Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY
MNRAS **447,** 3500–3511 (2015) doi:10.1093/mnras/stu2636

Clustering-based redshift estimation: comparison to spectroscopic redshifts

Mubdi Rahman,[1★] Brice Ménard,[1,2†] Ryan Scranton,[3] Samuel J. Schmidt[3] and Christopher B. Morrison[4]

[1]*Department of Physics and Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*
[2]*Kavli IPMU (WPI), the University of Tokyo, Kashiwa 277-8583, Japan*
[3]*Department of Physics, University of California, One Shields Avenue, Davis, CA 95616, USA*
[4]*Areglander-Institut für Astronomie, Auf dem Hügel 71, D-53121 Bonn, Germany*

Accepted 2014 December 11. Received 2014 December 9; in original form 2014 August 3

ABSTRACT

We investigate the potential and accuracy of clustering-based redshift estimation using the method proposed by Ménard et al. This technique enables the inference of redshift distributions from measurements of the spatial clustering of arbitrary sources, using a set of reference objects for which redshifts are known. We apply it to a sample of spectroscopic galaxies from the Sloan Digital Sky Survey (SDSS) and show that, after carefully controlling the sampling efficiency over the sky, we can estimate redshift distributions with high accuracy. Probing the full colour space of the SDSS galaxies, we show that we can recover the corresponding mean redshifts with an accuracy ranging from $\delta z = 0.001$ to 0.01. We indicate that this mapping can be used to infer the redshift probability distribution of a single galaxy. We show how the lack of information on the galaxy bias limits the accuracy of the inference and show comparisons between clustering redshifts and photometric redshifts for this data set. This analysis demonstrates, using real data, that clustering-based redshift inference provides a powerful data-driven technique to explore the redshift distribution of arbitrary data sets, without any prior knowledge of the spectral energy distribution of the sources.

Key words: methods: data analysis – surveys.

1 INTRODUCTION

Mapping of celestial objects occurs fundamentally in two dimensions. Exploring their physical properties, however, requires some knowledge of their distribution along the third dimension, distance. On extragalactic scales, distances are primarily estimated by combining measured redshifts with the expansion history of the Universe. The robustness of redshift estimation can vary dramatically depending on the type of source and the technique used. High-quality spectroscopic redshifts are available only when one can detect and identify a high-contrast spectroscopic feature (emission/absorption lines or spectral break) at sufficient resolution, but such observations are typically expensive and restricted to bright objects. For the vast majority of extragalactic sources, distance estimates rely on photometric redshifts that require a priori knowledge of the type of object observed. They depend on models or spectroscopic training sets and are prone to catastrophic outliers due to degeneracies between the observed colours and redshift. As new surveys will soon map out several billion galaxies, the lack of robust distance estimates is becoming a serious limitation, hindering our exploration of the Universe.

Redshifts can, in principle, be inferred through a different approach: using information encoded within the clustering of matter, rather than the spectral energy distribution of sources. This idea has been discussed for more than 30 years, with attempts made based on angular cross-correlation measurements, such as those by Seldner & Peebles (1979) and Phillipps & Shanks (1987). Landy, Szalay & Koo (1996) furthered the idea, demonstrating that a combination of auto- and cross-correlations between two populations of galaxies can be used to test whether the two samples overlap in redshift space. Along similar lines, Ho et al. (2008) used a combination of spatial auto- and cross-correlations with spectroscopic samples to constrain the redshift distribution of the NVSS radio survey with limited redshift resolution. Over the past eight years, several teams have designed methods aimed at characterizing redshift distributions from spatial correlations, including Schneider et al. (2006), Newman (2008), Matthews & Newman (2010) and McQuinn & White (2013). They primarily considered future (LSST-like) surveys and focused on large-scale clustering (where the galaxy and dark matter fields are related through a linear bias) to determine redshift

★ E-mail: mubdi@pha.jhu.edu
† Alfred P. Sloan Fellow.

distributions with per cent level accuracy. This has been motivated by the requirements of upcoming photometric surveys designed to constrain the properties of dark energy (which are not met by the photometric redshift techniques currently available). While all of these methods are promising, their applicability to real data sets at the promised level of accuracy still needs to be demonstrated.

