biases (W20; Abdalla et al. 2022). Moreover, our method produces biases that are consistent with zero within 1.5σ in all bins, whereas the SOM method produces biases that are inconsistent with zero at the level of $\sim 3.7\sigma$ in the fourth bin. As such, our method is intrinsically less biased given the same calibrating data and target wide-field population, while retaining a greater number of sources for scientific analysis.

5.4 Larger sample size for weak lensing analysis

In Table 5, we show the absolute numbers of galaxies obtained via the different tomographic bin assignment strategies and quality cuts. In bins 2, 4, and 5, the number of galaxies is higher when using *StratLearn* binning compared to z_B. In bins 1 and 3, the number of galaxies is slightly higher using the z_B binning. Overall, due to the improved binning accuracy of *StratLearn*, there is an approximately 10 per cent increase in the number of available galaxies for science (summing over all bins) when using *StratLearn* for tomography instead of z_B. *StratLearn* assigns substantially fewer galaxies to the right end bin than z_B (see Figs 5a and b), leading to a lower proportion of galaxies that are falsely removed from the analysis (the five tomographic bins), but also to a higher proportion of galaxies that actually are in the right end bin (having redshift greater than 1.2), but are assigned to one of the five tomographic bins (mostly to bin 4 and 5). Given the small biases of *StratLearn*-Bayes in bins 4 and 5 (Table 3), the inclusion of such high-redshift galaxies does not seem to have a negative impact on the calibration, but the positive effect of increasing the available data size within tomographic bins.

Table 5 also provides the number of galaxies within bin after applying the gold selection, as introduced by W20. We note that the gold selection cut is not needed when applying the *StratLearn*-Bayes approach, while it is a necessary step to obtain the SOM results. Thus, compared to the previously best combination of bin assignment and calibration method on these simulated data in W20, the *StratLearn*-Bayes approach leads to an increase of galaxies available for science of $\sim$18 per cent.

For weak lensing analyses, the relevant statistic is the increase in the effective number of sources incorporating the shape measurement weight. Heymans et al. (2012) derive the metric for effective number density of weak lensing sources as

$$n_{\rm eff} = \frac{1}{A}\frac{(\sum_N w)^2}{\sum_N (w^2)}, \tag{29}$$

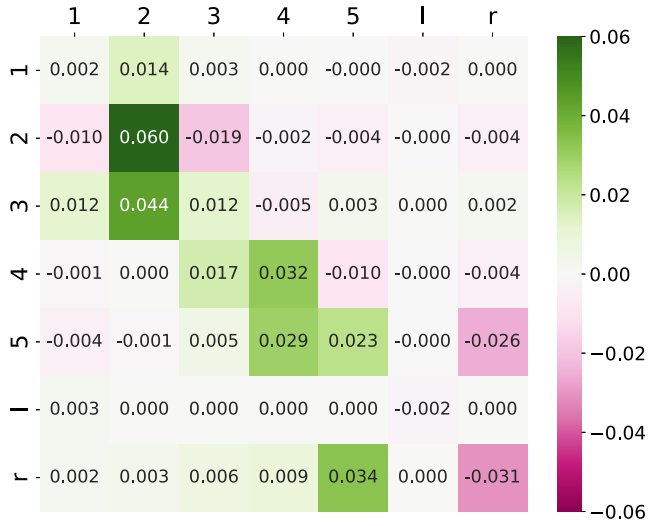


Figure 6. Heatmap of confusion matrix (accuracy) differences between *StratLearn* and z_B. On the diagonal, the difference of *StratLearn* – z_B accuracy's is shown. Off-axis, the difference of z_B - *StratLearn* is shown. Thus higher values (see colour scale at right) illustrate that *StratLearn* performs better than the z_B estimate.

where w is the shape-measurement weight for each source $i \in N$, and A is the survey area in square-arcmin. The change in the $n_{\rm eff}$ due to the SOM gold selection and quality control is described as $\Delta n_{\rm eff} = n_{\rm eff}^{\rm gold}/n_{\rm eff}^{\rm all}$. W20 quote this metric for SOM calibration with quality control in their table 2, finding values of ~ 0.8 in all bins. This suggests that, for a reanalysis of cosmic shear with our *StratLearn*-Bayes approach, we would increase the available lensing sample statistical power by a similar ~ 20 per cent in each tomographic bin.

5.5 Population distribution estimates

In the previous sections, we demonstrate the ability of the *StratLearn*-Bayes method to accurately and precisely estimate the redshift population means, which is most crucial for photo-z calibration in the weak lensing analysis. Since realistic estimates of the population distribution shapes will become more influential in cosmic shear analysis and for photometric galaxy clustering (as discussed in Section 1), here we numerically demonstrate how propensity scores can be employed via inverse-PS weighting (as introduced in Section 4) to improve estimation of the whole shape of the distribution.

In Fig. 7, we show the inverse-PS weighted redshift distributions per tomographic bin (in purple), obtained via the procedure described in Section 4, and based on the *StratLearn* tomographic bin assignment (following Section 3.2). The true redshift distributions per tomographic bin (not known in practice) are shown in black. The purple inverse-PS weighted distributions exhibit a similar shape as the black true redshift population distributions recovering reasonably well the true photometric population distribution shapes, particularly throughout tomographic bins 1 to 3. Fig. 7 further illustrates the SOM estimated population distributions (in orange), and its underlying true redshift population distributions (in light blue) obtained on the *StratLearn* tomographic binning after applying the gold selection quality cuts (W20). Fig. 7 presents the average (estimated) redshift population distributions across the 100 LoS per each tomographic bin.[10]

In Fig. 8, we assess the quality of the two estimation methods (inverse-PS and SOM) w.r.t. their underlying true distributions via probability–probability plots (pp-plots)[11]: the figure shows the average pp-plot (across the 100 LoS) for the inverse-PS estimated distributions versus the true (full) photometric redshift distributions per tomographic bin in purple lines, and the average pp-plot of the SOM estimated distributions versus the gold selected true distributions in orange dashed lines. The vertical bars gives 95 per cent intervals indicating the dispersion of the central 95 pp-plot lines from the 100 LoS.

Both estimates (inverse-PS and SOM) are close to the diagonal line throughout bins 1 to 3, with larger deviations in bins 4 and 5. Notably, the Inverse-PS and SOM pp-plot lines exhibit very similar deviation patterns from the diagonal line; both methods are based on reweighting of the spectroscopic samples (following 3), which explains similarities in their estimates. The purple (average) inverse-PS lines are closer to the diagonal line than SOM in tomographic bins 1,3,4, and 5, and almost identical with SOM in bin 2. In addition, the vertical 95 per cent intervals are generally smaller for inverse-PS compared to SOM (particularly in bins 1 to 3), indicating less variability in the estimate across the 100 LoS. Overall, the inverse-PS estimate thus approximates its underlying truth (the full binned photometric distribution) better than the SOM method its underlying (gold-selected) true distribution, with the additional advantage that no quality cuts are required for inverse-PS, leading to ~ 18 per cent more galaxies in the photometric sample available

[10]For illustration purposes, a mild Gaussian kernel density smoothing (with bandwidth 0.00294) was applied to the presented distributions in Fig. 7. The non-smoothed distributions are illustrated in Fig. D9 in the appendix.

[11]pp-plots are obtained by plotting two (empirical) cumulative distribution functions (CDF) against each other. The distributions are equal iff the pp-plot falls on the diagonal line from (0,0) to (1,1).

Table 3. Mean discrepancy (bias) computed over 100 lines of sight for different calibration methods, and different bin assignment strategies. We abbreviate the *StratLearn* bin assignment as *SL*. The add-on (gold) denotes quality cuts applied to the data according to W20. The Galaxies column shows the total number of galaxies (in millions) available in the five tomographic bins.

	Binning	Galaxies [M]	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	|Average|
StratLearn-Bayes (A)	*SL*	12.02	0.0123	−0.0076	0.0053	−0.0010	0.0001	0.0053
StratLearn-Bayes (B)	*SL*	12.02	0.0095	−0.0092	0.0047	−0.0013	0.0012	0.0052
SOM	*SL* (gold)	11.48	−0.0084	0.0022	0.0156	0.0117	0.0148	0.0105
StratLearn-Bayes (A)	z_B	10.90	0.0259	0.0127	0.0084	0.0003	−0.0231	0.0141
StratLearn-Bayes (B)	z_B	10.90	0.0228	0.0117	0.0071	−0.0002	−0.0236	0.0131
SOM	z_B (gold)	10.17	−0.0005	0.0036	0.0135	0.0147	−0.0102	0.0085

Table 4. As in Table 3, but showing standard deviation (SD) computed over 100 lines of sight for the different calibration methods, and different bin assignment strategies.

	Binning	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	|Average|
StratLearn-Bayes (A)	*SL*	0.0087	0.0065	0.0046	0.0052	0.0082	0.0066
StratLearn-Bayes (B)	*SL*	0.0089	0.0065	0.0045	0.0052	0.0082	0.0067
SOM	*SL* (gold)	0.0085	0.0064	0.0059	0.0058	0.0066	0.0066
StratLearn-Bayes (A)	z_B	0.0055	0.0048	0.0049	0.0037	0.0051	0.0048
StratLearn-Bayes (B)	z_B	0.0052	0.0048	0.0047	0.0036	0.0051	0.0047
SOM	z_B (gold)	0.0055	0.0061	0.0052	0.0040	0.0049	0.0051

Table 5. Sample sizes (in millions) within tomographic bins obtained via different bin assignment strategies and quality cuts (mean and standard deviation computed over 100 LoS). With (gold), we refer to the gold selection quality cuts described in W20.

