Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **444,** 129–146 (2014) doi:10.1093/mnras/stu1424

Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements

Carlos E. Cunha,[1,2⋆] Dragan Huterer,[1] Huan Lin,[3] Michael T. Busha[2,4] and Risa H. Wechsler[2,5,6]

[1]*Department of Physics, University of Michigan, 450 Church St, Ann Arbor, MI 48109-1040, USA*
[2]*Kavli Institute for Particle Astrophysics and Cosmology 452 Lomita Mall, Stanford University, Stanford, CA 94305, USA*
[3]*Center for Particle Astrophysics, Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*
[4]*Institute for Theoretical Physics, University of Zurich, CH-8057 Zurich, Switzerland*
[5]*Department of Physics, Stanford University, Stanford, CA 94305, USA*
[6]*SLAC National Accelerator Laboratory, 2575 Sand Hill Rd., MS 29, Menlo Park, CA 94025, USA*

Accepted 2014 July 14. Received 2014 July 14; in original form 2012 August 6

ABSTRACT

We use *N*-body-spectrophotometric simulations to investigate the impact of incompleteness and incorrect redshifts in spectroscopic surveys on photometric redshift training and calibration and the resulting effects on cosmological parameter estimation from weak lensing shear–shear correlations. The photometry of the simulations is modelled after the upcoming Dark Energy Survey and the spectroscopy is based on a low/intermediate-resolution spectrograph with wavelength coverage of $5500 < \lambda < 9500$ Å. Spectroscopic follow-up surveys suffer from both incompleteness (inability to obtain spectroscopic redshifts for certain galaxies) and wrong redshifts. Encouragingly, we find that a neural network-based approach can effectively describe the spectroscopic incompleteness in terms of the galaxies' colours, so that the spectroscopic selection can be applied to the photometric sample. Hence, we find that spectroscopic incompleteness yields no appreciable biases to cosmology, although the statistical constraints degrade somewhat because the photometric survey has to be culled to match the spectroscopic selection. Unfortunately, wrong redshifts have a more severe impact: the cosmological biases are intolerable if more than a per cent of the spectroscopic redshifts are incorrect. Moreover, we find that incorrect redshifts can substantially degrade the perceived accuracy of training set based photo-z estimators, though the actual accuracy is virtually unaffected. The main problem is the difficulty of obtaining redshifts, either spectroscopically or photometrically, for objects at $z > 1.3$. We discuss several approaches for reducing the cosmological biases, in particular finding that photo-z error estimators can reduce biases appreciably when the photo-z errors are correlated with the spectroscopic failures, but not otherwise.

Key words: galaxies: distances and redshifts – galaxies: photometry – cosmological parameters – large-scale structure of Universe.

1 INTRODUCTION

Large-scale structure surveys benefit enormously from the information about galaxy redshifts. The redshift information reveals the third spatial dimension of a galaxy survey, enabling a much more accurate mapping of the expansion and growth history of the Universe relative to the case when only angular information is available.

Unfortunately, obtaining spectroscopic redshifts for all galaxies is typically impossible in wide-field imaging surveys due to the large number ($\sim 10^8$–10^9) of galaxies and the high cost of spectroscopy, especially for the high-redshift galaxies. To circumvent this problem, the current approach in the community is to estimate redshifts using photometric measurements, i.e. fluxes from a few broad-band filters. These redshift estimates are known as photometric redshifts, or photo-zs, and are necessarily coarser than spectroscopic redshifts. Because of the intrinsically large errors, photo-zs typically cannot be used directly for cosmological analysis, unless the photo-z error distributions can be quantified precisely.

⋆ E-mail: ccunha@stanford.edu

The standard approach to quantify, or calibrate, the photo-z error distributions is to use a small subsample of galaxies with known redshifts. As discussed in detail in Cunha et al. (2012), spectroscopic samples used to train photo-zs (cf. Section 4.2.1) need to be locally (in the space of observables) representative subsamples of the photometric samples. For calibration of the photo-z *error distributions*, however, the spectroscopic sample must be globally representative. More specifically, the ideal spectroscopic survey should satisfy the following properties.

(i) *Large area.* A spectroscopic survey needs to span a large area to beat down sample variance, and has to have tens of thousands of galaxies to beat down shot noise in the photo-z error calibration (Cunha et al. 2012). In addition, the spectroscopic sample needs to be imaged under conditions that faithfully reproduce the variations in the full photometric sample (see e.g. Nakajima et al. 2012). Note that requirements might be alleviated with a correction to the individual galaxy redshift likelihoods (Bordoloi, Lilly & Amara 2010; Bordoloi et al. 2012). In the context of dark energy parameter constraints, however, a full analysis that goes beyond the overall redshift distribution and involves the full error matrix $P(z_s|z_p)$ is required (Bernstein & Huterer 2010, hereafter BH10; Hearin et al. 2010).

(ii) *High completeness.* The spectroscopic survey needs to span the same range of redshifts, galaxy types, and other observational selection parameters as the photometric survey. When this is not possible, we say that the survey is *incomplete*. In that case, the photometric survey has to be culled to ensure both surveys have matching selections. Alternatively, the galaxies in the spectroscopic survey can be weighted so as to reproduce the statistical properties of the photometric sample. Achieving high completeness in faint spectroscopic surveys is a major challenge.

(iii) *Few wrong redshifts.* We show in this paper that spectroscopic surveys need to have extremely accurate redshifts. As shown by many authors (e.g. Huterer et al. 2006; Ma, Hu & Huterer 2006; Amara & Refregier 2007; Abdalla et al. 2008; Kitching, Taylor & Heavens 2008; Ma & Bernstein 2008; Hearin et al. 2010) the photo-z calibration requires exquisite knowledge of the photo-z error distribution. Errors in the spectroscopic redshifts impair the characterization of the photo-z errors and severely degrade our ability to extract cosmological constraints from photometric surveys.

For fixed observing resources, there is a conflict between accurate redshifts and completeness goals: as we stretch the observational limits (i.e. by observing very faint galaxies) to sample redshifts that would mimic the distribution of the photometric sample, we increase the fraction of incorrect spectroscopic redshifts. As we will show, redshift accuracy is more important for the upcoming surveys. Similarly, for existing instruments, the requirements on large area and number of objects implies that redshifts will have to be obtained with low- and mid-resolution spectroscopy. With low resolution, is it harder to get redshifts for galaxies with narrow emission lines that fall in between the night sky lines. As a result, it is harder to reach high completeness with realistic integration times and the selection functions of low-resolution surveys are harder to quantify. Until the next generation of spectroscopic instruments becomes available, existing and upcoming photometric surveys will have to learn to deal with very incomplete spectroscopic samples for photo-z calibration.

The purpose of this paper is to assess the impact of spectroscopic selection, i.e. completeness and accuracy, on the training and calibration of photometric redshifts and the resulting impact on cosmological constraints derived from weak lensing (WL) shear–shear correlations. To achieve this goal, we combine N-body, photometric and spectroscopic simulations patterned after the proposed characteristics of the Dark Energy Survey (DES) and expected spectroscopic follow-up. We then propagate the errors due to imperfect photo-z calibration on the cosmological parameter constraints inferred from the weak gravitational lensing power spectrum observations forecasted for the DES.

The paper is organized as follows. In Section 2, we provide a pedagogical introduction to the main issues driving completeness and accuracy of a spectroscopic sample. In Section 3, we briefly describe the simulated catalogues we use, leaving the details of the catalogue generation to Appendix A. In Section 4, we give a step-by-step guide describing how we go from the simulated data to the cosmological constraints, detailing the methods used at each step. Results are presented in Sections 5 and 6. We discuss the robustness of our assumptions in Section 7 and the implications of our findings for spectroscopic survey design in Section 8. We present conclusions in Section 9.

2 BASICS OF LOW-RESOLUTION SPECTROSCOPY

In this section, we provide a brief pedagogical overview of issues in spectroscopy, targeted to theorists.

2.1 Key parameters of spectroscopic surveys

Spectroscopic redshifts are often derived by cross-correlating a library of galaxy templates with observed (or simulated) spectra. For fixed observing conditions (and in the absence of instrumental systematic effects), three main items determine the quality of the estimated spectroscopic redshifts.

(i) *Spectral coverage*. The wavelength range covered by the spectrograph needs to bracket a few significant spectral features. As shown in the bottom plot of Fig. A1, for our simulation the coverage is roughly from 5500 to 9500 Å, with decreasing sensitivity at longer wavelengths.

(ii) *Integration time*. The faintest galaxies detectable by upcoming optical surveys can be a few orders of magnitude fainter than the atmospheric emission. Thus, significant integration times, as well as careful subtraction of the sky background, are needed to obtain secure redshift measurements.

(iii) *Cross-correlation templates*. Having an accurate and representative set of galaxy spectral distribution templates is important in deriving accurate redshifts and associated uncertainties. As we discuss in the next section, this is particularly important for early-type galaxies and galaxies at $z > 1.5$ (also known as the *redshift desert*) because of the lack of strong emission features in the spectrograph window.

2.2 Principal spectral features

The two main emission lines used in optical spectroscopy are the [O II] (singly ionized oxygen) line at 3727 Å and the Hα (first transition in the Balmer series) line at 6563 Å. The main absorption feature is the 4000 Å break, caused by a confluence of absorption lines, particularly the H&K calcium lines. In high-resolution spectroscopy, [O II] is the most important line because it is actually a doublet – a pair of closely spaced lines. High-resolution observations – e.g. with DEEP2 (Newman et al. 2012), or SDSS (York et al. 2000) – can distinguish the doublet and hence confidently identify

[O II]. Low-resolution observations – e.g. as in the VIMOS-VLT Deep Survey (VVDS; Le Fèvre et al. 2005) and *z*COSMOS surveys (Lilly et al. 2007), rely on more than one feature. The limited spectral range of the instrument sets the regions of redshift space where one can confidently identify spectral features. In the case of VVDS, for example, there are roughly five different redshift regions:

(i) $z < 0.4$. The Hα can be detected, but [O II] cannot. There is risk of confusing Hα of a $z < 0.4$ galaxy for [O II] emission of a galaxy at $z > 0.8$. Fortunately, these galaxies are mostly brighter and thus the Hα line combined with less prominent spectral features is often sufficient to estimate a redshift. Similarly, for early-type galaxies, the 4000 Å break cannot be detected, and one relies on smaller absorption features.

(ii) $0.4 < z < 0.6$. Neither [O II] nor Hα can be detected. Redshifts have to be estimated based on [O III] and Hβ lines.

(iii) $0.6 < z < 0.9$. [O II] and other important lines ([O III] – 5007 Å, Hβ – 4861 Å) are detectable, but get progressively fainter towards higher redshift (due to increasing atmospheric noise and instrumental sensitivity).

(iv) $0.9 < z < 1.5$. [O III] and Hβ are out of the instrument range, but [O II] and the 4000 Å break are still detectable.

(v) $z > 1.5$ (the redshift desert). Only minor features in the spectra are available. Visual inspection to reduce incompleteness is essential in this range. Potential for wrong redshifts is increased because atmospheric emission lines can be mistakenly identified by the algorithm as real lines.

2.3 Additional systematics affecting the incompleteness

There are a few additional items contributing to the incompleteness of spectroscopic surveys that are not modelled in our simulations but that exist in real surveys.

(i) *Fibre collisions and slit overlaps*. If the angular separation between galaxies is too small, one may not simultaneously obtain their spectra (without using a multiple pass strategy). Since clustering of galaxies is type dependent, one has to be careful that fibre collisions and slit overlaps do not introduce selection biases.

(ii) *Optical distortions*. Geometric distortions due to the spectrograph optics may make extraction of spectra and subsequent measurement of redshifts more difficult near the edge of the instrument field of view.

(iii) *CCD fringing*. Spatial and wavelength dependent variations in the pixel response in the red end of the spectrograph. Fringing hinders measurement of the spectra and redshifts of faint galaxies.

(iv) *Stars and bright galaxies*. Light from nearby stars or bright galaxies can contaminate the spectra.

(v) *Cosmic rays*. Also can contaminate the spectra.

Issues such as stars, cosmic rays and edge effects will reduce the completeness, more or less randomly, resulting mostly in an increase in the shot noise, without galaxy type or redshift dependence. Note also, that photometric surveys are affected by countless other selection systematics. We are only concerned with systematic *differences* between the spectroscopic sample relative to the photometric sample, however it is defined.

3 SIMULATED DATA

We use cosmological simulations populated with galaxies and their photometric properties as described in Appendix A1. The photometric observations are patterned after the expected sensitivity of the DES and Vista Hemisphere Surveys, with galaxies imaged in the grizYJHKs filters over 5100 sq. degrees. For simplicity, we only use the observations on *griz* bands because they are imaged for longer periods of time, and hence are useful for all our sample. The imaging in these bands is expected to reach 10σ magnitude limits of 25.2, 24.7, 24.0, and 23.5 in *g, r, i* and *z*.

For computational efficiency, we select a subsample of approximately 1.3 million galaxies, hereafter our *photometric sample*, from the total 1 billion galaxies present in the simulation. We apply the same quality cuts as in Cunha et al. (2012), i.e. keep galaxies with $i < 24$ and at least 5σ detection in *grz*. This selection reduces our photometric sample to 726 824 galaxies.

Of this photometric sample, we randomly target a subset of 181 892 galaxies, hereafter the *spectroscopic sample* or *training set*, for the spectroscopic analysis. The generation of simulated spectra for this subsample is described in the Appendix A2.

4 FROM THE REDSHIFTS TO COSMOLOGY

In this section, we describe the step-by-step procedure we used for converting the simulated observations into cosmological constraints. The flowchart in Fig. 1 gives a pictorial version of the explanation below.