Another approach to clustering-based redshift inference has been proposed by Ménard et al. (2013, hereafter M13). By construction, their method does not aim at percent-level accuracy, but rather is designed to be directly applicable to existing data sets, optimizing the expected signal-to-noise ratio (S/N) by including small-scale clustering information (i.e. with $r <$ Mpc) to avoid systematic effects often affecting large-scale photometric calibration. Some demonstrative results were presented in Schmidt et al. (2013) with numerical simulations and in M13 with pilot studies using real data sets. In this paper, we implement the technique with a greater degree of sophistication (taking into account sampling considerations, cosmic variance, etc.) and test the reliability of our method by comparing clustering-based redshifts to spectroscopic redshift for galaxy populations selected from the Sloan Digital Sky Survey (SDSS). This work allows us to verify the expected accuracy of clustering-based redshift inference and demonstrate the potential of this new technique.

2 CLUSTERING-BASED REDSHIFT ESTIMATION

Our approach is based on the method introduced by M13. We refer the reader to this paper for the detailed description of the formalism. In this section, we briefly re-introduce the main concepts.

We consider two populations of extragalactic objects: (i) a *reference* population for which the angular positions and redshifts of each object are known. This population is characterized by a redshift distribution $\mathrm{dN_r}/\mathrm{d}z$, a mean surface density n_r, and a total number of sources N_r; and (ii) an *unknown* population for which angular positions are known but redshifts are not. Similarly, this population is characterized by the quantities $\mathrm{dN_u}/\mathrm{d}z$, n_u and N_u.

The basic principle is that if the two populations do not overlap in redshift, their angular correlation is expected to be zero (ignoring gravitational lensing effects). As discussed by M13, in the ideal case of an unknown sample located within a narrow redshift range, one can accurately probe its redshift distribution by splitting the reference population into redshift slices δz_i and measuring the angular or spatial correlations with the unknown population $w_{\mathrm{ur}}(\theta, z_i)$ for each subsample i:

$$\mathrm{dN_u}/\mathrm{d}z \propto w_{\mathrm{ur}}(\theta, z_i)\,. \tag{1}$$

The spatial correlation $w_{\mathrm{ur}}(\theta, z_i)$ is measured over some angular scale θ and is defined by

$$w_{\mathrm{ur}}(\theta, z_i) = \frac{\langle n_u(\theta, z_i)\rangle_r}{n_{\mathrm{u}}} - 1\,, \tag{2}$$

where $\langle n_u(\theta, z_i)\rangle_r$ denotes the mean density estimate of the unknown sample around reference objects at redshift z_i. Following M13, we optimize the S/N of our estimator by considering the integrated cross-correlation function

$$\overline{w}_{\mathrm{ur}}(z) = \int_{\theta_{\min}}^{\theta_{\max}} \mathrm{d}\theta\, W(\theta)\, w_{\mathrm{ur}}(\theta, z), \tag{3}$$

where $W(\theta)$ is a weight function, whose integral is normalized to unity, aimed at optimizing the overall S/N. As an example, if we set the weight function such that $\mathrm{d}W/\mathrm{d}\theta = 0$, the measurement reduces to the integrated overdensity in an annulus from $\theta_{\min}$ to $\theta_{\max}$. To probe the same range of physical scales as a function of redshift, we set $(\theta_{\min}, \theta_{\max})$ to match a fixed range of projected radii $(r_{\mathrm{p,\,min}}, r_{\mathrm{p,\,max}})$ in physical space. Once a cross-correlation signal is found, the amplitude of the redshift distribution is simply obtained through the normalization

$$\int \mathrm{d}z\, \frac{\mathrm{dN_u}}{\mathrm{d}z} = \mathrm{N_u}. \tag{4}$$

The normalization is dependent on all sources in the unknown sample existing within the redshift range of the reference sample.

As in M13, departing from this ideal situation will only cause a modest loss of accuracy in many cases, and redshift inference can still be made with sufficient precision for a large range of astrophysical applications. The degree of departure from an exact solution can be estimated by examining the relative contributions of the terms contributing to the angular correlation function: the redshift distribution $(\mathrm{dN_u}/\mathrm{d}z)$ and the bias-related clustering amplitude of each population. If, over the redshift range Δz, the relative variation of $\mathrm{dN_u}/\mathrm{d}z$ dominates over that of $b_u(z)$, i.e.

$$\frac{\mathrm{d}\log \mathrm{dN_u}/\mathrm{d}z}{\mathrm{d}z} \gg \frac{\mathrm{d}\log \overline{b}_u}{\mathrm{d}z}, \tag{5}$$

we approach the case where $\mathrm{dN_u}/\mathrm{d}z \rightarrow \mathrm{N_u}\, \delta_D(z - z_0)$, and equations (1) and (4) can be used to infer $\mathrm{dN_u}/\mathrm{d}z$. However, this inference is only valid up to a finite accuracy.