		Bin 1 (0.1,0.3]	Bin 2 (0.3, 0.5]	Bin 3 (0.5,0.7]	Bin 4 (0.7, 0.9]	Bin 5 (0.9, 1.2]	Total (0.1, 1.2]
StratLearn	mean	1.14	3.18	2.29	2.51	2.90	12.03
	(sd)	(0.112)	(0.207)	(0.189)	(0.204)	(0.220)	
StratLearn (gold)	mean	1.06	2.94	2.21	2.51	2.75	11.48
	(sd)	(0.089)	(0.158)	(0.171)	(0.205)	(0.223)	
z_B	mean	1.31	1.94	2.84	2.16	2.66	10.90
	(sd)	(−)	(−)	(−)	(−)	(−)	
z_B (gold)	mean	1.15	1.91	2.36	2.10	2.65	10.17
	(sd)	(0.034)	(0.007)	(0.071)	(0.047)	(0.003)	

for scientific analysis. For additional visualization of the distribution differences presented in Fig. 8, Fig. D8 in the appendix illustrates a slightly modified version of Fig. 8 by subtracting the *x*-axis values (the quantiles of the true distributions) from the *y*-axis values (the quantiles of the estimated distributions) in each tomographic bin.[12]

In Fig. 9, we assess the differences between the true full photometric redshift distribution and the true redshift distribution after gold selection, for each of the five tomographic bins. Fig. 9 presents a (modified) pp-plot, illustrating the full true photometric distributions (on the *x*-axis) versus the gold selected true distributions (on the *y*-axis); with the modification that the *x*-axis values (full true distribution quantiles) are subtracted from the *y*-axis (gold selection truth quantiles) for better visibility of the distribution differences. Fig. 9 illustrates that there are some mild changes in the underlying truth when applying the gold selection quality cuts (compared to the full true photometric sample) for bins 1,2,3, and 5. In bin 4, the true photometric distributions (before and after gold selection cuts) are approximately the same.

Finally, as noted in Section 2.2, direct redshift calibration methods are generally prone to high variance, in particularly in the presence of a small number of large weights. While we have demonstrated improvement of inverse-PS upon SOM for estimation of the redshift population distribution shapes on the *StratLearn*-based binning, it is true that the inverse-PS estimate can generally be affected by the same large variance instability. We note however that the formulation of the weights via propensity scores enables the use of methods developed in the (causal inference) propensity score literature to improve assessment and estimation of the propensity scores for direct redshift calibration (e.g. Imai & van Dyk 2004; Pirracchio, Petersen & van der Laan 2014; Austin & Stuart 2015; Ridgeway et al. 2022; Autenrieth et al. 2021), which will be the subject of a dedicated future work.

[12]While here we are mostly interested in the population estimates obtained for the newly proposed and more accurate *StratLearn*-based tomographic binning, we provide similar assessment of the population distribution estimates obtained for z_B-based tomographic binning in Figs D10 and D11 in the appendix.

6 DISCUSSION

This paper introduced a novel statistically principled method that improves photometric redshift calibration for weak lensing. The central plank of our approach is the estimation of individual galaxy photo-z conditional densities within a Bayesian hierarchical model, coupled with the *StratLearn* framework, a recently proposed statistically principled method for learning under non-representative source/training data in the presence of covariate shift (Autenrieth et al. 2024). The computation of galaxy-level conditional density estimates allows us to introduce an alternative tomographic binning strategy to the previously used z_B-based binning (Benitez 2000). We presented a hierarchical Bayesian framework, (*StratLearn*-Bayes), to model summaries of the conditional density estimates to obtain nearly unbiased photometric redshift population mean estimates within tomographic bins.

Before summarizing the main findings of our study, we briefly discuss some limitations associated with our analysis and methodology, and potential improvements for future work. First, we note that throughout the paper, we assume that the covariate shift assumption holds, i.e. we assume there are no unmeasured covariates that are associated with both the source/target selection and the redshift of a galaxy. As discussed in Appendix B, we plan to consider potential violations of this assumptions due to quality cuts (Newman et al. 2015; Hartley et al. 2020) for prevention of redshift failures in ongoing/future work, with the aim of further reducing bias. We

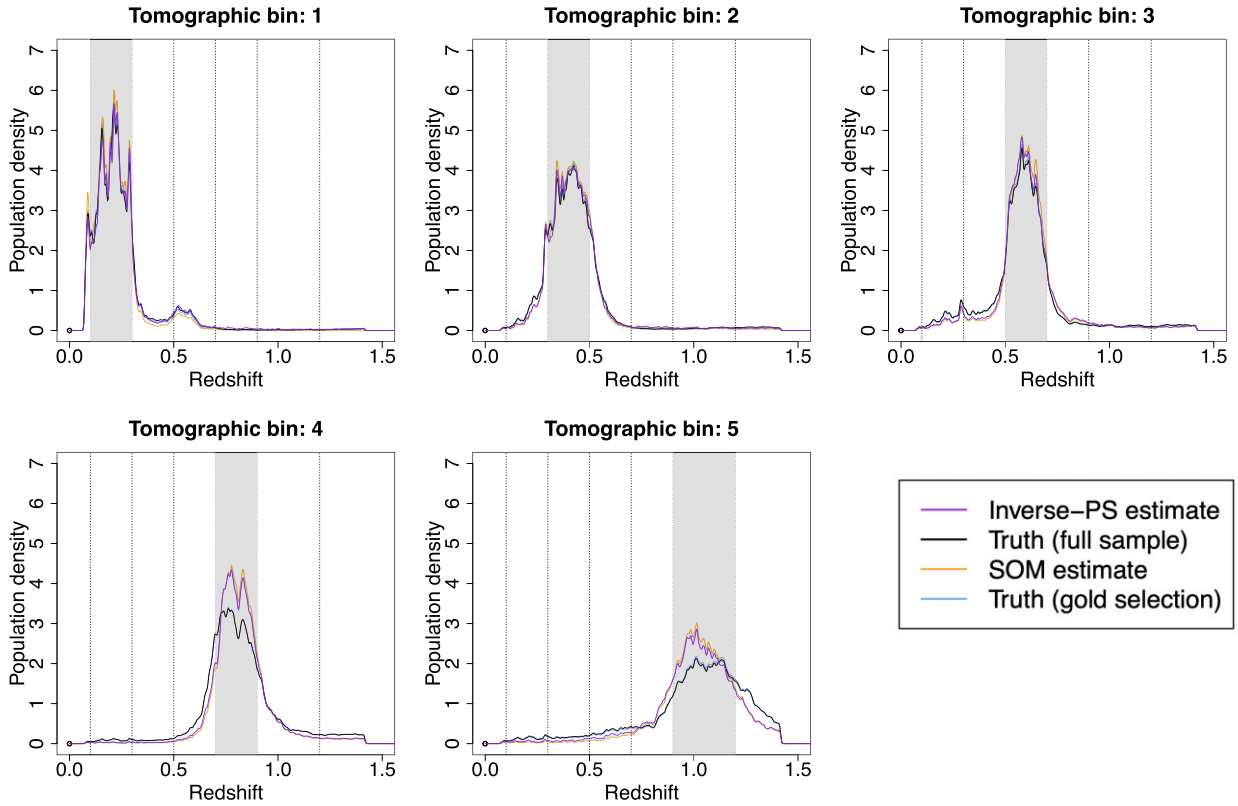


Figure 7. Redshift population distribution (estimates) per tomographic bin, with tomographic bins obtained as described in Section 3.2 via *StratLearn*-based binning. The figure illustrates the inverse-PS (purple) and SOM (orange) distribution estimates. The underlying true photometric redshift population distributions per tomographic bin (not known in practice) are illustrated in black for the full sample truth, and in light blue for the gold selected true distributions. The averaged (estimated) distributions across the 100 LoS are illustrated per tomographic bin.

further note that some of the remaining bias in the population mean estimates may be explained by effects introduced via the tomographic bin assignment described in Section 3.2. Employing a soft classification of galaxies based on the tomographic bin probabilities may lead to further bias improvement in future work.

We evaluated our method on a comprehensive and realistic simulation study (W20; Van Den Busch et al. 2020) mimicking the KiDS + VIKING-450 data set (Wright et al. 2019; Hildebrandt et al. 2020) with realistic photometric noise and spectroscopic incompleteness. The results of this study can be summarized in four points:

(i) The *StratLearn* conditional density-based tomographic binning strategy substantially improves upon the z_B tomographic bin assignment, with an overall binning accuracy of $\sim$62.2 per cent using *StratLearn*, compared to $\sim$52.6 per cent using z_B.