(i) The first step is to estimate spectroscopic redshifts for the sample for which we have spectra. We use the `rvsao.xcsao` spectral analyser algorithm described in Section 4.1. Not all spectra yield redshifts, and only the redshifts above certain confidence are kept. Even so, a fraction of the spectroscopic redshifts is incorrect.

(ii) The spectroscopic sample can only be used for calibration of the photo-*z* error distributions if it is a representative subsample of the photometric sample. Hence, we statistically match spectroscopic and photometric selection in one of two ways: by applying the spectroscopic selection to the photometric sample with neural networks (cf. Section 6.1), or by weighting the photometric sample so that its statistical properties match those of the spectroscopic sample (cf. Section 6.2).

(iii) Next, we calculate photo-*z*s for both the spectroscopic and photometric samples, cf. Section 4.2.

(iv) After the matching, we can calculate the photo-*z* error matrices required for cosmological analysis.

(v) Finally, we estimate fiducial constraints and biases in the cosmological parameters forecasted for the DES-type weak gravitational lensing survey. We break up the tests in two parts. In the first case, shown as the transparent hexagon in the flowchart, we only test the impact of the selection matching, by using only the correct value for redshifts. In the second case (grey hexagon), we use the actual value of the spectroscopic redshifts – thereby including the small fraction of wrong redshifts.

4.1 Analysing 1D spectra

Simulating spectroscopic redshift estimation is challenging because real spectroscopic surveys rely heavily on visual inspection. For our forecasts, visual inspection of thousands of spectra would be out of the question. Instead, we adopt a more reasonable strategy and apply an automated pipeline to all 1D spectra. We use the publicly available rvsao IRAF external package version 2.7.8 (Kurtz & Mink 1998). We run the cross-correlation tool `xcsao` on our simulated spectra. The algorithm performs a Fourier cross-correlation between the 'observed' (simulated) spectra and a user-defined library of template spectra. We obtain the template library used in the

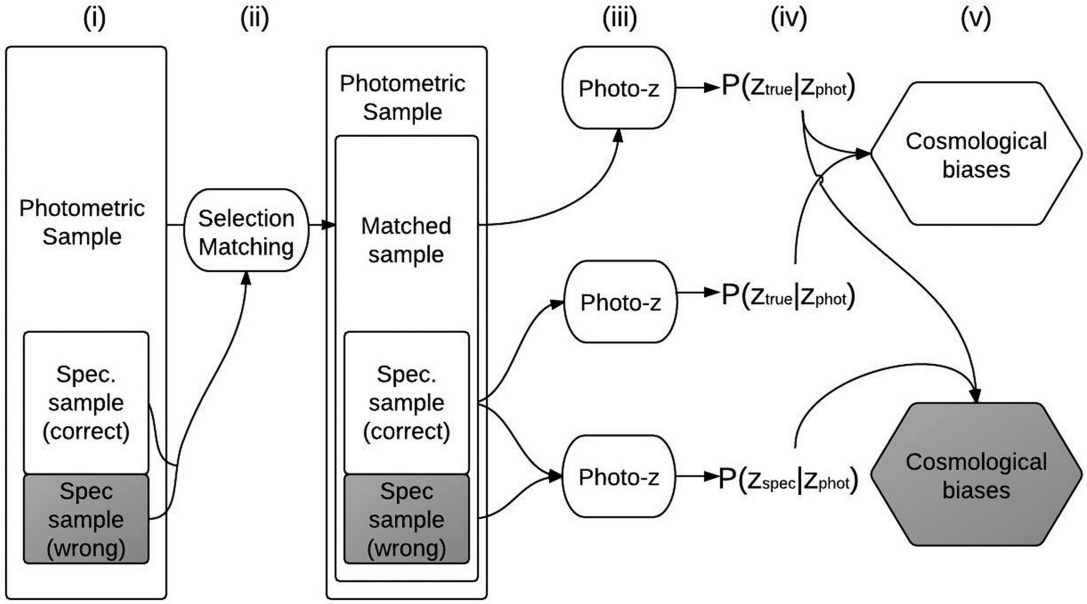

Figure 1. Flowchart describing our step-by-step procedure to go from the simulated observations to cosmological biases.

cross-correlation from the simulation itself. For the first pass, we pick six templates chosen to mimic the six galaxy templates used in the cross-correlation analysis of the SDSS spectroscopic pipeline.[1] Using templates from the simulation instead of the original SDSS templates improved the number of correct redshifts by 10 per cent. The limitation of the SDSS template basis is that it was chosen for low-redshift spectroscopy, and is not sufficient for redshifts greater than 1 or so. In the second pass, we added three templates from the simulations picked as the brightest templates above redshift 1.4 for which the cross-correlation coefficient – the R statistic described below – was less than 2.5. The additional templates doubled the number of correct redshifts above 1.4.

The cross-correlation analysis can be refined around certain wavelengths by giving it an initial redshift guess (by setting the parameter `czguess`) to start the search. We perform the analysis five times with: no guess, `czguess`=0.4, `czguess`=0.8, `czguess`=1.2 and `czguess`=1.6. We then choose which redshift estimate to keep based on the value of the R statistic, output by the pipeline. The R statistic, introduced by Tonry & Davis (1979) (cf. equation 23 of that work), is a measure of the strength of the cross-correlation given by the ratio of the height of the assumed true peak in the correlation to the average height of spurious peaks. R varies from 1 to several hundred in our simulation, and as we show later, $R > 6$ corresponds to >99 per cent correct redshifts.

We have performed our analysis for a number of settings of the spectroscopic pipeline, but only show results for three main cases, defined as follows.

(i) *Fiducial pipeline*. $z_{\rm spec}$s estimated using the $6 + 3 = 9$ templates and the five redshift guesses described above. Yields the highest completeness for $z > 1.4$.

(ii) *Comb2 pipeline*. $z_{\rm spec}$s estimated using the $6 + 3 = 9$ templates and only running `xcsao` twice, with `czguess=0.4` and `czguess=0.8`. Yields the highest overall completeness, but the lowest completeness at low and high redshift.

(iii) *Original pipeline*. $z_{\rm spec}$s estimated using the six original templates and only four redshift guesses: `czguess`=none, 0.4, 0.8 and 1.2.

4.2 Photometric redshifts

There exists a cornucopia of publicly available photometric redshift estimation algorithms. For recent reviews and comparison of methods see e.g. Hildebrandt et al. (2010) and Abdalla et al. (2011). We consider two different photo-z algorithms that broadly span the space of possibilities. We use a training-set fitting method with a very large training set and basic template-fitting code without any priors, which we briefly describe below.

4.2.1 *Training-set redshift estimators*

The basic setup of training set based redshift estimators is to use a sample with known spectroscopic redshifts to estimate the free parameters of a function relating the observables (in our case the magnitudes of the galaxies) to the redshifts. After the best-fitting free parameters have been determined, the function can be applied to the data for which no spectroscopic redshifts are available, known as the photometric sample. For this paper, we use an artificial neural network (ANN) as our training-set method, and we leave the details to Appendix B.

4.2.2 *Template-fitting redshift estimators*

Template-fitting estimators derive photometric redshift estimates by comparing the observed colours of galaxies to colours predicted from a library of galaxy spectral energy distributions (SED). We use the publicly available *LePhare* photo-z code[2] (Arnouts et al. 1999; Ilbert et al. 2006) as our template-fitting estimator. We chose the extended $\texttt{CWW}_{\rm KINNEY}$ template library, which comes with *Lephare* because it yielded the best photo-zs for our simulation. This library

[1] Templates 23 to 28 in the website: http://www.sdss.org/dr7/algorithms/spectemplates/index.html

[2] http://www.cfht.hawaii.edu/arnouts/LEPHARE/lephare.html

includes the four CWW templates (Coleman, Wu & Weedman 1980) extended to the IR and UV using templates from Bruzual & Charlot (2003) and six Kinney et al. (1996) starburst templates.

We note that a variety of public template-fitting codes are available (e.g. Coe et al. 2006; Feldmann et al. 2006), and each includes many options of template libraries, extinction laws, priors, etc. For a discussion on propagation of template-fitting uncertainties to redshift uncertainties see Abrahamse et al. (2011).

We emphasize that both categories of photo-z estimators should be thought of as performing the same function, of applying a model to describe the data. For training-set methods, the model is the training sample, whereas for template-fitting methods, the model is the set of templates and priors utilized. By construction, the training-set method we use has an excellent training set, and as a result, performs nearly optimally. Conversely, for the template-fitting method, we do not take many pains to optimize the template library nor to find appropriate priors, and thus, the template-fitting code performs substantially worse. With perfect templates and perfect priors, template fitting should equal the performance of training-set methods with perfect training. It is not our intention to suggest that any of these estimators is fundamentally superior, we just wished to test our analysis in optimistic and pessimistic regimes. Most importantly, as we show in later sections, *results are independent of the photo-z estimators used* despite the significant differences in performance.

4.3 Effect on the cosmological parameters

In Appendix C, we review the formalism from BH10 to calculate the biases in the observed WL power spectra, and hence in the cosmological parameters, given some arbitrary source systematic error. Here, we are of course interested in the systematics due to imperfect spectroscopic redshifts. It is beyond the scope of the paper to consider the biases for other cosmological probes, but we believe that the WL results will provide a useful baseline for what to expect from a single probe. In addition, we expect that a joint analysis might allow for self-calibration of the biases.

The fiducial WL survey corresponds to expectations from the DES, and assumes 5000 square degrees (corresponding to $f_{\rm sky} \simeq 0.12$) with tomographic measurements in $B = 30$ uniformly wide redshift bins extending out to $z_{\rm max} = 2.0$. The effective source galaxy density is 12 galaxies per square arcminute, while the maximum multipole considered in the convergence power spectrum is $\ell_{\rm max} = 1500$. The radial distribution of galaxies, required to determine tomographic normalized number densities n_i in equation (C1), is determined from the simulations and shown in Fig. 4.

We consider a standard set of six cosmological parameters with the following fiducial values: matter density relative to critical $\Omega_{\rm M} = 0.25$, equation of state parameter $w = -1$, physical baryon fraction $\Omega_{\rm B}h^2 = 0.023$, physical matter fraction $\Omega_{\rm M}h^2 = 0.1225$ (corresponding to the scaled Hubble constant $h \equiv H_0/(100\,{\rm km\,s^{-1}\,Mpc^{-1}}) = 0.7$), spectral index $n = 0.96$, and amplitude of the matter power spectrum $\ln A$ where $A = 2.3 \times 10^{-9}$ (corresponding to $\sigma_8 = 0.8$).

As detailed in Appendix C, we add the (unbiased) Planck forecasted constraints on the cosmological parameters to those of the DES. The fiducial (combined) constraint on the equation of state of dark energy assuming perfect knowledge of photometric redshifts is $\sigma(w) = 0.055$.

5 SPECTROSCOPIC SUCCESS AND FAILURE

In this section, we define the basic concepts regarding successful and failed galaxy spectra, and the accompanying rates.

5.1 Spectroscopic success rate

The spectroscopic analysis for the fiducial simulation parameters (16 200 s integration; 9 templates; no manual correction of spectra) yields about 74 per cent correct spectroscopic redshifts (defined as redshifts for which $|z_{\rm spec} - z_{\rm true}| < 0.01$). In a real survey, one can only choose redshifts based on some quality flag, which is the cross-correlation R statistic (described in Section 4.1) in our case. We thus define two success metrics.

(i) *True spectroscopic success rate* ($\rm SSR_T$): the fraction of galaxies with correct redshifts.

(ii) *Observed SSR* ($\rm SSR_O$): the fraction of galaxies with R greater than a certain value. Unless stated otherwise, we set the value to 6.0.

In Fig. 2, we show the $\rm SSR_T$ as a function of true redshift (left-hand panel), observed i-band magnitude (centre panel) and cross-correlation strength (right-hand panel). The left-hand panel shows that the $\rm SSR_T$ generally worsens with higher redshift, and the 'hiccups' in the curves are directly caused by different spectral lines

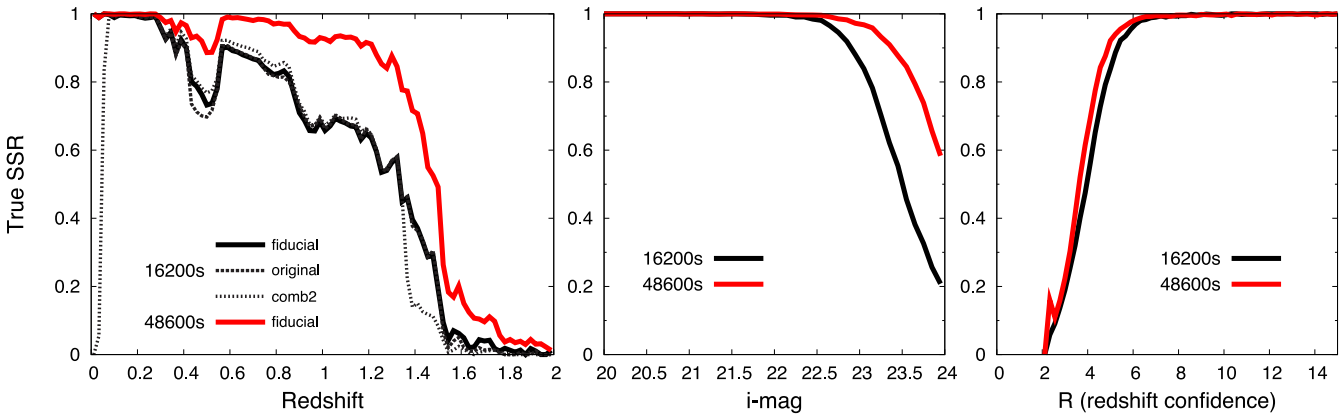

Figure 2. Left-hand panel: $\rm SSR_T$, defined as fraction of correct redshifts, as a function of true redshift. Central panel: $\rm SSR_T$ as a function of observed i-band magnitude. Right-hand panel: $\rm SSR_T$ as a function the cross-correlation strength statistic R, which is a measure of the redshift confidence. The black lines assume 16 200 s of integration time and the red (grey) lines assume 48 600 s. The solid, dashed and dotted lines correspond to different settings of the spectroscopic pipeline, described in Section 4.1.