The main limitation of the technique is the absence of clustering amplitude information, which places a limitation on its accuracy. Any effect due to gravitational lensing will be negligible in comparison to the signal induced from the clustering of matter, as discussed in M13. With our current approach, we only require the derivative of these clustering amplitudes with redshift, i.e. $\mathrm{d\bar{b}_r}/\mathrm{d}z$ and $\mathrm{d\bar{b}_u}/\mathrm{d}z$. In this analysis we will infer $\mathrm{d\bar{b}_r}/\mathrm{d}z$ from the measured auto-correlation of the reference population as a function of redshift, using the same range of scales and weighting as used in equation (3). The determination of this quantity is presented in Appendix A. We will then treat the clustering amplitude of the unknown population in two ways:

(i) First, we will neglect its contribution, i.e. we assume $\mathrm{d\bar{b}_u}/\mathrm{d}z = 0$.

(ii) Secondly, we assume that, on average, the clustering amplitude integrated over the range of scales considered evolves linearly with redshift: $\mathrm{d\bar{b}_u}/\mathrm{d}z = 1$. This is suggested by the redshift evolution of the amplitude of the galaxy auto-correlation function (see Appendix A for more detail).

Comparing the two estimators will enable an estimation of the uncertainty caused by the absence of clustering amplitude information for the unknown population. According to M13, even ignoring the evolution of the clustering amplitude of the unknown sample (i.e. using $\mathrm{d\bar{b}_u}/\mathrm{d}z = 0$), the technique is expected to provide an estimate of the redshift distribution with an accuracy of $\delta z \sim 0.01$ in the mean for $z < 1$ galaxy populations. We will test this below using a spectroscopic data set for which redshifts are accurately known.

3 DATA ANALYSIS

3.1 The data set

We use the Legacy Spectroscopic Galaxy Sample from the SDSS (York et al. 2000; Abazajian et al. 2009). This sample consists of nearly one million galaxies distributed over >8000 deg^2. It is

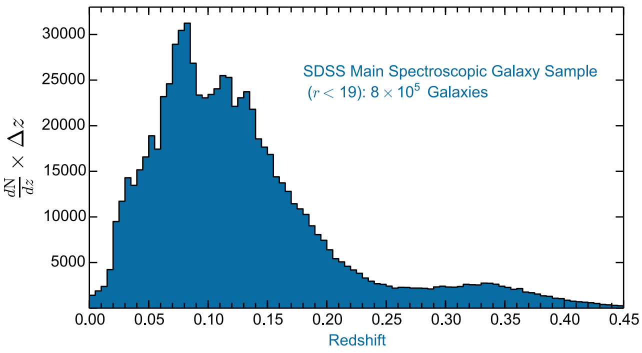

Figure 1. The redshift distribution of the SDSS legacy spectroscopic galaxy sample with $r_{\mathrm{model}} < 19$, presented with binwidth $\Delta z = 0.005$.

composed of two selections: the 'Main Galaxy' sample (Strauss et al. 2002), a complete sample selected with $r_{\mathrm{p}} < 17.77$ mag, and the Luminous Red Galaxy (LRG) sample (Eisenstein et al. 2001), extending to fainter magnitudes. To ensure on-sky uniformity of the sample, we use a limiting magnitude $r_{\mathrm{model}} = 19$ mag, restricting the entire sample to 791 546 galaxies. The corresponding redshift distribution is shown in Fig. 1. Typical spectroscopic redshift errors for these sources are of the order of 10^{-4}.

3.2 Density estimation

The ability to infer robust redshift distributions from the clustering-based technique relies on accurate source density estimation. This requires measuring densities within well-defined angular apertures, taking into account regions of unreliable photometric data coming from missing data, poor photometry, poor sky background estimation, and artefacts from bright stars, satellite tracks, etc. We handle this complexity and perform calculations on the celestial sphere with the astro-STOMP library.[1]

We minimize potential biases from Galactic dust extinction, as well as keep the overall footprint simple, by limiting the present analysis to a region in the Northern Galactic Cap, bounded by $100° < \alpha < 280°$ and $-11° < \delta < 80°$. This field covers about 5400 deg^2. We measure spatial cross-correlations within the range of physical scales 300 kpc $< r_p <$ 3 Mpc at the redshift of each reference subsample. This corresponds to an inner radius of 67 arcsec at $z = 0.3$ and ensures that the fibre collision exclusion (55 arcsec) is avoided in the measurement. As with M13, we choose a weight function $W(\theta) \propto \theta^{-0.8}$ (see equation 3), which is expected to mimic the spatial dependence of the galaxy correlation function. In order to reach the level of precision required in our analysis, we need to account for the small loss of area induced by the presence of additional fibres within the angular apertures (of the order of a few per cent) to obtain sufficiently accurate density estimates. We compensate for this by making the correction to the mean area estimate.