(ii) The *StratLearn*-Bayes model leads to the lowest bias in the estimation of tomographic redshift population means. On average across the five tomographic bins, the proposed *StratLearn*-Bayes method leads to an absolute bias of 0.0052, a substantial improvement over the previously best direct redshift calibration method employed on this simulation study, SOM with z_B binning (W20), of 0.0085 average absolute bias. The strong reduction of bias is accompanied by a slight increase in uncertainty, leading to an average standard deviation of 0.0067, compared to 0.0051. Using the *StratLearn*-Bayes framework, we find a maximum bias of $\Delta\langle z\rangle = 0.0095 \pm 0.0089$, slightly below the potentially critical bias value of $\Delta\langle z\rangle > 0.01$, compared to SOM based on z_B binning, which leads to maximum biases of 0.0135 ± 0.0052 and 0.0147 ± 0.0040.

(iii) While the SOM calibration method based on z_B binning, requires systematic quality cuts to define a gold sample (W20), our method does not require any cuts of the photometric sample. Thus, together with the improved tomographic bin assignment, the *StratLearn*-Bayes framework delivers an increase of $\sim$18 per cent in the galaxies available for the cosmic shear analysis.

(iv) We demonstrate how propensity scores can be employed via inverse-PS weighting in a direct redshift calibration approach to obtain realistic estimates of the redshift population distribution shapes per tomographic bin. Given the newly proposed *StratLearn* binning, we show that using inverse-PS leads to a better approximation of the true photometric population distributions compared to employing SOM for estimation of the gold selected population distributions (with the additional advantage of not requiring any quality cuts).

Finally, we believe that the improved tomographic binning assignment, the reduction of population mean bias within tomographic bin, and the increase in the number of galaxies available for cosmic shear analysis will have a substantial impact on the eventual scientific results and cosmological parameter inference. Analysing the final KiDS data release with our improved calibration method might lead to more precise and more accurate constraints on cosmological parameter estimates, particularly on S_8, the clustering strength of (predominantly dark) matter. We further believe that the proposed method might provide a powerful tool to improve the analysis

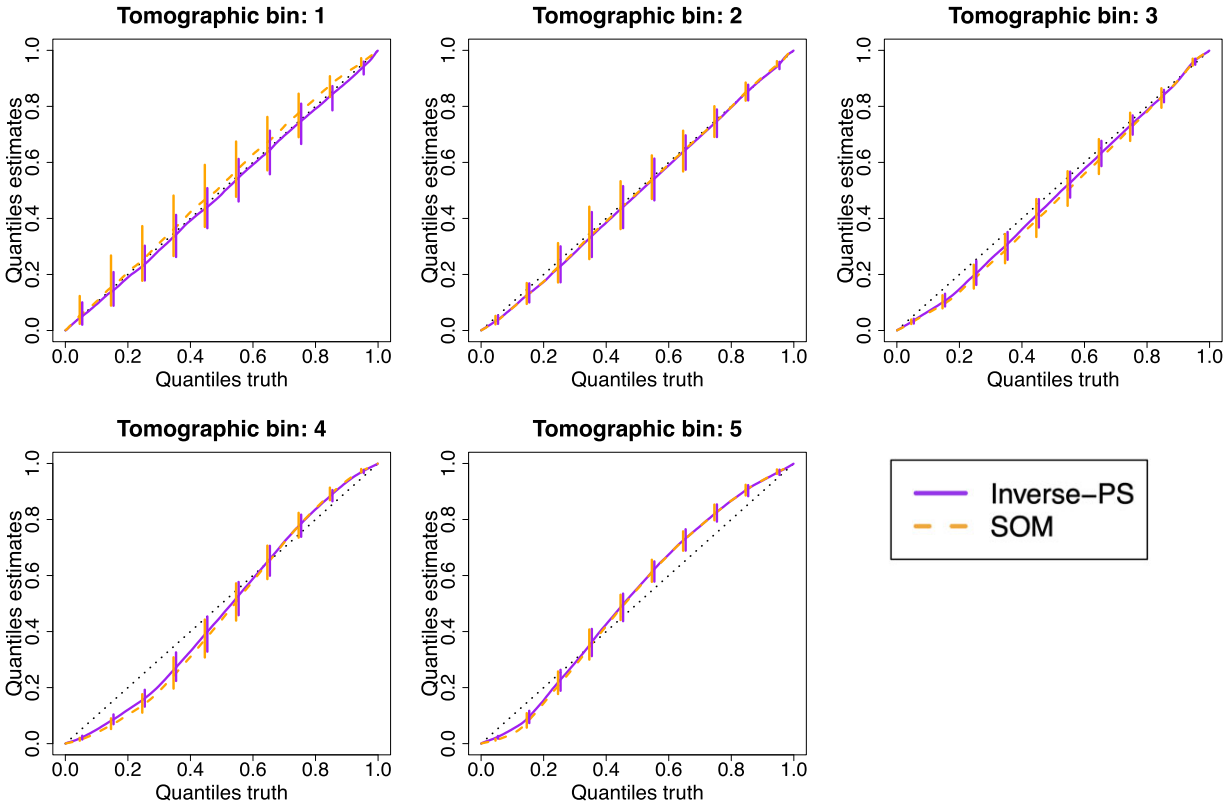


Figure 8. Probability–probability plots (pp-plot) for the inverse-PS estimated distributions versus the true (full) photometric redshift distributions in purple lines, and pp-plots of the SOM estimated distributions versus the gold selected true distributions in orange dashed lines, based on the *StratLearn* tomographic binning (following Section 3.2). For each tomographic bin, the averaged pp-plots across the 100 LoS are presented, with vertical bars illustrating 95 per cent intervals indicating the range of the central 95 pp-plot lines from the 100 LoS.

of present and upcoming cosmic shear analysis. We will investigate if the expected availability of larger spectroscopic source data sizes might allow further reduction of bias and variability to meet the stringent accuracy requirements of Euclid (Laureijs et al. 2011) and the Legacy Survey of Space and Time (LSST; Abell et al. 2009).

ACKNOWLEDGEMENTS

We would like to thank Alan Heavens and Andrew Jaffe for valuable discussions. DVD and MA acknowledge partial support from the UK Engineering and Physical Sciences Research Council [EP/W015080/1, EP/W522673/1]. RT's work was partially supported by STFC in the UK [ST/P000762/1, ST/T000791/1]; and DCS acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2021-03985]. DVD, DCS, and MA acknowledge support from the Marie-Skodowska-Curie RISE [H2020-MSCA-RISE-2019-873089] grant provided by the European Commission. RT acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1–Project FAIR 'Future Artificial Intelligence Research'. This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22]. RT is partially supported by the Fondazione ICSC, Spoke 3 'Astrophysics and Cosmos Observations', Piano Nazionale di Ripresa e Resilienza Project ID CN00000013 'Italian Research Center on High-Performance Computing, Big Data and Quantum Computing' funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di 'campioni nazionali di R&S (M4C2-19)' – Next Generation EU (NGEU). BJ acknowledges support by STFC Consolidated grant ST/V000780/1. AHW is supported by an European Research Council Consolidator grant (no. 770935), as well as by the Deutsches Zentrum für Luft- und Raumfahrt (DLR), made possible by the Bundesministerium für Wirtschaft und Klimaschutz, and acknowledges funding from the German Science Foundation DFG, via the Collaborative Research Center SFB1491 'Cosmic Interacting Matters – From Source to Signal'. 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. The MICE simulations have been developed at the MareNostrum supercomputer (BSC–CNS) thanks to grants AECT-2006-2-0011 through AECT-2015-1-0013. Data products have been stored at the Port d'Informació Científica (PIC), and distributed through the CosmoHub webportal (cosmohub.pic.es). Finally, this research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca) and the BC DRI Group. *The authors report there are no competing interests to declare.*

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

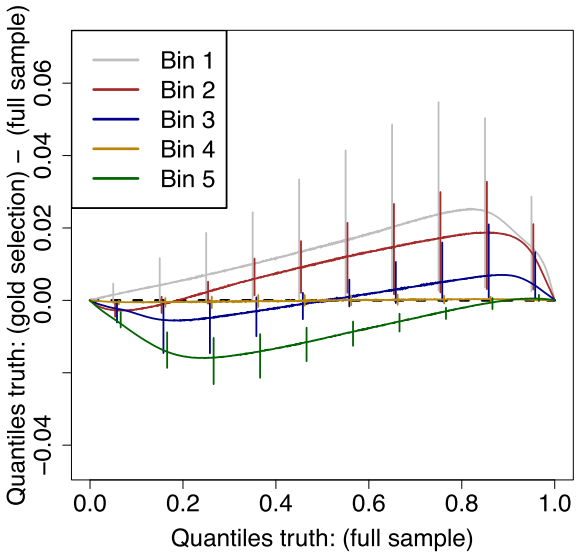


Figure 9. The figure illustrates (modified) pp-plots, comparing the true redshift distribution of the full sample (without quality cuts) with the true redshift distribution after applying gold selections, per tomographic bin (based on *StratLearn* tomograhic binning, following Section 3.2). For better readability, we illustrate 'modified' pp-plots, with the x-axis placing the quantiles of the full sample truth and the y-axis showing the quantiles of the gold selected truth subtracted by the x-axis values (the quantiles of the full sample true redshift distribution). For each tomographic bin, means of the pp-lines across the 100 LoS are illustrated, as well as 95 per cent intervals (vertical bars).