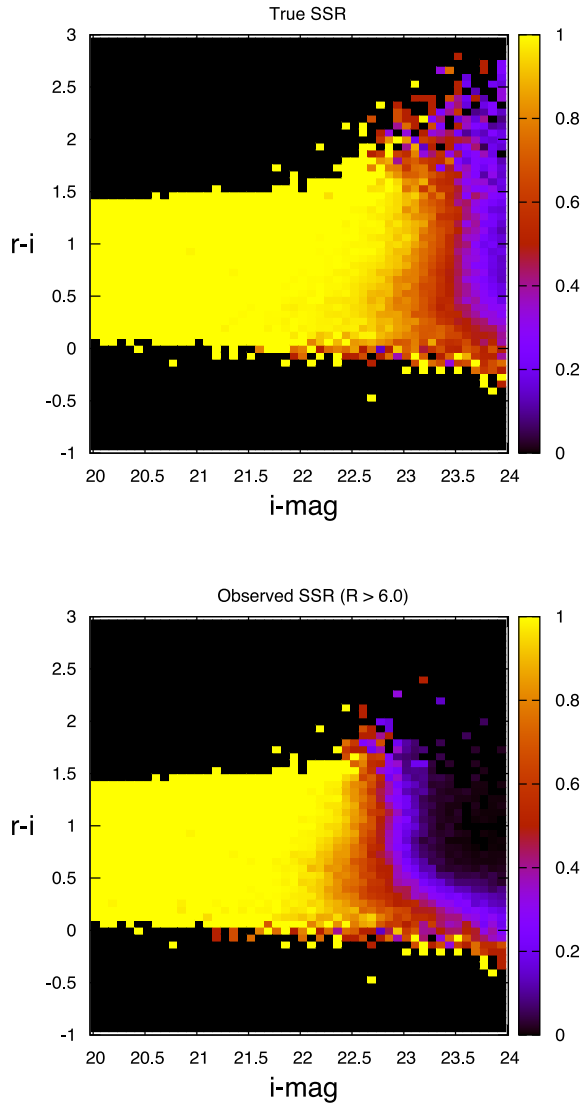

Figure 3. Top panel: true SSR (SSR_T), defined as fraction of correct redshifts as a function of true redshift. Bottom panel: observed SSR (SSR_O), defined as fraction of galaxies with correlation $R > 6.0$. Both results assume the Fiducial pipeline settings (cf. Section 4.1) of 16 200 s of integration time with the three additional templates.

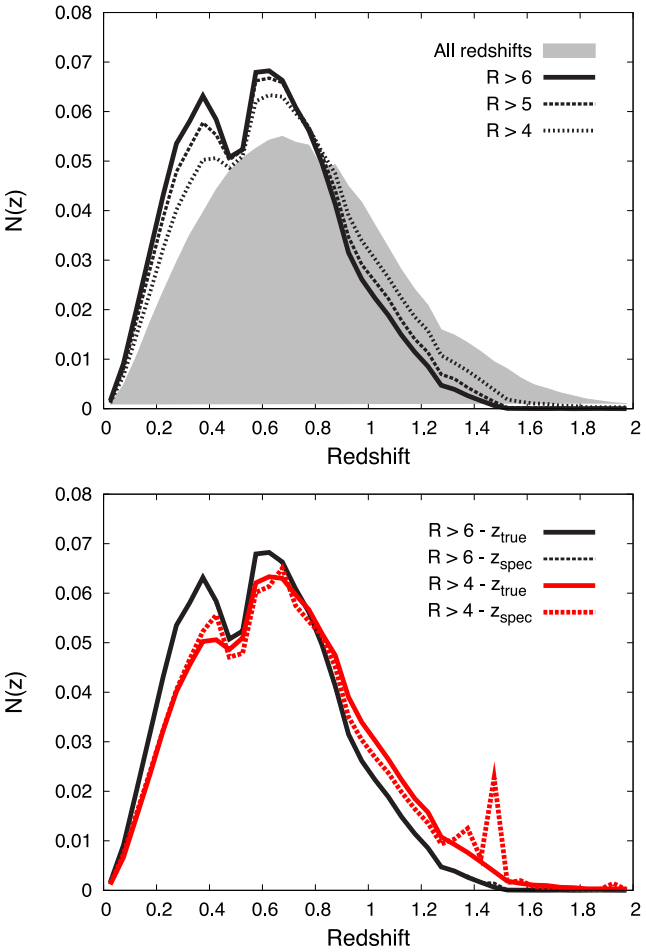

Figure 4. Top panel: distributions of true redshift for all galaxies (shaded area), galaxies with $R > 6$ (solid line), galaxies with $R > 5$ (dashed line) and galaxies with $R > 4$ (dotted line). Bottom panel: distribution of true redshift (solid lines) and spectroscopic redshift (dashed lines) for the $R > 6$ sample (black) and the $R > 4$ sample (red–grey).

which enter and leave the observed spectral range, as discussed in Section 2. The central panel shows the expected result that the SSR plunges beyond certain depth. Finally, the right-hand panel shows that the true SSR increases monotonically with cross-correlation statistic R, showing that we can use R to select an accurate redshift sample with high confidence.

In Fig. 3, we show the true and observed SSRs as a function of i-magnitude and $r-i$ colour. The top panel shows that virtually all the incorrect redshifts are at the faint end of the colour–magnitude diagram, with slight colour dependence. In particular, at the bluest end ($r–i \sim 0$) we see a region of low SSR extending to $i \sim 22$. This is typically caused by the lack of an appropriate template to describe certain galaxy populations.

The *observed* SSR, shown in the bottom panel of Fig. 3, shows a more pronounced colour variation. We can see that the bluer colours, corresponding to late spectral types, which have significant emission features, yields highest SSR_O. Conversely, the redder colours have the lowest SSR_O. As mentioned previously, early-type galaxies have virtually no emission lines, and hence are identified by absorption features. Intermediate types can have weak emission lines, but usually have weaker absorption features as well, which makes it difficult to determine a spectroscopic redshift for them.

Because of our stringent choice of cut, the sample with $R > 6.0$ contains a fraction 0.53 of the total galaxies and has 99.6 per cent correct spectroscopic redshifts. For comparison, if we define samples by the cuts $R > 5.0$ and 4.0 these would contain a fraction of 0.60 and 0.73 of total galaxies with 98.6 and 93.2 per cent correct redshifts, respectively. Faint, intermediate-type galaxy spectra yield the majority of the incorrect redshifts that escape the R selection.

In the top panel of Fig. 4 we show the effect of applying quality cuts based on the statistic R to the true-redshift distribution. More stringent (higher R) cuts preferentially remove galaxies from regions where less significant spectroscopic features fall inside the spectrograph window (as explained in Section 2). The bottom panel shows that the less stringent cuts allow for a higher fraction of incorrect redshifts, which have a visible impact in the redshift distribution even though 93.2 per cent of the redshifts are correct.

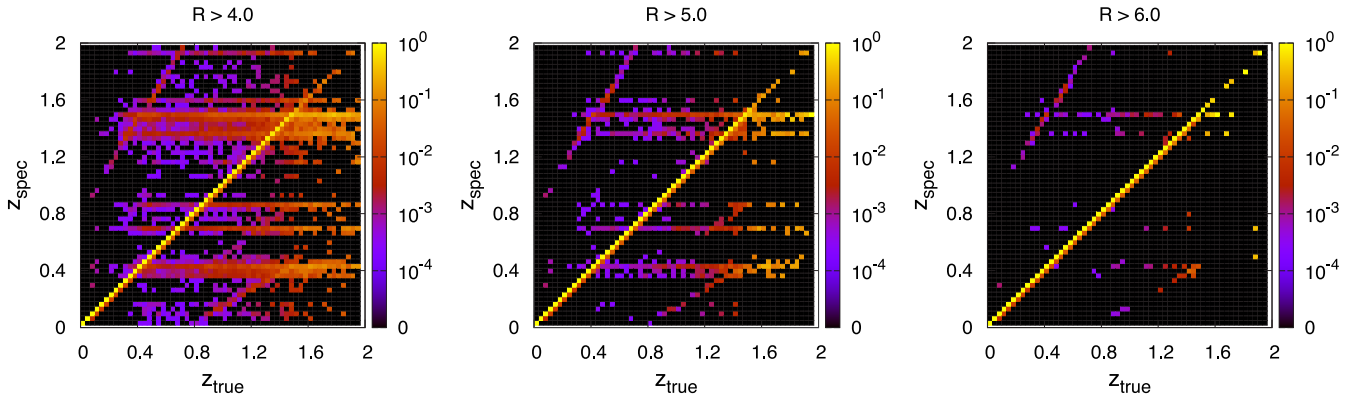

Figure 5. Leakage matrices ($P(z_{\rm spec}|z_{\rm true})$) for the training sets selected by the cuts $R > 4.0$ (left-hand panel), $R > 5.0$ (centre panel) and $R > 6.0$ (right-hand panel). The spectroscopic redshifts were calculated using 16 200 s exposures with the full set of nine templates in the spectroscopic pipeline, corresponding to our Fiducial pipeline.

5.2 Where do the wrong redshifts go?

We show the spectroscopic leakage matrices ($P(z_{\rm spec}|z_{\rm true})$) for several cuts in the R statistic for our Fiducial pipeline scenario in Fig. 5. The spectroscopic redshift errors, which correspond to any departures from the $z_{\rm spec} = z_{\rm true}$ (diagonal) line, clearly make interesting and definite patterns as follows.

(i) *Atmospheric line confusion*. Horizontal features in Fig. 5, when many different values of $z_{\rm true}$ are misinterpreted as a single $z_{\rm spec}$, correspond to cases where residuals from subtraction of atmospheric lines are confused with actual features in the galaxy spectrum.

(ii) *Galaxy line misidentification*. Diagonal lines in Fig. 5 (excepting the $z_{\rm spec} = z_{\rm true}$ diagonal, of course) correspond to the cases where the pipeline misidentifies lines of the galaxy itself due to limited spectroscopic coverage and Signal-to-Noise (S/N) (cf. Section 2.2). For example, the diagonal trend from $(z_{\rm true}, z_{\rm spec}) = (0, 0.8)$ to about $(0.7, 2.0)$ corresponds to the pipeline classifying Hα emission lines as [O II] lines. A corresponding feature due to [O II] being incorrectly classified as Hα can be seen starting at (0.8, 0) in the plots. Galaxy line misidentification seems to be a much smaller issue than atmospheric line confusion for our simulation.

The exact distribution of the wrong redshifts depends on the noise levels assumed and details of the spectroscopic analysis. As described in Appendix A2, we assumed a constant mean atmospheric emission and absorption, but in reality the observing conditions vary. The distribution of wrong redshifts also depends on details of the spectroscopic analysis. In Fig. 6, we show the $P(z_{\rm spec}|z_{\rm true})$ matrix for the Original pipeline, described in Section 4.1, which only uses the original six spectral templates (but not the three templates added to increase completeness for $z > 1.4$). In addition, it does not use the `czguess=1.6` results, which have the effect of increasing the probability that the pipeline will assign a high redshift to a galaxy. The Original pipeline is not optimized in any way towards high-z completeness, and as a result it finds no spectroscopic redshifts above $z = 1.6$. Conversely, the Fiducial pipeline (cf. right-hand plot in Fig. 5) does find some redshifts above $z = 1.6$, but at the cost of increasing the number of objects being incorrectly assigned very high values of spectroscopic redshifts and the number of objects at high redshifts being assigned very low redshifts. As we discuss in Section 6.1.2, the Original pipeline yields a bias in w a factor of 2 smaller than the Fiducial pipeline.

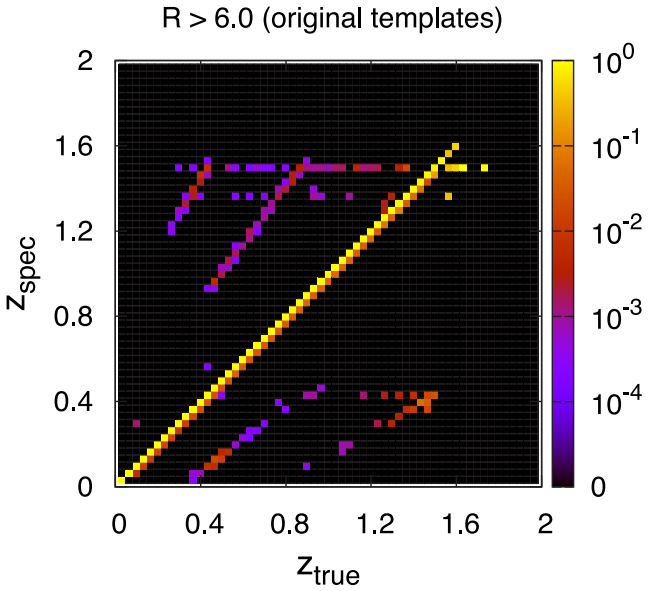

Figure 6. Same as Fig. 5 except for the Original pipeline, where only the six original templates were used, and only four different values of `czguess` (no guess, 0.4, 0.8 and 1.2) were used in the `rvsao` run. Without the three additional templates, no strong correlations were found for $z_{\rm spec} > 1.5$, which, in particular, implied that no galaxies were incorrectly assigned $z_{\rm spec} > 1.5$.

There are two points to take from this section. First, wrong spectroscopic redshifts occupy preferred regions of the $(z_{\rm true}, z_{\rm spec})$ plane. Since the exact redshift error distribution depends on the details of the spectroscopic analysis and observing conditions, it is challenging to accurately predict the spectroscopic redshift errors in real surveys. Hence, our conclusions concerning the impact of wrong redshift are necessarily only rough estimates. Secondly, increasing the completeness at high redshift can come at the expense of introducing more catastrophic spectroscopic redshifts. As we shall show in Section 6.1.2, this is a very high price to pay, and can severely increase biases in cosmological parameter constraints.

6 SPECTROSCOPIC SELECTION MATCHING

As can be inferred from the top panel in Fig. 4, spectroscopic failures alter the redshift distribution of the training set significantly, so that

one cannot use such a sample to estimate the error distributions of the photometric sample directly. We test two different approaches to correct for the selection effects in the training set.

One approach, presented in Section 6.1 is to cull the photometric sample to remove all galaxies that are not represented in the training set (the set of high-confidence spectroscopic redshift galaxies). The other approach, described in Section 6.2 is to apply weights to the spectroscopic sample so that the statistical properties of the weighted spectroscopic galaxies closely match those of the photometric sample. Thus, in the first approach, the photometric sample is modified to match the spectroscopic sample, whereas in the second approach, it is the spectroscopic sample that is modified.

6.1 Culling approach to selection matching

In the culling approach, we use a neural network (described in Appendix B) to accomplish the selection matching. What we want is to be able to classify galaxies in the photometric sample in the same way they were classified in the training set, that is, we need to estimate the cross-correlation strength R statistic for them.

To be more realistic, instead of using R, we map the R values into a new quality parameter Q. The Q parameter is *discrete*, and roughly matches the more standard quality flags of real surveys (e.g. VVDS, DEEP2). It also has the advantage of having a more limited range than the R statistic, which has no upper limit. The mapping we use is as follows:

$$\begin{aligned} R \geq 6 &\iff Q = 4 \\ 5 < R < 6 &\iff Q = 3 \\ 4 < R < 5 &\iff Q = 2 \\ 3 < R < 4 &\iff Q = 1 \\ 0 < R < 3 &\iff Q = 0. \end{aligned}$$

The relationship between $\mathrm{SSR_T}$ and Q classification depends slightly on the exposure time, but can be inferred from the right-hand plot of Fig. 2.

Following standard neural network procedure, we split the spectroscopic sample into two parts (of equal size), the training and validation samples. As described in Appendix B, we use the *griz* magnitudes as the inputs for the neural network, which then outputs an estimate for Q. For simplicity, we only perform a single neural net run, though the average of multiple runs is expected to yield best results.

After the neural net run converges, we apply the best-fitting function to the complete spectroscopic sample and to the photometric sample to obtain estimates of the Q coefficient, hereafter $Q_{\rm est}$, for all the galaxies. Note that $Q_{\rm est}$ is continuous, whereas Q is discrete.

The matching between Q and $Q_{\rm est}$ is very good for the $Q = 4$ galaxies. For the 16 200 s exposures, the σ_{68} width of the $Q - Q_{\rm est}$ distribution is 0.05, and only 15 per cent of the galaxies with $Q = 4$ have $|Q - Q_{\rm est}| > 0.5$. The matching is less accurate for the galaxies for the lower Q, with the worst case being the $Q = 3$ sample, for which $\sigma_{68} = 0.88$. Interestingly, the redshift success statistics for the Q-selected sample are very similar to those of a $Q_{\rm est}$ selected sample. Most importantly, the SSR increases monotonically with $Q_{\rm est}$ (not shown), making it an accurate classifier of redshift confidence. We leave more detailed comparisons between Q and $Q_{\rm est}$ for future work.

We apply cuts on $Q_{\rm est} = 1.5$, 2.5, and 3.5 to both spectroscopic and photometric samples. With the 16 200 s exposures, the corresponding True SSR for the galaxy samples is 0.996, 0.978 and 0.914, respectively, with a corresponding fraction of objects relative to the total of 0.463, 0.586 and 0.751 in the three cases. For the 48 600 s exposures, we find True SSRs of 0.996, 0,978 and 0.936, respectively, with corresponding fractions of objects retained of 0.655, 0.808 and 0.960.

The next step is to investigate the impact of the selection to the WL analysis. We break up the process into several parts, for clarity.

(i) If a training set based method is to be used for calculating photo-zs, the first step is to use the training sample with the desired $Q_{\rm est}$ cut to derive photometric redshifts for the matched photometric sample (cf. Section 6.1.1). This step may be skipped if a pure template-based algorithm is being used.

(ii) Next, we calculate the WL constraints for the photometric sample selected with the $Q_{\rm est}$ cut and compare that to what we get for the full sample. Constraints degrade both from the reduction in the total number of objects as well as with the shift of the redshift distribution towards lower redshifts (cf. Section 6.1.2).

(iii) The next step is to assess the bias resulting from differences in the selection of the spectroscopic and photometric samples as well as the biases due to wrong redshifts. (cf. Section 6.1.2).

6.1.1 Photo-z training

We use a neural network photo-z estimator to exemplify the impact of selection matching and wrong redshifts on training set based photo-z estimation (cf. Section 4.2.1). For simplicity, we assume that the photo-zs for the photometric sample should only be calculated for the subset of galaxies surviving the selection cuts of the previous section. In other words, we require that the spectroscopic training sample and the photometric sample have matching selections. We thus define three sets of spectroscopic and photometric samples, specified by the spectroscopic quality cuts on $Q_{\rm est}$ of $Q_{\rm est} > 3.5$, 2.5, or 1.5.

To separate the effects of selection matching from the effect of wrong redshifts, we estimate the photo-zs twice. First, we assume we have the true redshifts for all galaxies passing the $Q_{\rm est}$ cuts, to isolate potential biases due to the spectroscopic selection matching. Then, we perform the photo-z training on the actual spectroscopic redshifts, to gauge the additional impact of wrong redshifts.

Table 1 shows the 1σ photo-z scatter for the samples defined by the $Q_{\rm est}$ cuts. The two $z_{\rm true}$ columns correspond to the scenarios where the true redshifts were used in the training. The scatter is defined as the dispersion in the distribution of $(z_{\rm true} - z_{\rm phot})$ for both the training sample and photometric sample. As expected, the photo-z scatter of the training sample is in excellent agreement with the scatter of the photometric sample, suggesting that both samples have close to identical photo-z properties and that the selection matching does not introduce any biases. Furthermore, the scatter

Table 1. Rms scatter of neural network photo-zs for the samples selected by the cuts on estimated $z_{\rm spec}$ quality, $Q_{\rm est} > 1.5$, 2.5, and 3.5. Note that the scatter for the Train*/$z_{\rm spec}$ column is defined as the dispersion in the $z_{\rm spec} - z_{\rm phot}$ distribution, whereas it is defined as the dispersion in the $z_{\rm true} - z_{\rm phot}$ for the other columns.

Photo-z scatter and training-set size					
	$z_{\rm true}$		$z_{\rm spec}$		
Selection	Train	Photo	Train	Photo	Train*
$Q_{\rm est} > 1.5$	0.121	0.121	0.149	0.149	0.214
$Q_{\rm est} > 2.5$	0.098	0.099	0.105	0.106	0.142
$Q_{\rm est} > 3.5$	0.082	0.083	0.081	0.082	0.098

improves as we apply more stringent cuts on $Q_{\rm est}$. The decrease in scatter is as expected, since the objects with low $Q_{\rm est}$ are typically the faintest.

The three $z_{\rm spec}$ columns in Table 1 show the more realistic case where the actual spectroscopic redshifts (wrong redshifts included) was used to train the photo-zs. In the $z_{\rm spec}$(Train) we show the scatter in the training set calculated as the dispersion in the $(z_{\rm true} - z_{\rm phot})$ distribution, which we can see is in excellent agreement with the scatter of the photometric sample shown in the fifth column. Comparing the dispersion of the $z_{\rm spec}$(Photo) and $z_{\rm true}$(Photo) cases, we see that the presence of wrong redshifts degrades the photo-zs of the photometric sample by as much as 20 per cent in the case of the $Q_{\rm est} > 1.5$ cut. The degradation is reduced for the more stringent cuts as the fraction of wrong redshifts is reduced.

In reality, one does not know the true redshifts for the training set, but only the spectroscopic redshifts. Hence, the scatter in the training-set photo-zs would be estimated using the spectroscopic redshifts, as the dispersion in the $(z_{\rm spec} - z_{\rm phot})$ distribution. We show this estimate of the scatter in the $z_{\rm spec}$(Train*) column. We see substantially larger values of the scatter compared to the $z_{\rm spec}$(Photo) column, for all $Q_{\rm est}$ cuts. The point is that the neural network is not substantially affected by the wrong redshifts, so that the true scatter does not degrade. However, our estimate of the scatter using the spectroscopic redshifts is strongly affected, as the wrong spectroscopic redshifts show up as catastrophically incorrect redshifts, which we can often remove.

We therefore conclude that the use of training samples is still justified in the presence of incorrect redshifts. For the sake of comparison, the template-fitting photo-zs without any priors have rms scatter of 0.229, 0.215 and 0.203 for the samples selected with $Q_{\rm est} > 1.5$, 2.5 and 3.5, respectively. If the scatter were estimated using the spectroscopic redshifts, then the rms would be 0.313, 0,246 and 0.211 for the same cases. A more careful analysis of the template-fitting photo-zs could certainly produce better results. However, as we show in the next section, the cosmological biases are very similar whether template or neural net photo-zs are used, and we leave more detailed analyses for future work. For more comparisons between these two methods using a catalogue with similar photometry, see Cunha et al. (2012).

6.1.2 *WL constraints and biases*

In this section, we examine the constraints and biases in the dark energy equation of state w inferred from WL shear–shear correlations. The errors in w are caused by our inability to characterize the photometric redshift error distribution of our sample. In other words, we must know the $P(z_{\rm true}|z_p)$ error matrix for our photometric sample to high accuracy. When we rely on a spectroscopic sample to characterize the error distribution, we are actually estimating $P(z_s|z_p)$, but this distribution differs from the true error matrix $P(z_{\rm true}|z_p)$ because of issues in spectroscopic selection matching and wrong spectroscopic redshifts. We now investigate how these spectroscopic redshift errors affect the dark energy equation of state measurements.

Table 2 shows the 1σ constraints on w and systematic errors for several different sample selections. The results shown used template-fitting photo-zs described in Section 4.2.2. For clarity, we artificially separate the issues due to selection matching from that of the wrong redshifts as follows: we perform the cosmological parameter forecast analysis assuming that all redshifts that passed the $Q_{\rm est}$ selection cut were the correct, true redshifts, thereby explicitly isolating the selection matching systematics. The results are presented under the $z_{\rm true}$ column in Table 2. We can see that biases in w are negligible compared to the statistical constraints, demonstrating that the neural network can accurately match the spectroscopic selection to the photometric sample. The table also shows the fraction of galaxies surviving the selection cuts. For example, for the 16 200 s exposures, we see that the $Q_{\rm est} > 3.5$ cut removes more than half of the sample, which results in nearly a factor of 2 degradation in the statistical constraints relative to what is achievable with the full sample ($\sigma(w) = 0.055$). The degradation is so severe because most of the objects removed by the cut are at high redshifts.

Table 2. Statistical and systematic errors in the dark energy equation of state w for the different $Q_{\rm est}$-selected samples. The bias results shown used the template-fitting photo-zs. The Gal. Frac. column indicates the fraction of galaxies from the full data set that passed the selection cut, and the $\rm SSR_T$ indicates the fraction of correct redshifts (i.e. fraction for which $|z_{\rm spec} - z_{\rm true}| < 0.01$) in the sample. The true redshifts $z_{\rm true}$ column assumes, artificially, that all galaxies in the spectroscopic sample that passed the $Q_{\rm est}$ cut had perfect spectroscopic redshifts. The $z_{\rm spec}$ column shows the more realistic case where the actual spectroscopic redshifts (including the small fraction of wrong redshifts) were used in the calibration of the photo-z error distributions. Recall that the statistical, marginalized, error in w for perfect redshifts is $\sigma(w) = 0.055$.

Constraints on w (template-fitting photo-zs)					
16 200 s				bias(w)	
Selection	Gal. frac.	$\rm SSR_T$ (per cent)	$\sigma(w)$	$z_{\rm true}$	$z_{\rm spec}$
$Q_{\rm est} > 1.5$	0.75	91.4	0.07	0.004	−0.52
$Q_{\rm est} > 2.5$	0.59	97.8	0.09	0.002	−0.13
$Q_{\rm est} > 3.5$	0.46	99.6	0.10	−0.001	−0.02
48 600 s					
$Q_{\rm est} > 1.5$	0.96	93.6	0.06	0.004	−0.39
$Q_{\rm est} > 2.5$	0.81	97.8	0.07	0.005	−0.15
$Q_{\rm est} > 3.5$	0.66	99.6	0.08	0.003	−0.03

Next, we examine the impact of wrong redshifts. As the last column of Table 2 shows, wrong redshifts can be devastating to the WL constraints. The bias in w is, perhaps, tolerable only in the $Q_{\rm est} > 3.5$ cases. In the other scenarios one can see that the biases in w are greater than the 1σ constraints even with close to 98 per cent correct redshifts ($\rm SSR_T \simeq 0.98$).