REFERENCES

Abbott T. M. C. et al., 2022, Phys. Rev. D, 105, 023520
Abdalla E. et al., 2022, J. High Energy Astrophys., 34, 49
Abell P. A. et al., 2009, Lsst science book, version 2.0.
Alarcon A., Sánchez C., Bernstein G. M., Gaztañaga E., 2020, MNRAS, 498, 2614
Amara A., Refregier A., 2008, MNRAS, 391, 228
Amon A. et al., 2022, Phys. Rev. D, 105, 023514
Asgari M. et al., 2021, A&A, 645, A104
Austin P. C., Stuart E. A., 2015, Stat. Medi., 34, 3661
Autenrieth M., Levine R. A., Fan J., Guarcello M. A., et al., 2021, J. Educ. Data Min., 13, 24
Autenrieth M., van Dyk D. A., Trotta R., Stenning D. C., 2024, Stat. Anal. Data Min.: ASA Data Sci. J., 17, e11643
Benitez N., 2000, ApJ, 536, 571
Buchs R. et al., 2019, MNRAS, 489, 820
Capaccioli M., Mancini D., Sedmak G., 2005, The Messenger, 120, 10
Carretero J., Castander F., Gaztañaga E., Crocce M., Fosalba P., 2015, MNRAS, 447, 646
Cochran W. G., 1968, Biometrics, 24, 295
Crocce M., Castander F., Gaztañaga E., Fosalba P., Carretero J., 2015, MNRAS, 453, 1513
Dalton G. B. et al., 2006, in McLean I. S., Iye M., eds, Ground-based and Airborne Instrumentation for Astronomy. Vol. 6269, SPIE, Belingham, p. 62690X
Dark Energy Survey and Kilo-Degree Survey Collaboration et al., 2023, Open J. Astrophys., 6, 36
De Vicente J., Sánchez E., Sevilla-Noarbe I., 2016, MNRAS, 459, 3078
Dey B., Zhao D., Newman J. A., Andrews B. H., Izbicki R., Lee A. B., 2022, preprint (arXiv:2205.14568)
Edge A., Sutherland W., Kuijken K., Driver S., McMahon R., Eales S., Emerson J. P., 2013, The Messenger, 154, 32
Emerson J., McPherson A., Sutherland W., 2006, The Messenger, 126, 41
Fosalba P., Crocce M., Gaztañaga E., Castander F., 2015, MNRAS, 448, 2987
Freeman P. E., Izbicki R., Lee A. B., 2017, MNRAS, 468, 4556
Gatti M. et al., 2018, MNRAS, 477, 1664
Gatti M. et al., 2022, MNRAS, 510, 1223
Gelman A., Carlin J. B., Stern H. S., Rubin D. B., 1995, Bayesian Data Analysis. Chapman and Hall/CRC, New York/US, p. 552
Hamana T. et al., 2020, PASJ, 72, 16
Hartley W. G. et al., 2020, MNRAS, 496, 4769
Heymans C. et al., 2012, MNRAS, 427, 146
Hikage C. et al., 2019, PASJ, 71, 43
Hildebrandt H. et al., 2010, A&A, 523, A31
Hildebrandt H. et al., 2012, MNRAS, 421, 2355
Hildebrandt H. et al., 2016, MNRAS, 463, 635
Hildebrandt H. et al., 2020, A&A, 633, A69
Hoffmann K., Bel J., Gaztanaga E., Crocce M., Fosalba P., Castander F. J., 2015, MNRAS, 447, 1724
Hoyle B. et al., 2018, MNRAS, 478, 592
Hu W., 1999, ApJ, 522, L21
Imai K., van Dyk D. A., 2004, J. Am. Stat. Assoc., 99, 854
Izbicki R., Lee A. B., 2016, J. Comput. Graph. Stat., 25, 1297
Izbicki R., Lee A. B., Freeman P. E. et al., 2017, Ann. Appl. Stat., 11, 698
Jones D. M., Heavens A. F., 2019, MNRAS, 483, 2487
Joudaki S. et al., 2020, A&A, 638, L1
Kohonen T., 1982, Biol. Cybern., 43, 59
Kuijken K. et al., 2019, A&A, 625, A2
Van der Laan M. J., Polley E. C., Hubbard A. E., 2007, Stat Appl Genet Mol Biol, 6
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258
Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118
Malz A. I., Hogg D. W., 2022, ApJ, 928, 127
Masters D. et al., 2015, ApJ, 813, 53
McQuinn M., White M., 2013, MNRAS, 433, 2857
Moreno-Torres J. G., Raeder T., Alaiz-RodríGuez R., Chawla N. V., Herrera F., 2012, Pattern Recognit., 45, 521
Morrison C. B., Hildebrandt H., Schmidt S. J., Baldry I. K., Bilicki M., Choi A., Erben T., Schneider P., 2017, MNRAS, 467, 3576
Myers J. A., Rassen J. A., Gagne J. J., Huybrechts K. F., Schneeweiss S., Rothman K. J., Joffe M. M., Glynn R. J., 2011, Am. J. Epidemiology, 174, 1213
Myles J. et al., 2021, MNRAS, 505, 4249
Naimi A. I., Balzer L. B., 2018, Eur. J. Epidemiology, 33, 459
Newman J. A., 2008, ApJ, 684, 88
Newman J. A., Gruen D., 2022, ARA&A, 60, 363
Newman J. A. et al., 2015, Astropart. Phys., 63, 81
Pirracchio R., Petersen M. L., van der Laan M., 2014, Am. J. Epidemiology, 181, 108
Planck Collaboration et al., 2020, A&A, 641, A1
Rau M. M., Wilson S., Mandelbaum R., 2020, MNRAS, 491, 4768
Rau M., Morrison C., Schmidt S., Wilson S., Mandelbaum R., Mao Y., 2022, MNRAS, 509, 4886
Rau M. M. et al., 2023, MNRAS, 524, 5109
Reischke R., 2024, MNRAS, 530, 4412
Ridgeway G., McCaffrey D., Morral A., Burgette L., Griffin B. A., 2022, Santa Monica, CA: RAND Corporation
Rosenbaum P. R., Rubin D. B., 1983, Biometrika, 70, 41
Rubin D. B., 1997, Ann. Intern. Med., 127, 757
Salvato M., Ilbert O., Hoyle B., 2019, Nat. Astron., 3, 212
Sánchez C., Bernstein G. M., 2019, MNRAS, 483, 2801
Schneider M., Knox L., Zhan H., Connolly A., 2006, ApJ, 651, 14
Secco L. F. et al., 2022, Phys. Rev. D, 105, 023515
Sugiyama S. et al., 2023, Phys. Rev. D, 108, 123517
Tanaka M. et al., 2018, PASJ, 70, S9
Troxel M. A. et al., 2018a, Phys. Rev. D, 98, 043528
Troxel M. A. et al., 2018b, MNRAS, 479, 4998
Van Den Busch J. et al., 2020, A&A, 642, A200
Wolpert D. H., 1992, Neural Netw., 5, 241

Wright A. H. et al., 2019, A&A, 632, A34
Wright A. H., Hildebrandt H., Van den Busch J. L., Heymans C., 2020, A&A, 637, A100 (W20)

APPENDIX A: ADDITIONAL DETAILS FOR CONDITIONAL DENSITY ESTIMATION

In this section, we provide derivations of the *generalized* risk under the L^2-loss given in (4). Following Izbicki et al. (2017), we start with the risk based on the general L^2-loss via

$$R_S(\hat{f}) = \iint (\hat{f}(z|x) - f(z|x))^2 \, \mathrm{d}P_S(x) \, \mathrm{d}z, \tag{A1}$$

with $\hat{f}(z|x)$ being the full conditional density estimate of redshift z given the covariates at point x, $f(z|x)$ being the true conditional density of z given x, and $P_S(x)$ being the distribution of the source covariates. In extended form, (30) can be written as

$$R_S(\hat{f}) = \iint \hat{f}^2(z|x) \, \mathrm{d}P_S(x) \, \mathrm{d}z - 2 \iint \hat{f}(z|x) f(z|x) \, \mathrm{d}P_S(x) \, \mathrm{d}z + \underbrace{\iint f^2(z|x) \, \mathrm{d}P_S(x) \, \mathrm{d}z}_{= \text{constant } C}, \tag{A2}$$

which up to the constant C is equal to

$$R_S(\hat{f}) = \iint \hat{f}^2(z|x) \, \mathrm{d}P_S(x) \, \mathrm{d}z - 2 \iint \hat{f}(z|x) \, \mathrm{d}P_S(x, z). \tag{A3}$$

From (31) to (32), the equality $\mathrm{d}P_S(x, z) = f(z|x) \, \mathrm{d}P_S(x) \, \mathrm{d}z$ (via Radon–Nikodym derivative) is employed. Given the labelled source samples (x_S, z_S), we can get an estimate of (32) via

$$\hat{R}_S(\hat{f}) = \frac{1}{n_S} \sum_{k=1}^{n_S} \int \hat{f}^2 \left(z|x_S^{(k)} \right) \mathrm{d}z - 2 \frac{1}{n_S} \sum_{k=1}^{n_S} \hat{f} \left(z_S^{(k)}|x_S^{(k)} \right), \tag{A4}$$

as presented in (4).

APPENDIX B: COVARIATE SHIFT ASSUMPTION

In this section, we discuss the covariate shift assumption and its potential violation due to spectroscopic quality cuts as indicated by Hartley et al. (2020). Following a number of previous studies (e.g. Lima et al. 2008; Hildebrandt et al. 2020; Wright et al. 2020), and as described in Section 2, we assume throughout this paper that the covariate shift assumption holds, i.e. $p_S(x) \neq p_T(x)$, but $p_S(z|x) = p_T(z|x)$. That means there are no unmeasured covariates that are associated to the selection of galaxies into the spectroscopic source set, and also predictive for redshift z.

As discussed in Hartley et al. (2020), additional selection cuts might lead to a violation of this assumption. More precisely, in some cases, the spectroscopic redshift measurement/estimate (for galaxies in the spectroscopic source set) may disagree significantly with the true redshift, a situation referred to as 'redshift failure' (Hartley et al. 2020). To avoid contamination of subsequent analyses, quality/confidence flags are introduced to indicate and remove galaxies with suspected redshift failure.

These quality flags are primarily determined based on characteristics of the galaxy spectra (e.g. the S/N of emission and absorption lines, and the strength of the 4000 Å break; Hartley et al. 2020). Here, we denote the spectroscopic features/covariates used to obtain the quality flags for a galaxy i as y_i. Based on y_i, a galaxy is assigned a high- or low-quality flag, and galaxies with low-quality flags are removed from the analysis pipeline, so we can assume that the total fraction of spectroscopic redshift failures will only be around ~ 1 per cent (Hildebrandt et al. 2020).

Of course, spectral information is obtained only for galaxies in the spectroscopic source set, but not for galaxies in the photometric target set. The spectral features y associated to the selection of galaxies into the final spectroscopic source set are thus unmeasured for galaxies in the photometric target set. As demonstrated in the causal inference literature (e.g. Rubin 1997; Myers et al. 2011; Austin & Stuart 2015), a correction of such covariates is only necessary/important if they are also associated to the outcome variable.[13] A possible justification of our covariate shift assumption is that, given the photometric covariates x, the additional spectroscopic covariates y (employed to indicate quality flags) are not further predictive/informative of redshift z, i.e. $p(z|x, y) = p(z|x)$. A thorough analysis of this assumption along with possible correction strategies is a topic of ongoing/future work.

APPENDIX C: ADDITIONAL MODEL DETAILS

C1 Posterior derivations

This section provides the theoretical justification of the posterior derivations in Section 3.3.

Deriving (20) from (19) is obtained by integrating over the product of normal densities in (19), which can analytically be done via completing the squares. We note that flipping the z_i and $\hat{\xi}_i$ in (19) constitutes the standard normal–normal hierarchical model with Gaussian measurement errors on latent z_i with Gaussian population. The analytical derivation of this model is a standard result in Bayesian statistics (see e.g. Gelman et al. 1995, page 117), demonstrating that

$$\int_{\mathbb{R}} N\left(\hat{\xi}_i|z_i, \hat{\tau}_i^2\right) N(z_i|\mu, \sigma^2) \, \mathrm{d}z_i = N\left(\hat{\xi}_i|\mu, \hat{\tau}_i^2 + \sigma^2\right). \tag{C1}$$

Mathematically, the densities $N(\hat{\xi}_i|z_i, \hat{\tau}_i^2)$ and $N(z_i|\hat{\xi}_i, \hat{\tau}_i^2)$ are identical, due to the symmetry of the normal distribution. It thus directly follows that

$$\int_{\mathbb{R}} N\left(z_i|\hat{\xi}_i, \hat{\tau}_i^2\right) N(z_i|\mu, \sigma^2) \, \mathrm{d}z_i = N\left(\hat{\xi}_i|\mu, \hat{\tau}_i^2 + \sigma^2\right), \tag{C2}$$

which concludes (20) from (19).

With a uniform conditional prior density $p(\mu_b|\sigma_b)$, the Gaussian conditional posterior of μ_b given $\sigma_b, \hat{\mathbf{X}}_{n_{Tb}}$ in (21) then follows directly as another standard result (see e.g. Gelman et al. 1995, page 117). More precisely, (20) is a product of Gaussian densities, which yields a Gaussian density (the log-posterior is quadratic in μ_b). The parameters of the Gaussian conditional posterior in (21) are obtained by considering the $\hat{\xi}_i$ as independent estimates of μ_b with variances $(\hat{\tau}_i^2 + \sigma_b^2)$.

[13] In the causal inference literature, covariates associated with the source/target (control/treatment) selection but not with the outcome variable are denoted as *instrumental variables*. In fact, it has been shown that the inclusion of instrumental variables in the propensity score analysis does not improve bias of the target estimates, but may lead to increased variability.

C2 Posterior distribution of σ

In our hierarchical Bayesian model, the marginal posterior distribution of the population variance σ_b can be obtained via

$$p(\sigma_b|\hat{\mathbf{X}}_{\mathbf{n_{Tb}}}) \propto p(\sigma_b) V_\mu^{1/2} \prod_{j=1}^{J} \left(\tau_j^2 + \sigma_b^2\right)^{-1/2} \exp\left(-\frac{(\hat{\xi}_j - \tilde{\mu}_b)^2}{2\left(\hat{\tau}_j^2 + \sigma_b^2\right)}\right) \tag{C3}$$

with $\tilde{\mu}_b$ and V_{μ_b} as defined in (22). Choosing a uniform prior on σ_b, $p(\sigma_b) \propto 1$, makes (36) a proper posterior density (see e.g. Gelman et al. 1995, page 117).

C3 Justification of stacked population variance estimate

In this section, we justify the stacked estimator of the marginal redshift population distribution $p_b(z)$ in (23). Precisely, we can express $p_b(z)$ via

$$p_b(z) = \int p_b(z|x) p_b(x)\, \mathrm{d}x, \tag{C4}$$

with x being photometric magnitudes/colours, $p_b(x)$ being the distribution of the covariates (magnitudes/colours) of bin b, and $p_b(z|x)$ being the conditional distribution of redshift z given covariates x of bin b. By assuming the set of photometric magnitudes/colours are finite, and assuming that the conditional distribution of z given x is the same for all bins (i.e. $p_b(z|x) = p(z|x)$), we obtain

$$p_b(z) = \sum_i p(z|x = x_i) p_b(x = x_i). \tag{C5}$$

We have an estimate of the conditional densities $p(z_i|x_i) \approx f(z_i|x_i)$. Since $z_i \overset{\text{iid}}{\sim} p(z)$, it holds that $p(z|x = x_i) = p(z_i|x = x_i) \approx \hat{f}(z_i|x_i)$. Further, we can approximate $p_b(x = x_i)$ by counting occurrences of x_i in the sample of observed magnitudes/colours within tomographic bin b. Alternatively, averaging over all estimated conditional densities $\hat{f}(z_i|x_i)$ of galaxies in bin b directly incorporates these occurrence frequencies, leading to the estimator in (23).

C4 Posterior sampling without Gaussian replacement

In this section, we detail an alternative Bayesian model that directly employs the non-parametric conditional density estimates $\hat{f}(z_i|x_i) \approx p(z_i|x_i)$, instead of replacing $\hat{f}(z_i|x_i)$ with a Gaussian approximation as described in the methods developed in Section 3.3.

C4.1 Model description

For each tomographic bin b, given the photometric data $\mathbf{X}_{\mathbf{n_{Tb}}}$, the joint posterior distribution of the population mean μ_b (our parameter of interest) and the population standard deviation σ_b, can be written via

$$p(\mu, \sigma|\mathbf{X}_{\mathbf{n_{Tb}}}) \propto p(\mu_b, \sigma_b) \prod_i \int \hat{f}(z_i|x_i) N\left(z_i|\mu_b, \sigma_b^2\right) \mathrm{d}z_i. \tag{C6}$$

In practice, we compute the conditional density estimates $\hat{f}(z_i|x_i)$ as histograms on a fine equidistant grid. More precisely, we obtain $\hat{f}(z_i|x_i) = \sum_{g=1}^{G} \hat{f}^{(g)}(z_i|x_i)\mathbb{1}_g$, where $g = 1, \ldots, G$ denote the disjoint and equidistant grid points, and $f^{(g)}$ is the density at grid location g. Evaluated on a fine grid ($G = 201$ grid points in practice), (39) can then be approximated via

$$p(\mu_b, \sigma_b|\mathbf{X}_{\mathbf{n_{Tb}}}) \propto p(\mu_b, \sigma_b) \times \prod_i \sum_g \Delta_{\text{grid}} f^{(g)}(z_i|x_i) N\left(z_i^{(g)}|\mu_b, \sigma_b^2\right), \tag{C7}$$

with $N(z_i^{(g)}|\mu_b, \sigma_b^2)$ denoting the normal density at grid point g, and Δ_{grid} being the width of each grid bin.