Comparing the 48 600 and 16 200 s results we see that the magnitude of the biases in w are set entirely by the $\rm SSR_T$, regardless of the level of completeness. This is another reminder that the emphasis must be on accuracy over completeness.

We investigated the dependence of the results on the photo-z estimator by performing the WL analysis with the neural network photo-zs instead of the template photo-zs. The resulting biases in w are shown in the third column of Table 3. Comparing to the fourth column, where we reproduce the template photo-z biases from Table 2, we see that the magnitude of the bias is very similar for the two photo-z estimators, despite noticeable differences in the photo-z error distributions of both (see e.g. Cunha et al. 2012).

We also tested the possibility of decreasing the biases by culling photo-z outliers. In the presence of wrong spectroscopic redshifts, the culling could remove not only catastrophic photometric redshifts, but perhaps also identify the wrong $z_{\rm spec}$s. We used the nearest-neighbour error estimator (NNE; Oyaizu et al. 2008a), to cull 10 per cent of the sample selected as the galaxies with largest NNE error, ($e_{\rm NNE}$). Since the fraction of objects to be culled was fixed, the value of the $e_{\rm NNE}$ cut varied for each catalogue and photo-z

Table 3. Biases in the dark energy equation of state w for both the training-set and template-fitting photo-z estimates when the NNE estimator is used to cull outliers in $|z_{\rm phot} - z_{\rm spec}|$ space. The last column assumes that the template-fitting photo-zs were culled based on the template-fitting error estimates. The 'G. Frac.' column indicates the fraction of galaxies from the full data set that passed the selection cut. Recall that the statistical marginalized errors in w for the three $Q_{\rm est}$ cases are 0.07, 0.09 and 0.10, respectively, as shown in Table 2.

	Biases in w (culling catastrophics with NNE)					
16 200 s		No NNE cut		NNE cut		T. cut
Selection	G. frac.	Neural	Templ.	Neural	Templ.	Templ.
$Q_{\rm est} > 1.5$	0.75	−0.27	−0.52	−0.19	−0.35	−0.41
$Q_{\rm est} > 2.5$	0.59	−0.13	−0.13	−0.06	−0.11	−0.09
$Q_{\rm est} > 3.5$	0.46	−0.02	−0.02	−0.01	−0.02	−0.01

estimator. The results are presented in the last two columns of Table 3. For simplicity, we did not recalculate the fiducial constraints when deriving the biases for the culled samples; given the qualitative nature of this analysis, this is a reasonable approximation. The NNE cut seems quite effective for the neural network photo-zs, typically reducing the biases by half. When the NNE culling was applied to the template-fitting estimator, the effect was negligible for the $Q_{\rm est} > 3.5$ case, and relatively small for the other cases, suggesting that the NNE is only effective for identifying spectroscopic outliers when a training set based procedure is used. This is by no means obvious since the NNE is very efficient at identifying photo-z outliers even when template-fitting methods are used (Oyaizu et al. 2008a). For comparison, we also tested the effect of applying the same 10 per cent cut using an error estimator from the template-fitting code itself[3] – shown in the last column of Table 3. We find that the biases due to wrong redshifts for the $Q_{\rm est} > 1.5$, 2.5 and 3.5 cases are reduced to −0.41, −0.086 and −0.014, showing that culling using this error estimator is also beneficial. We emphasize that the improvement from culling outliers is *not* due to the improvement of the photo-z statistics, but only due to the removal of spectroscopic failures with the culling. The effect on the neural net photo-zs was more significant simply because a larger fraction of failures was removed. In contrast, note that, in Cunha et al. (2012), we found that culling based on photo-z error estimates had little impact on cosmological biases due to sample variance in calibration sample, despite the effective identification of the photo-z outliers.

What about removing the wrong redshifts by comparing the photo-zs to spectroscopic redshifts? This is an intriguing option, but one must be very careful. The outlier islands of the photo-z error distribution often cause the largest biases in cosmological parameters (BH10; Hearin et al. 2010). By removing catastrophic objects from the spectroscopic sample, one all but ensures that catastrophic islands will not be calibrated properly, thereby compromising the cosmological constraints. Hence, even though the comparison to photo-zs can produce a cleaner spectroscopic sample, its selection will no longer match that of the photometric sample. There are two ways to salvage the situation. Either one can apply the neural network procedure once again to match the selection, or one can directly remove the regions of observable space occupied by the objects with the corresponding photo-zs. The latter approach is essentially what we have done by performing the NNE cuts described in the previous paragraph. Regions of observable space for which the difference between $z_{\rm spec}$ and $z_{\rm phot}$ are large were removed from both the spectroscopic and photometric samples, thereby removing the spectroscopic outliers and simultaneously matching the corresponding selection in the photometric sample.

Finally, we investigated the dependence of the results on the details of our spectroscopic pipeline, described in Section 4.1. We find that our Fiducial pipeline, despite giving the best high-redshift completeness, yielded the largest biases in w, shown in the Table 2. The different pipelines yielded consistent trends, and we focus on one particular case, that highlights the importance of the settings. The Original pipeline had a factor of 2 smaller bias for the $Q_{\rm est} > 3.5$ sample. In the Original setting, recall that only six templates were used. As can be seen by comparing the right-hand plot in Fig. 5 with Fig. 6, the three additional templates increased the redshift completeness above $z > 1.4$ but resulted in leakage from the high $z_{\rm true}$ bins to low $z_{\rm spec}$ bins. In particular, some galaxies at $z_{\rm true} \sim 1.9$ were assigned $z_{\rm spec}$s of ∼0.5 and ∼0.7. This failure mode was responsible for about 2/3 of the increase in bias in going from the Original to the Fiducial pipeline. The remainder of the difference was due to the fact that the Fiducial pipeline uses `czguess=1.6` which has the effect of increasing the probability that a galaxy will be assigned a high redshift. As a result, the Fiducial pipeline yields $z_{\rm spec}$s above 1.5 for several galaxies with $z_{\rm true} < 0.8$.

We conclude that the commonly adopted approach of maximizing the completeness is not recommended because it leads to the increase of the fraction of wrong redshifts which in turn implies worse dark energy parameter biases.

6.2 Weights approach to selection matching

In Section 6.1, we matched the selection of the spectroscopic and photometric samples by culling the photometric sample. That is, we selectively removed galaxies from the photometric sample so that it statistically matched, as closely as possible, the spectroscopic sample. In this section, we try a more aggressive approach that allows us to keep nearly the full photometric sample. Our technique is to weight galaxies in the spectroscopic sample using the `probwts` method of Lima et al. (2008) and Cunha et al. (2009), so that the statistical properties of these weighted spectroscopic galaxies match those of the photometric sample. For convenience of reference, we briefly describe the `probwts` technique in Appendix D.

We select a training set by picking galaxies from the spectroscopic sample with Q above some threshold $Q_{\rm crit}$. We test the reconstruction for several values of $Q_{\rm crit}$. Following standard `probwts` procedure, we remove the (small) part of the photometric sample that is determined to have zero overlap with the spectroscopic sample. This removes at most a few per cent of the photometric sample, with negligible impact on the statistical constraints.

Note that, in the first approach, with the neural net, all the spectroscopic sample is used to characterize the spectroscopic selection in observable space. The cosmological analysis is then only performed on the sample that matches the estimated selection. In the second, we only use reliable spectra, which we re-weight to match the full photometric sample. Then, the full photometric sample is used on the cosmological analysis. The first approach is the more conservative one as it throws away photometric data, to keep only the most reliable sample. The second approach is more aggressive as it tries to keep most of the data and only rescale the training set.

As the top plot of Fig. 7 shows, the weights improve the estimate of the overall redshift distribution when true redshifts are used. One can see that the weights roughly fix the broadest discrepancies, but

[3] The error estimate we use is the difference between the `Z_BEST68_HIGH` and `Z_BEST68_LOW` outputs of the *LePhare* code.

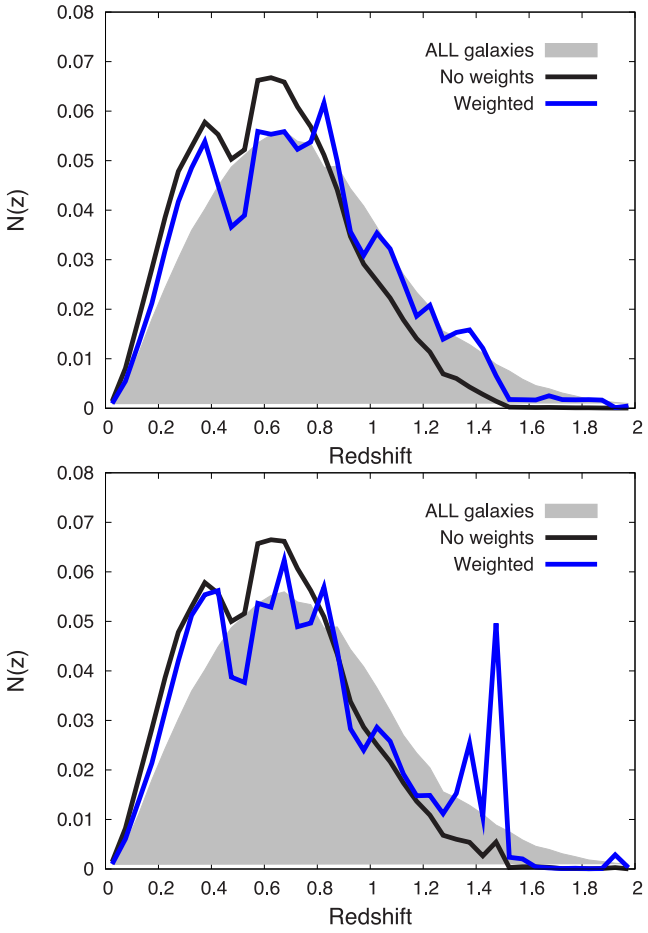

Figure 7. Top plot: the true-redshift distribution of the full photometric sample (shaded grey), of the spectroscopic sample with $Q \geq 3$ with no weights (black line), and with weights (blue–dark grey line). Bottom plot: same as above, but showing weighted and unweighted distributions of spectroscopic redshifts. One can see that, because wrong redshifts occupy regions of low completeness in observable space, the weights boost their impact enormously.

cannot correct sharper features. For example, the dip in the training sample from around $0.4 < z < 0.6$ gets rescaled, but its rough shape persists. What this suggests is that objects in this redshift range occupied the same region of observable space, and the weighting affected them all similarly.

The bottom plot shows the results when the spectroscopic redshifts are used. We see that even a speck of wrong redshifts (2.4 per cent in this case) can have dramatic impact depending on where they are located (cf. bottom plot). Comparing, the bottom plot of Fig. 7 with the middle plot of Fig. 5, we see that the spikes in the Weighted estimated of the redshift distribution at $z \sim 1.5$, 1.4, 0.8, 0.7 and 0.4 all correspond to the regions of concentration of wrong redshifts seen in Fig. 5. However, whereas the spikes below $z = 1$ are not particularly prominent, the spikes around $z = 1.4$ and 1.5 are enormous. There are a couple of factors contributing to the problem. As can be inferred from the left-hand plot in Fig. 2, the completeness drops precipitously above $z > 1.4$. Hence, the few spectroscopic redshifts above $z > 1.4$ typically receive large weights to compensate for the incompleteness. In addition, as shown in the middle plot of Fig. 5, the fraction of correct redshifts for galaxies with $z_{\rm true} > 1.4$ is very small, and many of these are incorrectly assigned a spectroscopic redshift of $z_{\rm spec} = 1.4$ or 1.5. The large weights magnify the impact of the wrong redshifts, resulting in the large spikes, and in large bias in the cosmological parameters, as we show in the next section.

6.2.1 *Weak lensing constraints and biases with weights*

Table 4 shows the 1σ constraints and biases on w when one uses the weights technique to match the spectroscopic selection to the photometric sample. As in Section 6.1.2, we separate the analysis into two parts. First, in the $z_{\rm true}$ column, we show only the effect of matching the selection between the spectroscopic and photometric samples. Afterwards, in the $z_{\rm spec}$ column, we use the actual spectroscopic redshifts to show the impact of wrong redshifts.

As shown in the table the weights perform reasonably for all cases when one considers only the true redshifts. The biases are typically smaller than the statistical errors on w, and the statistical constraints are better than for the culling approach of Section 6.1.2 since almost all of the photometric sample was usable for analysis. It is interesting to note that more rigorous cuts ($Q > 6$ and 5) yielded the smallest biases even though the completeness of the spectroscopic sample was smaller than for the $Q < 4$ case. It is instructive to re-examine the top plot of Fig. 7. Comparing the blue solid line with the shaded grey region, we see that the weights reconstruction of the overall redshift distribution is quite jagged, yet it yields small biases in w. The reason for this is that the w is sensitive to a very broad redshift range, and an estimator can over- or underestimate the overall redshift distribution in small redshift ranges as long as there are no net biases. This point is explored in much more detail in section 5.3.1 of Cunha et al. (2012). Unfortunately, the $z_{\rm spec}$ column in Table 4 shows that the presence of wrong redshifts severely compromises the weights approach.

Because the wrong redshifts are tightly associated with the regions of high incompleteness, particularly at high redshift, and because the variations in completeness are so sharp, the wrong redshifts received very large weights resulting in large cosmological biases. A major part of the problem is the sharp *change* in completeness with redshift shown on the left-hand plot of Fig. 2. We find that the results for the weights do not improve for the 48 600 s cases because the steep variations in the completeness with redshift become even larger for that case since the increased exposure time did not yield significant increase in completeness above z of 1.4.