For the reasons as described in Section 3.3, we choose an empirical Bayesian approach, employing the stacked estimate $\hat{\sigma}_b^2$ for the population variance parameter σ_b^2. With $\sigma_b = \hat{\sigma}_b$ fixed, and adopting a uniform conditional prior density $p(\mu_b|\sigma_b) \propto 1$ (as in Section 3.3), we obtain

$$p(\mu_b|\mathbf{X}_{\mathbf{n_{Tb}}}, \hat{\sigma}_b) \propto \prod_i \sum_g \Delta_{\text{grid}} f^{(g)}(z_i|x_i) N\left(z_i^{(g)}|\mu_b, \hat{\sigma}_b^2\right). \tag{C8}$$

C4.2 Computation

We implement a simple Metropolis algorithm to obtain a posterior sample of (41). For computational reasons, we use a subset of 5 per cent (600 000 galaxies) of the target set for our analysis in this section. Galaxies are assigned to tomographic bins using the *StratLearn* binning described in Section 3.2. For each tomographic bin, we then use the Metropolis sample to obtain estimates $\hat{\mu}_b$ of the posterior means, which are then used as the point estimates of the population means μ_b.[14]

Table C1 presents the bias and SD computed over 100 lines of sight. Employing the conditional densities $f(z_i|x_i)$ led to substantially larger bias, averaging 0.0140 across the five tomographic bins, compared to the *StratLearn*-Bayes models employing the Gaussian replacements of the conditional density estimates, which yielded an average bias across the five bins of 0.0052 (see Table 3). The increase in average bias is attributable to larger biases in tomographic bins two and five. We note that we found a similar pattern with various simulation settings, using different subsets of the shear-measurement reweighted photometric sample D_T, and the non-reweighted photometric sample D_T^u (introduced in Section 5.1). This illustrates that the model with Gaussian replacements appears to perform better in various simulation settings.

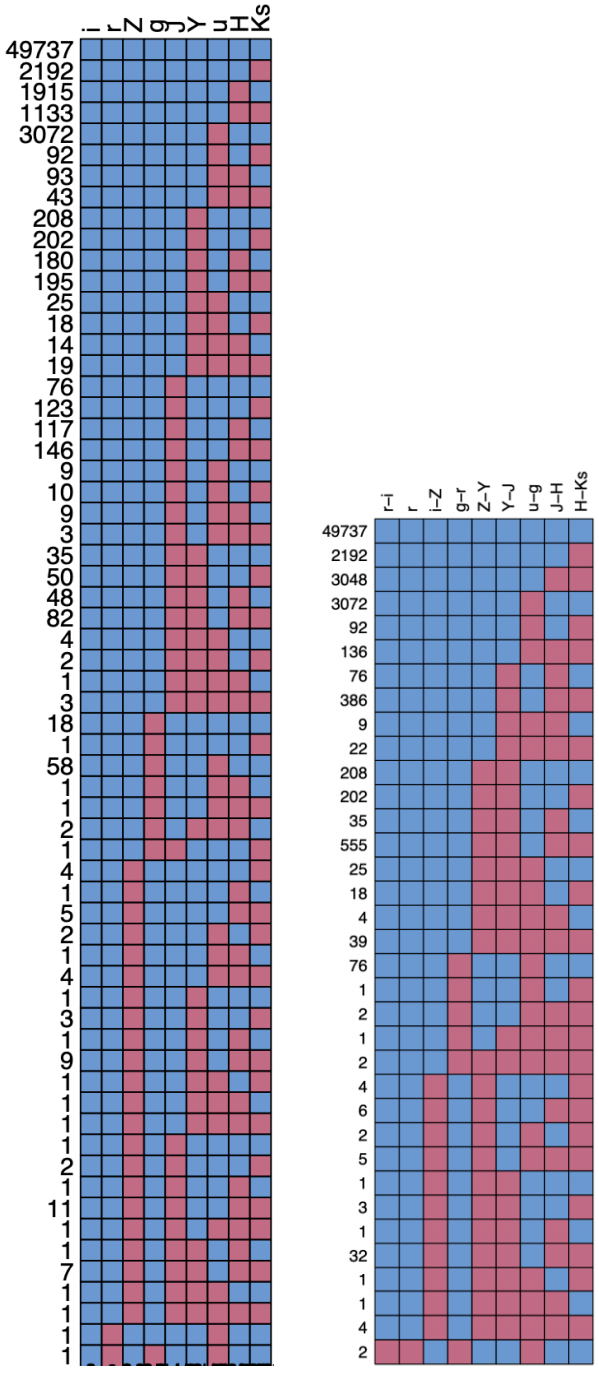
Table C1. Mean discrepancy (bias) and standard deviation (SD) computed over 100 lines of sight obtained for the Bayesian model described in Section C4, using the conditional densities $f(z_i|x_i)$ directly for the object level distributions instead of the Gaussian replacements.

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	|Average|
Bias	−0.0069	−0.0223	−0.0046	−0.0082	0.0279	0.0140
SD	0.0101	0.0084	0.0054	0.0061	0.0087	0.0077

APPENDIX D: ADDITIONAL FIGURES

This section presents additional figures, as previously referred to in the main paper. More precisely, the figures provide additional data/simulation study details, such as Figs D1 and D2; and additional numerical results, Figs D3, D4, D5, D6, D7, D8, D9, D10, and D11.

[14]For each posterior sample (each tomographic bin and line of sight), the Metropolis sampler was run for 4000 iterations, leading to an effective sample size of around 400–800, treating the first 1000 iterations as burn-in. Visual inspections of trace plots and auto-correlation plots indicate well converged chains.



(a) Magnitudes. (b) Colors.

Figure D1. 'Missing data' pattern for (a) magnitudes, and (b) colours, of a random subsample of 60 K galaxies from the photometric survey.

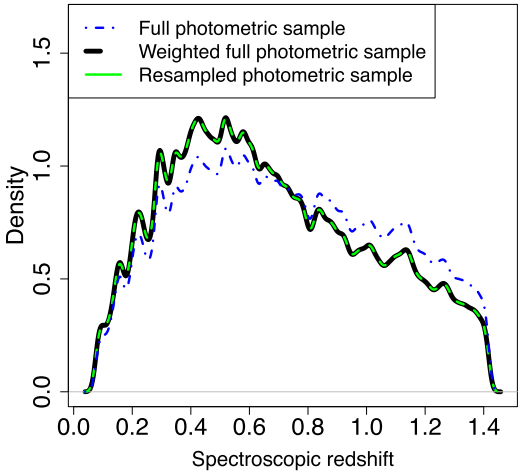


Figure D2. Spectroscopic (true) redshift distributions of the photometric survey (not known in practice), with and without incorporation of shear-measurements weights. The blue dashed line shows the redshift distribution of the full photometric sample D_T^u (before resampling). The black line shows the redshift density of the full photometric density weighted by the shear-measurement weights $\hat{w}$, and the green line illustrates the redshift density of the resampled sample D_T. The black and green line perfectly match, with a mean difference of $\sim 10^{-4}$, illustrating that there is negligible difference of targeting the resampled distribution (as in this study) and targeting the weighted distribution (as in W20).

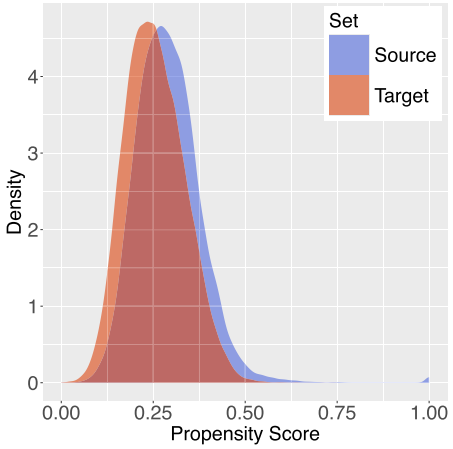


Figure D3. Propensity score distributions of source and target data in Section 5.

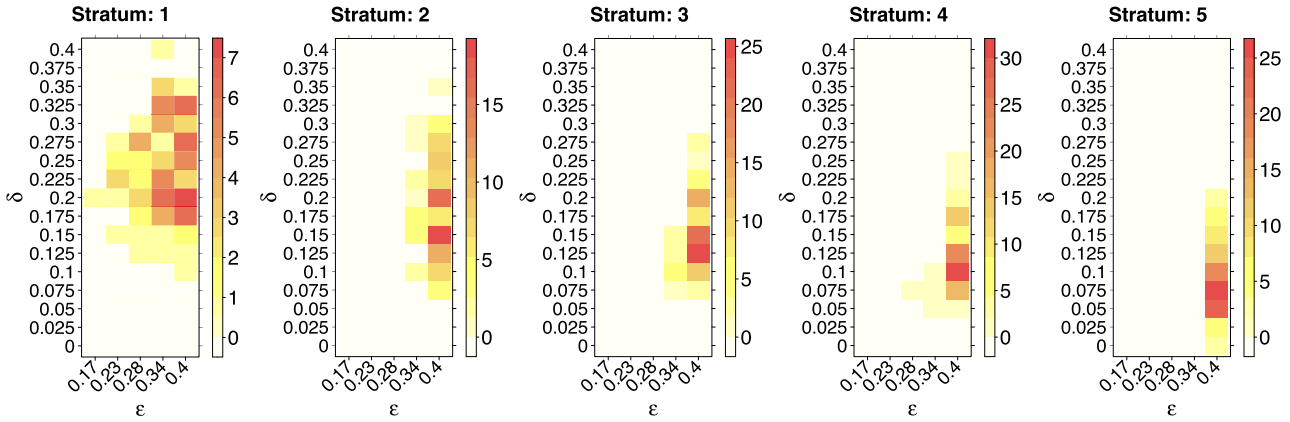


Figure D4. Illustration of the optimized hyperparameters for the Series conditional density estimator (described in Section 3.1). The heatmaps illustrate the prevalence of the various hyperparameter combinations across the 100 LoS for each stratum, respectively. Hyperparameters were optimized as described in Sections 3.1 and 5.1. We note that for the ϵ hyperparameter two additional grid values (0.05 and 0.11) were available, but never selected as the optimal hyperparameter, and thus not illustrated here for better readability. In addition, we note that our preliminary investigation showed that values for ϵ which were greater than 0.4 did not (or only marginally) lead to risk improvements; we thus used 0.4 as the maximum value for the ϵ grid.