Table 4. Statistical and systematical errors in w when the weights technique for selection matching is used. Results are shown assuming that the spectroscopic sample was selected with different cuts of the cross-correlation strength parameter Q, described in Section 6.1. The bias results shown used the template-fitting photo-zs. The Gal. Frac. column indicates the fraction of galaxies from the *spectroscopic sample* that passed the selection cut, and the $\rm SSR_T$ indicates the fraction of correct redshifts (i.e. fraction for which $|z_{\rm spec} - z_{\rm true}| < 0.01$) in the sample. Essentially, all of the photometric sample was used in the analysis, hence the statistical constraints are the same for all samples.

Constraints on w (template-fitting photo-zs and weights)					
16 200 s				bias(w)	
Selection	G. frac.	$\rm SSR_T$ (per cent)	$\sigma(w)$	$z_{\rm true}$	$z_{\rm spec}$
$Q \geq 3$	0.73	93.2	0.06	0.070	−0.7
$Q \geq 4$	0.60	98.6	0.06	0.034	−0.5
$Q \geq 5$	0.53	99.6	0.06	−0.036	−0.3

In summary, we find that the weights approach needs to be considered with care in the presence of wrong redshifts, and that the more conservative approach of culling using the neural network is the safest. In practice, the weights are often needed to account for other types of incompleteness (see e.g. Cunha et al. 2009), so both approaches should be used in tandem.

7 DISCUSSION: ROBUSTNESS OF ASSUMPTIONS AND RESULTS

We now discuss the dependence of our results on the key assumptions and numerical tools used in this work.

(i) *N-body/photometric simulations*. The success rate statistics are affected by luminosity function and distribution of galaxy types in the simulation. However, the main conclusions of our paper, concerning selection matching and impact of wrong redshifts, should not be affected. We tested the selection matching for a variety of situations (several of which we do not show), including varying atmospheric noise models and spectrograph resolution. For all cases, the matching worked well, incurring no additional bias. In addition, Soumagnac et al. (in preparation) obtain similar results using a very different set of spectro/photometric simulations described in Jouvel et al. (2009).

The distribution of wrong redshifts in ($z_{\rm true}$, $z_{\rm spec}$) space could also change for a different simulation, but the preferred loci where the failures concentrate should not vary appreciably, since they are based on confusion between galaxy or atmospheric spectral lines that do not depend on any details of the simulation. Furthermore, the fact that a small fraction of spectroscopic failures can cause severe biases is not likely to change.

(ii) *Sky noise model*. Our model for sky subtraction is idealized as it assumes a perfect shot noise model. Sky subtraction is often not as efficient, and observing conditions vary from the median. In addition, there are issues such as CCD fringing (cf. Section 2.3) which are difficult to model. Other effects we did not model include contamination from nearby stars or bright galaxies, and cosmic rays. These other effects, however, are only expected to affect the overall completeness, without galaxy type or redshift dependence.

(iii) *Simulated spectra*. As discussed in Appendix A2, the simulated spectra we use are based on the five eigenspectra of `kcorrect`, which are derived based on about 1600 SDSS main sample galaxies, 400 luminous red galaxies (LRGs) and a photometric sample of several thousands of galaxies imaged in the UV, optical and IR. Is this enough? Yip et al. (2004) showed that a set of three eigentemplates were sufficient to describe about 98 per cent of the variance in the 170 000 galaxies in the Strauss et al. (2002) SDSS sample. Additional templates improved coverage very slowly, with a set of 500 eigentemplates needed to account for 99 per cent of the sample variance (cf. table 1 in that work). Yip et al. (2004) show that the missing variance was due mainly to extreme line emission galaxies. We roughly confirm this trend for our simulated spectra by looking at the distribution of equivalent widths of the [O II] emission line for our simulated galaxies. We find that our equivalent widths reach at most 30 Å. For comparison, Cooper et al. (2006) find, for the DEEP2 sample, a distribution of [O II] equivalent widths reaching as much as 100 Å.

In addition, Yip et al. (2004) showed that one needs a random subsample of about 10 000 galaxies to obtain convergence for the first 10 eigentemplates. These results suggest that the `kcorrect` basis should be sufficient to characterize all but a few per cent of the low-redshift galaxies.[4] However, a few per cent of 'oddball' galaxies could potentially cause problems for cosmological analysis if they cannot be disentangled from the rest of the sample using colours and if their redshift distribution differs significantly from the rest of the sample with similar colours. The problem is expected to become worse at high redshift. To properly quantify the impact of the outliers, observing campaigns targeted at the spectroscopic failures of existing spectroscopic surveys are crucial.

In some sense, our choice of template library used for deriving spectroscopic redshifts is pessimistic for the high-redshift galaxies: as discussed in Appendix A2, the `kcorrect` templates are based on *GALEX* colours for the bluer frequencies. Hence, parts of the spectra of high-z objects were simulated using purely photometric data, resulting in excessively featureless spectra in the UV frequencies, which implied lower-than-expected completeness for $z > 1.4$.

(iv) *Spectroscopic redshift pipeline*. The `rvsao.xcsao` code uses cross-correlation techniques in Fourier space to derive redshifts from spectra. The disadvantage of this approach relative to a standard χ^2 method is that one does not include any information about the noise. One can disregard certain regions of the spectrum in the analysis, thereby removing at least the most prominent atmospheric lines. We found that the removal of some lines did not increase the completeness of the sample noticeably, and changed the distribution of the wrong redshifts. We leave more extensive tests on the optimal techniques for spectroscopic redshift estimation for a future work.

(v) *Culling approach to selection matching*. In a real spectroscopic survey, the success of the neural network culling approach to selection matching is likely to require a more careful redshift confidence classification. In our simulation, the atmospheric conditions were fixed for the full sample, and the noise subtraction was perfect. In a real survey, the variability of observing conditions and other problems such as slit placement and cosmic ray contamination will complicate the mapping between magnitudes in the photometric surveys and the redshift confidence in the spectroscopic survey. This difficulty can likely be overcome if some measure of the observing conditions is used as input when training the neural network. We leave these investigations for a future work.

8 IMPLICATIONS FOR SURVEY DESIGN

Given the findings of this paper and Cunha et al. (2012), what should survey planners do to optimize their spectroscopic surveys?

The first step is obvious: one needs to optimize the allocation of time observing different kinds of galaxies. Specifically, one can use colour information to preselect galaxies that will require longer exposure times to obtain accurate redshifts. For example, in Section 6.1.2, we saw that tripling the exposure time improved the completeness from 0.46 to 0.66 for the $Q_{\rm est} > 3.5$ cut. If the 20 per cent of the sample that yielded additional redshifts could be known in advance, one would only target this sample for additional observation, which would only require an increase of 40 per cent in the observing time, instead of the naive 200 per cent additional time if the full sample was targeted for follow-up observation. With an optimized observing strategy, one would be able to save precious

[4] The Yip et al. (2004) analysis was based on principal component analysis, whereas Blanton & Roweis (2007) used non-negative matrix factorization to determine their respective eigenbasis. Thus, comparisons between Blanton & Roweis (2007) and Yip et al. (2004) are only meant as ballpark estimates.

telescope time and still achieve redshift accuracy that does not degrade the cosmological constraints appreciably. We leave a more detailed analysis for future work.

We showed in this paper that the tolerance for wrong redshifts is extremely low. It is, however, possible to get away with a higher fraction of wrong spectroscopic redshifts by modelling their effects on the cosmological parameters. Then one would need to, in analogy to the photo-z case, fully characterize the spectroscopic error matrix $P(z_{\rm spec}|z_{\rm true})$. However, determining the matrix $P(z_{\rm spec}|z_{\rm true})$ from observations is likely to be very challenging in practice, as in order to control the sample variance of galaxies used for the calibration, one would likely have excessively high requirements on the area of the follow-up (Cunha et al. 2012).

It is also possible that one can use spatial cross-correlations to estimate the spectroscopic error matrix. Since correlations between different redshift bins should be very close to zero, any correlation has to be due to wrong redshifts. Several works have explored this fact for photo-z calibration (Schneider et al. 2006; Erben et al. 2009; Benjamin et al. 2010; Zhang, Pen & Bernstein 2010). Schneider et al. (2006), for example, found cross-correlations to work well only in the simplest Gaussian cases. But for spectroscopic failures, the excess correlation signal should be due to a few big outliers, and hence might be more easily detectable.

9 CONCLUSIONS

We investigated the impact of spectroscopic failures on the training and calibration of photometric redshifts, and the consequent impact on the forecasted dark energy parameter constraints from weak gravitational lensing. Our tests were based on N-body/spectrophotometric simulations patterned after the DES and expected spectroscopic follow-up observations loosely patterned after the VVDS survey.

Spectroscopic failures consist of two types of issues: the inability to obtain spectroscopic redshifts for certain galaxies, and incorrect redshifts.

The inability to obtain redshifts introduces incompleteness in the spectroscopic sample – i.e. missing redshifts in some region of parameter space (e.g. at faint magnitudes) represented in the full photometric population of galaxies. This incompleteness must be accounted for before one can use the spectroscopic sample to calibrate photo-zs – i.e. characterize the photo-z error matrices, e.g. the $P(z_s|z_p)$, of the sample.

We studied two approaches to account for the incompleteness in the spectroscopic sample. In the first approach, we used an ANN to estimate the spectroscopic selection function for the photometric sample. This selection function was then used to cull the photometric sample so that its statistical properties matched the spectroscopic sample. We found that this approach works extremely well, yielding only insignificant bias in the WL constraints using the culled sample (refer to $z_{\rm true}$ column in Table 2). However, the statistical constraints did degrade substantially as, typically, a large fraction of the sample was culled. In the second approach, we accounted for the incompleteness in the spectroscopic sample by applying weights to the galaxies with spectroscopic redshifts, following the approach of Lima et al. (2008), so that the statistical properties of the spectroscopic and photometric samples match. This approach was also successful (cf. $z_{\rm true}$ column in Table 4) – as expected, because most of the photometric sample could be used – yielding tolerable cosmological biases while obtaining the maximum statistical constraints. Overall, we found that the effects of spectroscopic incompleteness are well under control.

Unfortunately, on the other hand, we found that wrong redshifts can significantly degrade cosmological constraints and >99 per cent of correct spectroscopic redshifts seems to be needed (cf. $\rm SSR_T$ and $z_{\rm spec}$ columns in Tables 2 and 4). We found the results to be independent of the photo-z estimators used, but somewhat dependent on the settings of the spectroscopic pipeline. In particular, we found that attempts to increase the completeness of the spectroscopic sample during the spectral analysis can result in more catastrophic spectroscopic redshift failures, which will increase cosmological biases.

We tested a couple of approaches to identify wrong spectroscopic redshifts, finding that the NNE (Oyaizu et al. 2008a) is able to reduce the bias in the measured dark energy equation of state by half while removing only 10 per cent of the photometric sample. Slightly less improvement in the w bias was obtained using the template-fitting error estimator.

In summary, we find that wrong redshifts are by far the main issue affecting calibration of photo-z error distributions with spectroscopic samples. Future follow-up spectroscopic observations of the planned and ongoing wide-area photometric surveys must focus primarily on the accuracy of the spectroscopic redshifts even if that implies sacrificing the spectroscopic completeness.

ACKNOWLEDGEMENTS

CEC would like to thank Joerg Dietrich, Stephanie Jouvel, Anja von der Linden, Jeff Newman and Peter Norberg for discussions about spectroscopic surveys. We also thank G. Bernstein, J. Cohn, and S. Lilly for detailed comments on the draft. This paper has gone through internal review by the DES collaboration. CEC is supported by a Kavli Fellowship at Stanford University. DH is supported by the DOE OJI grant under contract DE-FG02-95ER40899. DH is additionally supported by NSF under contract AST-0807564, and NASA under contract NNX09AC89G. RHW received support from the US Department of Energy under contract number DE-AC02-76SF00515. MTB was supported by Stanford University and the Swiss National Science Foundation under contract 2000 124835/1. This research was supported in part by the National Science Foundation under Grant no. PHY05-51164, Grant no. 1066293 and the hospitality of the Aspen Center for Physics. Fermilab is operated by Fermi Research Alliance, LLC under Contract no. DE-AC02-07CH11359 with the US Department of Energy.