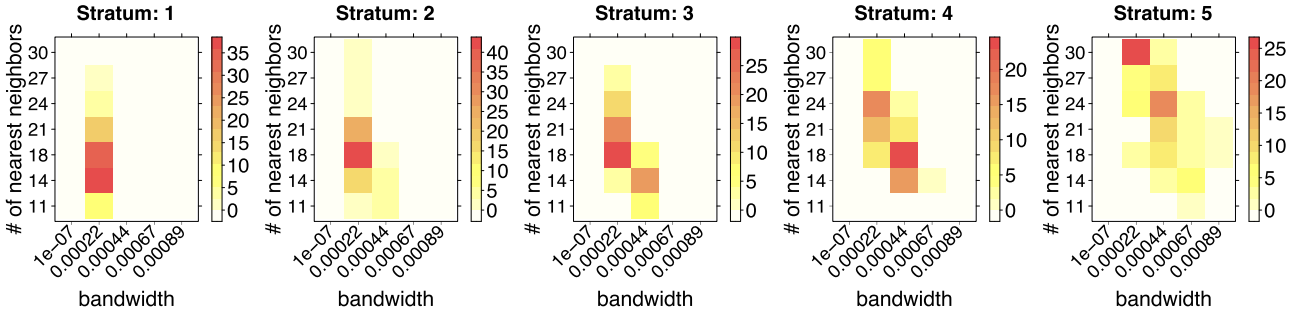


Figure D5. Illustration of the optimized hyperparameters for the Ker-NN conditional density estimator (described in Section 3.1). The heatmaps illustrate the prevalence of the various hyperparameter combinations across the 100 LoS for each stratum, respectively. Hyperparameters were optimized as described in Sections 3.1 and 5.1. For the 'bandwidth' hyperparameter five additional grid values (0.0011 0.0013 0.0015 0.0018 0.002) were available, but never selected as the optimal hyperparameter. For the '# of nearest neighbours parameter' parameter three additional grid values (2,5,8) were available but never selected as the optimal hyperparameter. These grid values are not illustrated in the heatmaps for better readability.

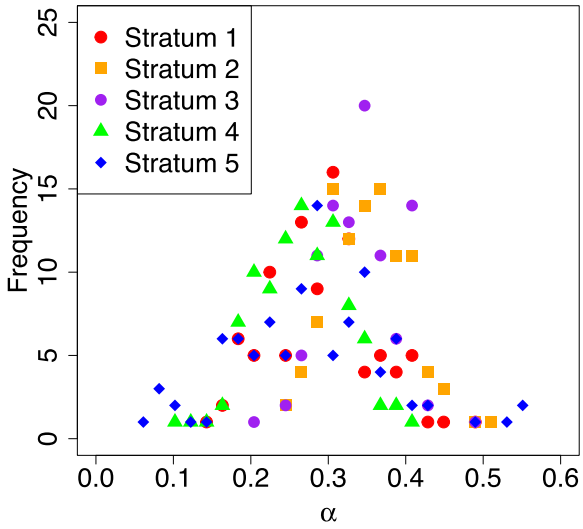


Figure D6. Frequency of the optimized α hyperparameter for the Comb conditional density estimator (described in Section 3.1) across the 100 LoS, for the five strata respectively. The α hyperparameter was optimized as described in Sections 3.1 and 5.1.

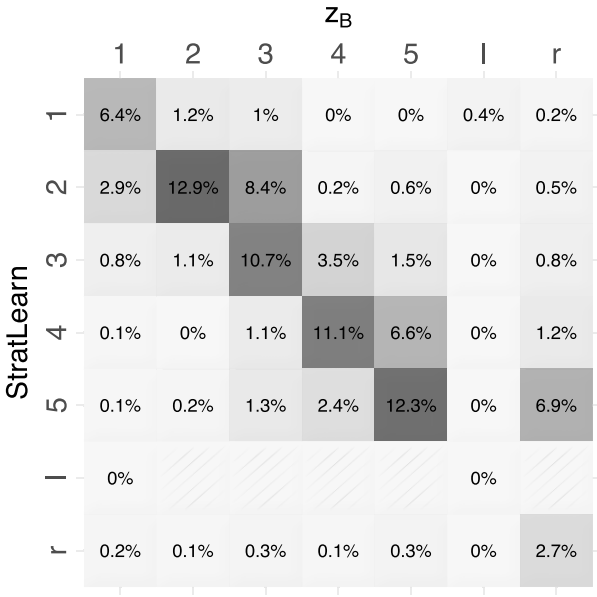


Figure D7. Changes in bin assignment using *StratLearn* versus z_B binning, averaged across the 100 LoS.

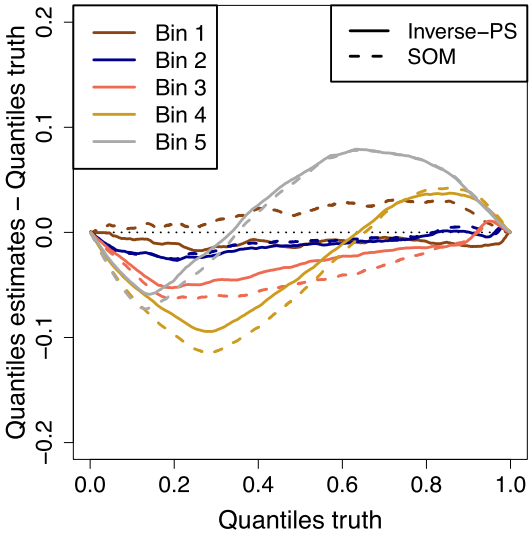


Figure D8. Modification of the pp-plots illustrated in Fig. 8 to visualize differences between the estimated distributions (Inverse-PS and SOM) with their underlying ground truth (full photometric truth and gold selected truth); modified by subtracting the *x*-axis values (the quantiles of the true distributions) from the *y*-axis values (the quantiles of the estimated distributions) in each tomographic bin illustrated in Fig. 8. Solid lines illustrate the inverse-PS results, and dashed lines illustrate the SOM results. The 95 per cent intervals (vertical bars) are omitted for clarity.

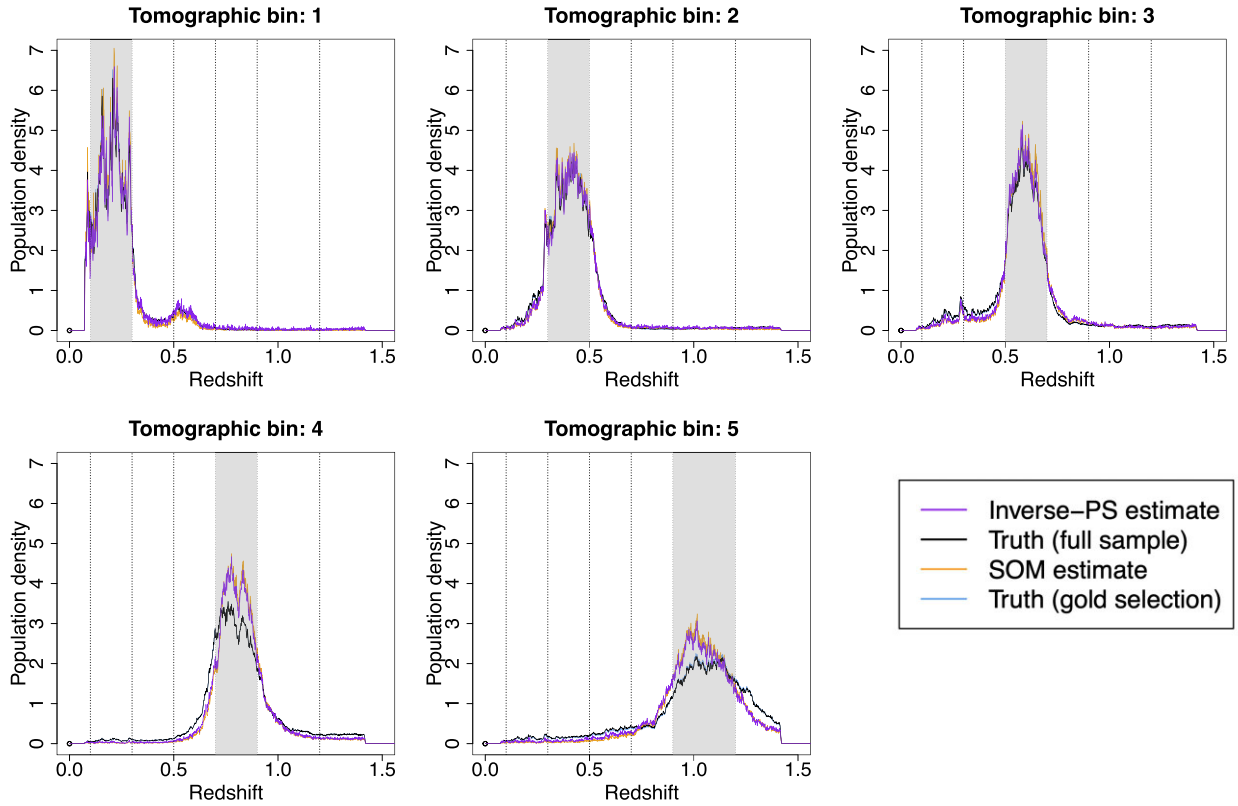


Figure D9. The same as Fig. 7, but without additional Gaussian Kernel smoothing of the population distributions. More precisely, the figure illustrates the redshift population distribution (estimates) per tomographic bin, with tomographic bins obtained as described in Section 3.2 via *StratLearn*-based binning. The figure illustrates the inverse-PS (purple) and SOM (orange) distribution estimates. The underlying true photometric redshift population distributions per tomographic bin (not known in practice) are illustrated in black for the full sample truth, and in light blue for the gold selected true distributions. The averaged (estimated) distributions across the 100 LoS are illustrated per tomographic bin.