REFERENCES

Abdalla F. B., Amara A., Capak P., Cypriano E. S., Lahav O., Rhodes J., 2008, MNRAS, 387, 969
Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, MNRAS, 417, 1891
Abrahamse A., Knox L., Schmidt S., Thorman P., Tyson J. A., Zhan H., 2011, ApJ, 734, 36
Amara A., Refregier A., 2007, MNRAS, 381, 1018
Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Benjamin J., Van Waerbeke L., Menard B., Kilbinger M., 2010, MNRAS, 408, 1168
Bernstein G., Huterer D., 2010, MNRAS, 401, 1399 (BH10)
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Blanton M. R. et al., 2003, ApJ, 592, 819
Bordoloi R., Lilly S. J., Amara A., 2010, MNRAS, 406, 881
Bordoloi R. et al., 2012, MNRAS, 421, 1671
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Busha M. T., Wechsler R. H., Behroozi P. S., Gerke B. F., Klypin A. A., Primack J. R., 2011, ApJ, 743, 117

Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, AJ, 132, 926
Coleman G. D., Wu C. C., Weedman D. W., 1980, ApJS, 43, 393
Collister A. A., Lahav O., 2004, PASP, 116, 345
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Cooper M. C. et al., 2006, MNRAS, 370, 198
Cooper M. C., Tremonti C. A., Newman J. A., Zabludoff A. I., 2008, MNRAS, 390, 245
Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, MNRAS, 396, 2379
Cunha C. E., Huterer D., Busha M. T., Wechsler R. H., 2012, MNRAS, 423, 909
Eisenstein D. J. et al., 2001, AJ, 122, 2267
Erben T. et al., 2009, A&A, 493, 1197
Feldmann R. et al., 2006, MNRAS, 372, 565
Hearin A. P., Zentner A. R., Ma Z., Huterer D., 2010, ApJ, 720, 1351
Hildebrandt H. et al., 2010, A&A, 523, A31
Huterer D., Linder E. V., 2007, Phys. Rev. D, 75, 023519
Huterer D., Turner M. S., 2001, Phys. Rev. D, 64, 123527
Huterer D., Takada M., Bernstein G., Jain B., 2006, MNRAS, 366, 101
Ilbert O. et al., 2006, A&A, 457, 841
Jouvel S. et al., 2009, A&A, 504, 359
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001, ApJ, 556, 121
Kinney A. L., Calzetti D., Bohlin R. C., McQuade K., Storchi-Bergmann T., Schmitt H. R., 1996, ApJ, 467, 38
Kitching T. D., Taylor A. N., Heavens A. F., 2008, MNRAS, 389, 173
Knox L., Scoccimarro R., Dodelson S., 1998, Phys. Rev. Lett., 81, 2004
Kurtz M. J., Mink D. J., 1998, PASP, 110, 934
Le Fèvre O. et al., 2005, A&A, 439, 845
Lilly S. J. et al., 2007, ApJS, 172, 70
Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118
Loveday J. et al., 2012, MNRAS, 420, 1239
Ma Z., Bernstein G., 2008, ApJ, 682, 39
Ma Z., Hu W., Huterer D., 2006, ApJ, 636, 21
Nakajima R., Mandelbaum R., Seljak U., Cohn J. D., Reyes R., Cool R., 2012, MNRAS, 420, 3240
Newman J. A. et al., 2012, ApJS, 208, 5
Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008a, ApJ, 689, 709
Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., Sheldon E. S., 2008b, ApJ, 674, 768
Schneider M., Knox L., Zhan H., Connolly A., 2006, ApJ, 651, 14
Strauss M. A. et al., 2002, AJ, 124, 1810
Tonry J., Davis M., 1979, AJ, 84, 1511
Wetzel A. R., White M., 2010, MNRAS, 403, 1072
Yip C. W. et al., 2004, AJ, 128, 585
York D. G. et al., 2000, AJ, 120, 1579
Zhang P., Pen U. L., Bernstein G., 2010, MNRAS, 405, 359

APPENDIX A: THE SIMULATIONS

In this section, we describe the construction of the simulations used in our analysis.

A1 *N*-body/photometric simulations

The simulated galaxy catalogue used for this work was generated using the Adding Density Determined GAlaxies to Lightcone Simulations (ADDGALS) algorithm (Busha et al., in preparation; Wechsler et al., in preparation). This algorithm attaches synthetic galaxies to dark matter particles in a lightcone output from a dark matter *N*-body simulation. The model is designed to match the luminosities, colours, and clustering properties of galaxies.

The simulations used here start with a dark matter lightcone which spans the redshift range from $0 < z < 2$, over one octant of sky (5156 sq. degrees). The lightcone is constructed from three distinct *N*-body simulations, which range in resolution from a few 10^{10} to a few $10^{11}\,\mathrm{M}_{\odot}$ particles and box sizes ranging from 1 to 4 Gpc h^{-1}. The simulations were run with the LGADGET code and modelled a flat Λcold dark matter cosmology using parameters consistent with *Wilkinson Microwave Anisotropy Probe* 7 results.

The ADDGALS algorithm used to create the galaxy distribution consists of two steps: galaxies based on an input luminosity function are first assigned to particles in the simulated lightcone, after which multiband photometry is added to each galaxy using a training set of observed galaxies. For the first step, we begin by defining the relation $P(\delta_{\rm dm}|M_r, z)$ – the probability that a galaxy with magnitude M_r and a redshift z resides in a region with local density $\delta_{\rm dm}$, defined as the radius of a sphere containing $1.8 \times 10^{13} h^{-1}\,\mathrm{M}_{\odot}$ of dark matter. This relation can be tuned to reproduce the luminosity-dependent galaxy 2-point function by using a much higher resolution simulation combined with the technique known as subhalo abundance matching. This is an algorithm for populating very high resolution dark matter simulations with galaxies based on halo and subhalo properties that accurately reproduce properties of the observed galaxy clustering (Conroy, Wechsler & Kravtsov 2006; Behroozi, Conroy & Wechsler 2010; Wetzel & White 2010; Busha et al. 2011). The relationship $P(\delta_{\rm dm}|M_r, z)$ can be measured directly from the resulting catalogue. Once this probability relation has been defined, galaxies are added to the simulation by integrating a (redshift dependent) *r*-band luminosity function to generate a list of galaxies with magnitudes and redshifts, selecting a $\delta_{\rm dm}$ for each galaxy by drawing from the $P(\delta_{\rm dm}|M_r, z)$ distribution, and attaching it to a simulated dark matter particle with the appropriate $\delta_{\rm dm}$ and redshift. The advantage of ADDGALS over other commonly used approaches based on the dark matter haloes is the ability to produce significantly deeper catalogues using simulations of only modest size. When applied to the present simulation, we populate galaxies as dim as $M_r \approx -14$, compared with the $M_r \approx -21$ completeness limit for a standard halo occupation approach.

While the above algorithm accurately reproduces the distribution of satellite galaxies, central objects require explicit information about the mass of their host haloes. Thus, for haloes with more than 100 particles, we assign central galaxies using the explicit mass–luminosity relation determined from our calibration catalogue. We also measure $\delta_{\rm dm}$ for each halo, which is used to draw a galaxy from the integrated luminosity function with the appropriate magnitude and density to place at the centre.

For the galaxy assignment algorithm, we choose a luminosity function that is similar to the SDSS luminosity function as measured in Blanton et al. (2003), but evolves in such a way as to reproduce the higher redshift observations (e.g., SDSS-Stripe 82, AGES, GAMA, NDWFS and DEEP2). In particular, ϕ_* and M_* are varied as a function of redshift in accordance with the recent results from GAMA (Loveday et al. 2012).

Once the galaxy positions have been assigned, photometric properties are added. We begin with a training set of spectroscopic galaxies and the simulated set of galaxies with *r*-band magnitudes generated earlier. For each galaxy in both the training set and simulation we measure Δ_5, the distance to the fifth nearest galaxy on the sky in a redshift bin. Each simulated galaxy is then assigned an SED based on drawing a random training-set galaxy with the appropriate magnitude and local density, *k*-correcting to the appropriate redshift, and projecting on to the desired filters. When doing the colour assignment, the likelihood of assigning a red or a blue galaxy is smoothly varied as a function of redshift in order to

simultaneously reproduce the observed red fraction at low and high redshifts as observed in SDSS and DEEP2.

Differences between the training-set and simulated galaxy sample complicate the process of colour assignment. In order to compile a sufficiently large training set, we use a magnitude-limited sample of SDSS spectroscopic galaxies brighter than $m_r = 17.77$ with $z < 0.2$. The simulated sample, on the other hand, is a volume-limited sample, spanning a broader redshift range. When measuring Δ_5 we restrict ourselves to neighbours brighter than $M_r = -19.7$ in the simulation sample, while using all objects in the observational catalogue. To mitigate differences in luminosity and redshift, each galaxy is rank ordered according to its density in its redshift bin, and require that objects be in the same percentile bin in each sample rather than having the same the absolute value of Δ_5. This is similar to the method used in Cooper et al. (2008).

The final step for producing a realistic simulated catalogue is the application of photometric errors. While the photometric errors generated here are particular to DES, the algorithm can be generalized for any survey. For each galaxy, we add a noise term to the intrinsic galaxy flux, where the noise is drawn from a Gaussian of width

$$\text{noise} = \sqrt{t_e n_p n_s + f_{g,i} t_e}, \quad \text{(A1)}$$

where t_e is the exposure time, n_p the number of pixels covered by a galaxy, n_s the flux of the sky in a single detector pixel, and $f_{g,i}$ is the intrinsic flux of the galaxy. Here, galaxies are assumed to have the same angular size, hence n_p is identical for all objects. Application of the above relation to objects from the SDSS catalogue shows that it is able to faithfully reproduce the reported errors of the survey.

A2 Creating simulated spectra

We use the `kcorrect v4_1` code (Blanton et al. 2003) to derive simulated spectra. The `kcorrect` code includes a set of five eigenspectra derived using a non-negative matrix factorization (NMF) technique (Blanton & Roweis 2007). To derive the eigenspectra, the authors start out with a basis of 450 star formation history templates from Bruzual & Charlot (2003) as well as 35 templates from Kewley et al. (2001). The method uses this basis to derive the non-negative linear combination of templates that best described the observations. In this case, the observations consist of a sample of several thousand photometrically and/or spectroscopically observed galaxies, from the far-UV to the near-IR (Blanton & Roweis 2007). The spectroscopic part of the training data consisted of 400 SDSS LRGs with $0.15 < z < 0.5$ (Eisenstein et al. 2001) and 1600 SDSS main sample galaxies with $0.0001 < z < 0.4$ (Strauss et al. 2002), with both sets of data observed in the range $3800\,\text{Å} < \lambda < 9000\,\text{Å}$.

We use the `kcorrect` subroutine to convert the true redshift and error-free magnitudes of a simulated galaxy from our photometric simulation into a best-fitting SED. The SED is characterized by the coefficients of the five eigentemplates, and are output as the variable `coeffs`. The `coeffs` are then passed into the subroutine `k_reconstruct_spec`, which produces a simulated spectrum with a resolution, in units of velocity dispersion, of 300 km s^{-1}.

We pattern our mock survey loosely on the VVDS (VVDS; Le Fèvre et al. 2005). The characteristics of the instrument that we assume are: collecting area of $16\pi\text{m}^2$, aperture of $5 \times 0.5\,\text{arcsec}^2$. For simplicity, we assume a constant resolution and a dispersion of $\Delta\lambda = 7.14\,\text{pixel}^{-1}$ over the entire spectrograph range of 5500−9500 Å. Comparing the spectrograph window of 5500−9500 Å to the spectroscopic coverage of the training set used to create the simulated spectra, we see that for objects below redshift

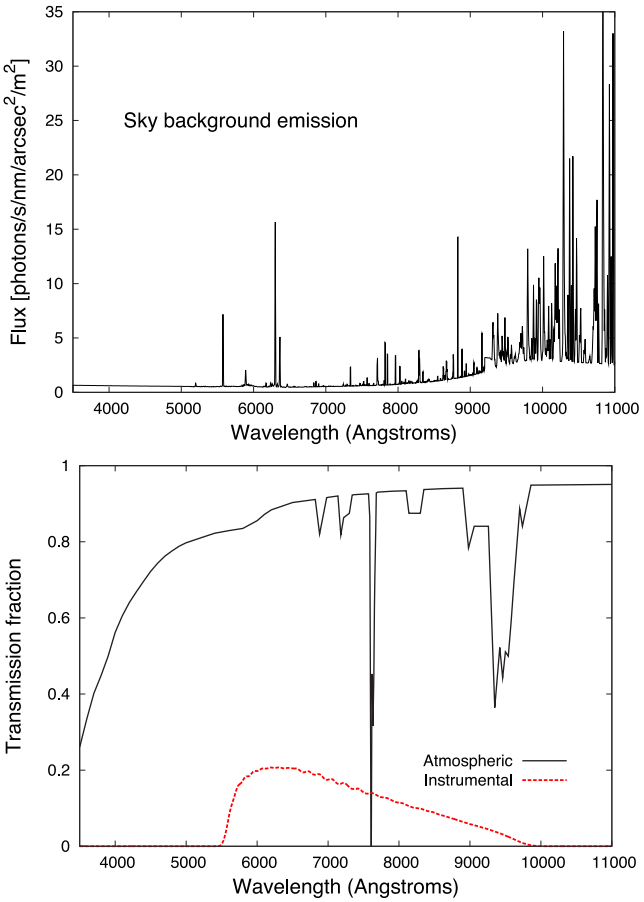

Figure A1. Top panel: atmospheric emission in units of photons s^{-1} nm^{-1} m^{-2} arcsec^{-2}. Bottom panel: atmospheric and instrumental transmission fractions, i.e. fraction of photons that reach the focal plane, used in our simulation. The total transmission function is given by the product of the two transmissions.

of 0.05, there is no spectroscopic representation of the training-set galaxies in the range 9000−9500 Å. More problematic is the fact that the spectroscopic training set has wavelength coverage starting at 3800 Å, and only goes to $z = 0.4$. As a result, for galaxies at about $z > 1.0$, the blue side of the simulated spectra are based solely on photometric data. Considering that most of the SDSS main sample is below redshift of 0.2, the simulated spectra should begin to lose resolution in the blue end for $z > 0.73$. These limitations in the simulated spectra result in higher-than-expected incompleteness above $z = 1.4$, but do not affect the overall conclusions.

We use a Palomar sky extinction model (courtesy of Oke and Gunn) with 1.3 airmasses and altitude of 2635 m to calculate the atmospheric transmission fraction (the solid black line in the bottom panel of Fig. A1). The instrument transmission is based on the VIMOS instrument transmission function[5] and is shown as the dashed red line in the bottom panel of Fig. A1. The total transmission is the product of the atmospheric and instrumental transmissions. We assume 16 200 s exposures for the fiducial observation strategy and also investigate a scenario with 48 600 s exposures.

[5] http://www.eso.org/observing/etc/bin/gen/form?INS.NAME=VIMOS+INS.MODE=SPECTRO

We add atmospheric emission based on the sky spectrum[6] shown in the top panel of Fig. A1. The total noise is given by the rms sum of the atmospheric noise, shot noise from the galaxy spectrum itself and readout noise per pixel, which we take to be a constant 5 photons. In reality, we only simulate the sky-subtracted spectrum, as follows. First, we convert the different spectra into photon counts for each pixel. We then assume that the atmospheric and galaxy noise follow a Poisson distribution, so that the uncertainty in the produced noise is the square root of the number of photons emitted. The readout noise is taken to be Gaussian. We calculate the total noise, N as

$$N = \sqrt{n_{\rm atm} + n_{\rm gal} + n_{\rm read}^2}, \tag{A2}$$

where $n_{\rm atm}$, $n_{\rm gal}$ and $n_{\rm read}$ are the number of photons from the atmosphere, the galaxy and the readout noise, respectively. The expected signal is simply the total number of photons from the galaxy. The expectation value of the error in the flux, δF is then given by

$$\delta F = F\frac{N}{S}. \tag{A3}$$

To obtain the sky-subtracted galaxy spectrum we, at each pixel, sample from a Gaussian distribution with mean given by the flux and width given by the error in the flux δF.

APPENDIX B: ARTIFICIAL NEURAL NETWORKS

We use an ANN method to both estimate the spectroscopic redshift quality and photometric redshifts, using an implementation based on Collister & Lahav (2004) and Oyaizu et al. (2008b). Despite the fancy name, an ANN is simply a function which relates redshifts (or any quantity we wish to estimate) to photometric observables. The training set is used to determine the best-fitting value for the free parameters of the ANN. The best-fitting parameters are found by minimizing the overall scatter of the photo-zs determined for the training-set galaxies. The ANN configurations are not unique in the sense that different sets of parameters can result in the same overall scatter. The best-fitting parameters found after minimizing the scatter depend on where in parameter space the optimization run begins. Hereafter, we refer to an ANN function using a given set of best-fitting parameters as a neural network solution.

The technical details are as follows. We use a particular type of ANN called a Feed Forward Multilayer Perceptron, which consists of several nodes arranged in layers through which signals propagate sequentially. The first layer, called the input layer, receives the input photometric observables (magnitudes, colours, etc.). The next layers, denoted hidden layers, propagate signals until the output layer, whose outputs are the desired quantities, in this case the photo-z estimate or the redshift quality Q estimate. Following the notation of Collister & Lahav (2004), we denote a network with k layers and N_i nodes in the ith layer as $N_1 : N_2 : \ldots : N_k$.

A given node can be specified by the layer it belongs to and the position it occupies in the layer. Consider a node in layer i and position α with $\alpha = 1, 2, \ldots, N_i$. This node, denoted $P_{i\alpha}$, receives a total input $I_{i\alpha}$ and fires an output $O_{i\alpha}$ given by

$$O_{i\alpha} = F(I_{i\alpha}), \tag{B1}$$

where $F(x)$ is the activation function. The photometric observables are the inputs $I_{1\alpha}$ to the first layer nodes, which produce outputs $O_{1\alpha}$. The outputs $O_{i\alpha}$ in layer i are propagated to nodes in the next layer $(i + 1)$, denoted $P_{(i+1)\beta}$, with $\beta = 1, 2, \ldots N_{i+1}$. The total input $I_{(i+1)\beta}$ is a weighted sum of the outputs $O_{i\alpha}$

$$I_{(i+1)\beta} = \sum_{\alpha=1}^{N_i} w_{i\alpha\beta} O_{i\alpha}, \tag{B2}$$

where $w_{i\alpha\beta}$ is the weight that connects nodes $P_{i\alpha}$ and $P_{(i+1)\beta}$. Iterating the process in layer $i + 1$, signals propagate from hidden layer to hidden layer until the output layer. In our implementation, we use a network configuration N_m: 10: 10: 10: 1, which receives N_m magnitudes and outputs a photo-z or a spectroscopic redshift quality. We use hyperbolic tangent activation functions in the hidden layers and a linear activation function for the output layer.

APPENDIX C: BIASES IN WEAK LENSING MEASUREMENTS OF DARK ENERGY

To assess the impact of the spectroscopic failures on the cosmological parameters, we closely follow the formalism used in our previous work on the impact of sample variance to photo-z calibration (Cunha et al. 2012). We consider a WL survey, and for simplicity only study the shear–shear correlations. The observable quantity we consider is the convergence power spectrum

$$C_{ij}^{\kappa}(\ell) = P_{ij}^{\kappa}(\ell) + \delta_{ij}\frac{\langle\gamma_{\rm int}^2\rangle}{\bar{n}_i}, \tag{C1}$$

where $\langle\gamma_{\rm int}^2\rangle^{1/2}$ is the rms intrinsic ellipticity in each component, $\bar{n}_i$ is the average number of galaxies in the ith redshift bin per steradian and ℓ is the multipole that corresponds to structures subtending the angle $\theta = 180^\circ/\ell$. For simplicity, we drop the superscripts κ below. We take $\langle\gamma_{\rm int}^2\rangle^{1/2} = 0.26$.

We follow the formalism of BH10, where the photometric redshift errors are algebraically propagated into the biases in the shear power spectra. These biases in the shear spectra can then be straightforwardly propagated into the biases in the cosmological parameters. We now review briefly this approach.

Let us assume a survey with the (true) distribution of source galaxies in redshift $n_{\rm t}(z)$, divided into B bins in redshift. Let us define the following terms.

(i) *Leakage* $P(z_p|z_t)$ (or l_{tp} in BH10 terminology): fraction of objects from a given true-redshift bin that are placed into an incorrect (non-corresponding) photometric bin.

(ii) *Contamination* $P(z_t|z_p)$ (or c_{tp} in BH10 terminology): fraction of galaxies in a given photometric bin that come from a non-corresponding true-redshift bin.

When specified for each tomographic bin, these two quantities contain the same information. Note in particular that the two quantities satisfy the integrability conditions

$$\int P(z_p|z_t){\rm d}z_p \equiv \sum_p l_{tp} = 1 \tag{C2}$$

$$\int P(z_t|z_p){\rm d}z_t \equiv \sum_t c_{tp} = 1. \tag{C3}$$

A fraction l_{tp} of galaxies in some true-redshift bin n_t 'leak' into some photo-z bin n_p, so that l_{tp} is the fractional perturbation in the true-redshift bin, while the contamination c_{tp}

[6] Sky spectrum obtained from http://www.gemini.edu/sciops/ObsProcess/obsConstraints/atm-models/skybg_50_10.dat

is the fractional perturbation in the photometric bin. The two quantities can be related via

$$c_{tp} = \frac{N_{\rm t}}{N_{\rm p}}\, l_{tp}, \tag{C4}$$

where $N_{\rm t}$ and $N_{\rm p}$ are the absolute galaxy numbers in the true and photometric redshift bins, respectively. Then

$$n_t \rightarrow n_t \tag{C5}$$

$$n_p \rightarrow (1 - c_{tp})\, n_p + c_{tp}\, n_t \tag{C6}$$

and the photometric bin normalized number density is affected (i.e. biased) by photo-z catastrophic errors. The effect on the cross power spectra is then

$$\begin{aligned}
C_{pp} &\rightarrow (1 - c_{tp})^2 C_{pp} + 2c_{tp}(1 - c_{tp})C_{tp} + c_{tp}^2 C_{tt} \\
C_{mp} &\rightarrow (1 - c_{tp})C_{mp} + c_{tp}\, C_{mt} \qquad (m < p) \\
C_{pn} &\rightarrow (1 - c_{tp})C_{pn} + c_{tp}\, C_{tn} \qquad (p < n) \\
C_{mn} &\rightarrow C_{mn} \qquad \text{(otherwise)}
\end{aligned} \tag{C7}$$

(since the cross power spectra are symmetrical with respect to the interchange of indices, we only consider the biases in power spectra C_{ij} with $i \leq j$). Note that these equations are exact for a fixed contamination coefficient $c_{\rm tp}$.

The bias in the observable power spectra is the right-hand side–left-hand side difference in the above equations.[7] The cumulative result due to all contaminations in the survey (or, $P(z_t|z_p)$ values for each z_t and z_p binned value) can be obtained by the appropriate sum

$$\begin{aligned}
\delta C_{pp} &= \sum_t (-2c_{tp} + c_{tp}^2)C_{pp} + 2c_{tp}(1 - c_{tp})C_{tp} + c_{tp}^2 C_{tt} \\
\delta C_{mp} &= \sum_t \left(-c_{tp} C_{mp} + c_{tp}\, C_{mt}\right) \\
\delta C_{pn} &= \sum_t \left(-c_{tp} C_{pn} + c_{tp}\, C_{tn}\right)
\end{aligned} \tag{C8}$$

for each pair of indices (m, p), where the second and third line assume $m < p$ and $p < n$, respectively.

The bias in cosmological parameters is given by using the standard linearized formula (Knox, Scoccimarro & Dodelson 1998; Huterer & Turner 2001), summing over each pair of contaminations (t, p)

$$\delta p_i \approx \sum_j (\mathbf{F}^{-1})_{ij} \sum_{\alpha\beta} \frac{\partial \bar{C}_\alpha}{\partial p_j} (\mathbf{Cov}^{-1})_{\alpha\beta}\, \delta C_\beta, \tag{C9}$$

where $\mathbf{F}$ is the Fisher matrix and $\mathbf{Cov}$ is the covariance of shear power spectra (see just below for definitions). This formula is accurate when the biases are 'small', that is, when the biases in the cosmological parameters are much smaller than statistical errors in them, or $\delta p_i \ll (\mathbf{F}^{-1})_{ii}^{1/2}$. Here i and j label cosmological parameters, and α and β each denote a *pair* of tomographic bins, i.e. α, $\beta = 1, 2, \ldots, B(B + 1)/2$, where recall B is the number of tomographic redshift bins. To connect to the C_{mn} notation in equation (C7), for example, we have $\beta = mB + n$.

[7] We have checked that the quadratic terms in c_{tp} are unimportant, but we include them in any case.

We calculate the Fisher matrix $\mathbf{F}$ assuming perfect redshifts, and following the procedure used in many other papers (e.g. Huterer & Linder 2007). The WL Fisher matrix is then given by

$$\mathbf{F}_{ij}^{\rm WL} = \sum_\ell \frac{\partial \mathbf{C}}{\partial p_i}\, \mathbf{Cov}^{-1}\, \frac{\partial \mathbf{C}}{\partial p_j}, \tag{C10}$$

where p_i are the cosmological parameters and $\mathbf{Cov}^{-1}$ is the inverse of the covariance matrix between the observed power spectra whose elements are given by

$$\begin{aligned}
\mathbf{Cov}\left[C_{ij}(\ell), C_{kl}(\ell')\right] &= \frac{\delta_{\ell\ell'}}{(2\ell + 1)\, f_{\rm sky}\, \Delta\ell} \\
&\times \left[C_{ik}(\ell)C_{jl}(\ell) + C_{il}(\ell)C_{jk}(\ell)\right],
\end{aligned} \tag{C11}$$

where $\Delta\ell$ is the width of the binning in multipole ℓ. The survey specifications and the cosmological parameter set is described in Section 4.3. Finally, we add the information expected from the Planck survey given by the Planck Fisher matrix (Hu, private communication). The total Fisher matrix we use is thus

$$\mathbf{F} = \mathbf{F}^{\rm WL} + \mathbf{F}^{\rm Planck}. \tag{C12}$$

The fiducial constraint on the equation of state of dark energy assuming perfect knowledge of photometric redshifts is $\sigma(w) = 0.055$.

Our goal is to estimate the biases in the cosmological parameters due to imperfect knowledge of the photometric redshifts. In particular, the relevant photo-z error will be the difference between the inferred $P(z_s|z_p)$ distribution for the calibration (or, training) set – using spectroscopic redshifts as a proxy for the true redshifts – and the $P(z_t|z_p)$ distribution for the actual survey. Therefore, we define

$$\delta C_{ij} = C_{ij}^{\rm train} - C_{ij}^{\rm phot} \tag{C13}$$

$$= \delta C_{ij}^{\rm train} - \delta C_{ij}^{\rm phot}, \tag{C14}$$

where the second line trivially follows given that the true, underlying power spectra are the same for the training and photometric galaxies. All of the shear power spectra biases δC can straightforwardly be evaluated from equation (C8) by using the contamination coefficients for the training and photometric samples, respectively. Therefore, the effective error in the power spectra is equal to the difference in the biases of the training set (our *estimates* of the biases in the observable quantities) and the photometric set (the actual biases in the observables).

APPENDIX D: PROBWTS

In this section, we briefly review the weighting method[8] of Lima et al. (2008) and Cunha et al. (2009). We define the weight, w, of a galaxy in the spectroscopic training set as the normalized ratio of the density of galaxies in the photometric sample to the density of training-set galaxies around the given galaxy. These densities are calculated in a local neighbourhood in the space of photometric observables, e.g. multiband magnitudes. In this case, the DES *griz* magnitudes are our observables. The hypervolume used to estimate the density is set here to be the Euclidean distance of the galaxy to its Nth nearest neighbour in the training set. We set $N = 2$, to derive the most localized estimates possible.

[8] The weights code is available at http://www.stanford.edu/~ccunha/nearest/. The codes can also be obtained in the git repository probwts at http://github.com

The weights can be used to estimate the redshift distribution of the photometric sample using

$$N(z)_{\rm wei} = \sum_{\beta=1}^{N_{\rm T,tot}} w_\beta N(z_1 < z_\beta < z_2)_{\rm T}, \quad \text{(D1)}$$

where the weighted sum is over all galaxies in the training set. Lima et al. (2008) and Cunha et al. (2009) show that this provides a nearly unbiased estimate of the redshift distribution of the photometric sample, $N(z)_{\rm P}$, provided the differences in the selection of the training and photometric samples are solely done in the observable quantities used to calculate the weights.

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