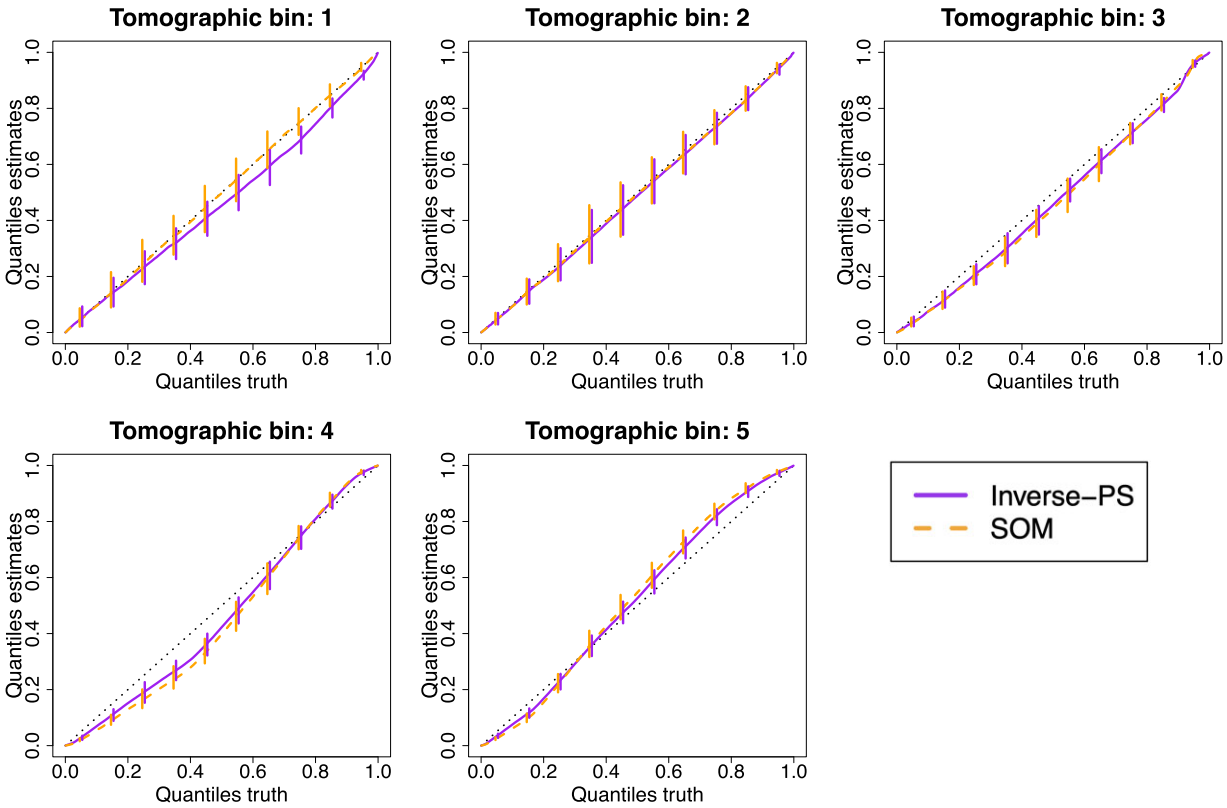


Figure D10. The same as in Fig. 8, but on z_B-based binning instead of *StratLearn*-based binning. More precisely, the figure presents pp-plots for the inverse-PS estimated distributions versus the true (full) photometric redshift distributions in purple lines, and pp-plots of the SOM estimated distributions versus the gold selected true distributions in orange dashed lines, based on the z_B tomographic binning (following Section 3.2). For each tomographic bin, the averaged pp-plots across the 100 LoS are presented, with vertical bars illustrating 95 per cent intervals indicating the range of the central 95 pp-plot lines from the 100 LoS. In bin 1, the SOM pp-plot line is closer to the 45° line, while in tomographic bins 3 to 5 the inverse-PS pp-plot line is slightly closer to the 45° line, with almost identical performance in tomographic bin 2. Given the z_B-based binning none of the estimators (inverse-PS or SOM) is thus consistently closer to its underlying ground-truth (throughout the tomographic bins).

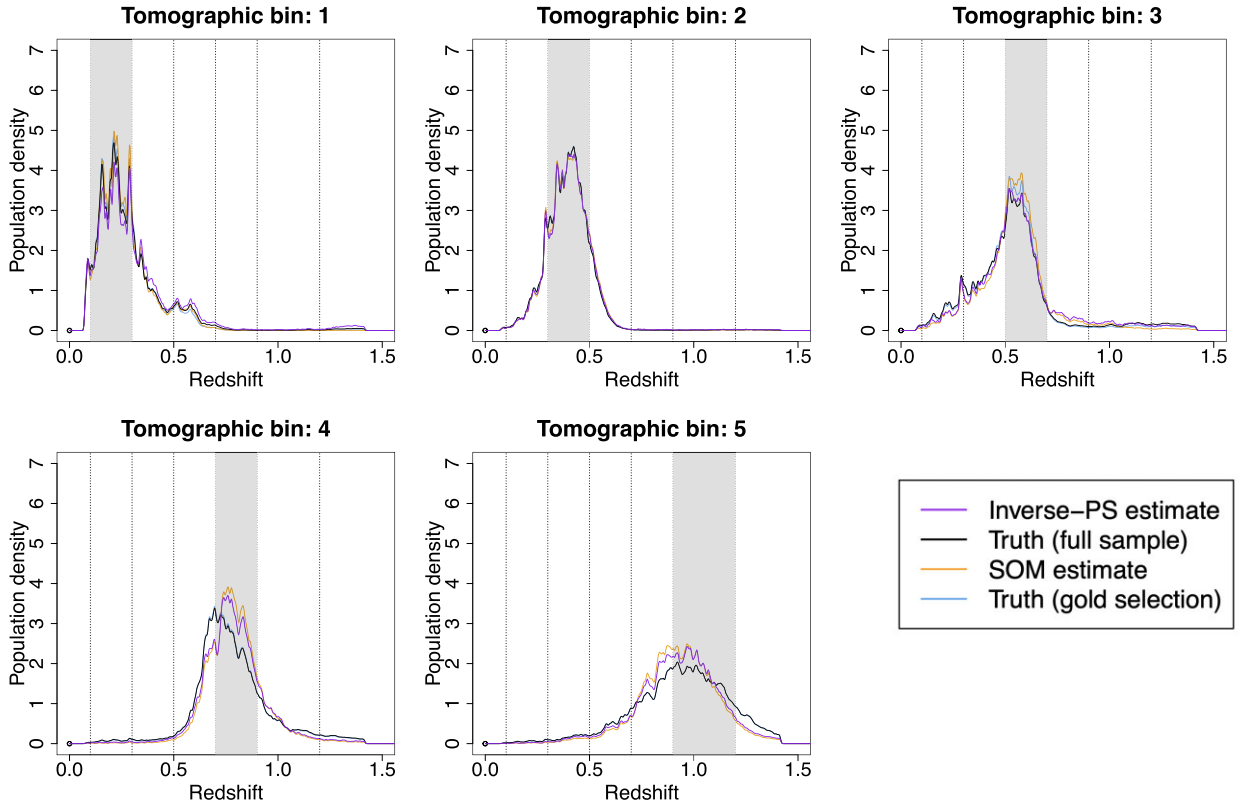


Figure D11. The same as in Fig. 7, but on z_B-based binning instead of *StratLearn*-based binning. More precisely, the figure illustrates the redshift population distribution (estimates) per tomographic bin, with z_B-based tomographic binning. The figure illustrates the inverse-PS (purple) and SOM (orange) distribution estimates. The underlying true photometric redshift population distributions per tomographic bin (not known in practice) are illustrated in black for the full sample truth, and in light blue for the gold selected true distributions. The averaged (estimated) distributions across the 100 LoS are illustrated per tomographic bin.

This paper has been typeset from a TEX/LATEX file prepared by the author.