We begin by exploring the ideal regime for clustering-based redshift inference: selecting populations of galaxies in a narrow redshift bin limit. As discussed in Section 2, this regime is expected to provide nearly exact results. We use galaxies between $0.03 < z < 0.3$ and define a set of 'selected' subsamples with $\Delta z_{\mathrm{sel}} = 8 \times 10^{-4}$. Similarly, we define a sequence of reference subsamples with $\Delta z_{\mathrm{ref}} = 2 \times 10^{-4}$, amounting to 1400 redshift bins. In this narrow bin limit, we can use clustering measurements to locate the test samples in redshift space. This can be done without any additional information of the clustering amplitude of each population. Consequently, we use $\mathrm{d}\bar{b}_{\mathrm{u}}/\mathrm{d}z = 0$. For each test sample, we measure the weighted, integrated cross-correlation with the reference subsamples (equation 3) and estimate the normalized redshift distribution using equation (4).

We present the corresponding set of measurements in Fig. 2, showing the estimated density of selected galaxies, $\mathrm{d}^2\mathrm{N}/\mathrm{d}z_{\mathrm{cl}}\mathrm{d}z_{\mathrm{sel}}$, as a function of clustering redshift. Each column of this figure shows the estimated redshift distribution for a population of spectroscopic objects selected within a redshift bin $\Delta z = 8 \times 10^{-4}$. Overall, we find a good agreement between the redshift of each selected population and the clustering-based redshift estimate as indicated by the $z_{\mathrm{cl}} = z_{\mathrm{sel}}$ line. Away from this line, our technique does not indicate any problematic signal at a level greater than a few per cent. The width of the signal in the vertical direction appears to increase towards lower redshifts. This effect, due to the redshift evolution of the clustering amplitude of the selected objects, defines the response function of our technique, i.e. the redshift distribution observed for an input sample located at a single redshift. This quantity is characterized and discussed in more detail in Appendix B. The accuracy of the estimate of the peak position is actually higher than the size of each redshift bin. More than 90 per cent of the signal in these distributions is located within a few resolution elements around the centre of the distribution, with $\Delta z_r < 10^{-3}$.

3.3 Noise properties

To highlight low-level features in Fig. 2, we present a clipped version of the distribution in Fig. 3, limiting the maximum value of $\mathrm{d}^2\mathrm{N}/\mathrm{d}z_{\mathrm{cl}}\mathrm{d}z_{\mathrm{sel}}$ to 60, compared to about 1600 in the previous version. Consequently, these density maps show fluctuations at the few per cent level. The figure demonstrates that, away from the $z_{\mathrm{cl}} = z_{\mathrm{sel}}$ track, the statistically estimated $\mathrm{dN}/\mathrm{d}z$ oscillates around zero with a level in agreement with the expected Poisson noise. While the typical source density of each selected sample is less than one object per square degree, the angular aperture used in the cross-correlation measurements covers about 0.1 deg^2 at $z = 0.3$. In regions where we expect no clustering signal (such as away from $z_{\mathrm{cl}} = z_{\mathrm{sel}}$) the mean angular correlation, and equivalently the measured overdensity, is expected to be consistent with zero. As the resolution of the density map increases, our sampling of the underlying Poisson distribution becomes finer and, given its asymmetry, we expect most of the pixels to display values that are slightly below the overall average. This gives rise to the faint red background in the figure. However, averaging over larger groups of pixels leads to values consistent with zero. This effect becomes stronger when the sample size decreases or when the aperture becomes smaller; the probability of having no object falling within the selected aperture is higher.

Towards the lower left corner of the figure, structure appears away from the $z_{\mathrm{cl}} = z_{\mathrm{sel}}$ track, significantly above the noise level. These are due to three types of effects.

(i) The presence of a massive galaxy cluster in the reference sample leads to a substantial increase in the sampling of the corresponding region of the sky. This creates a spurious signal each time a selected sample presents an overdensity in the same region of the sky. This gives rise to a series of correlations distributed horizontally, located at the redshifts of the massive clusters in the reference sample. This effect can be seen in Fig. 3: a horizontal feature is

[1] The open-source STOMP library is available at https://code.google.com/p/astro-stomp/

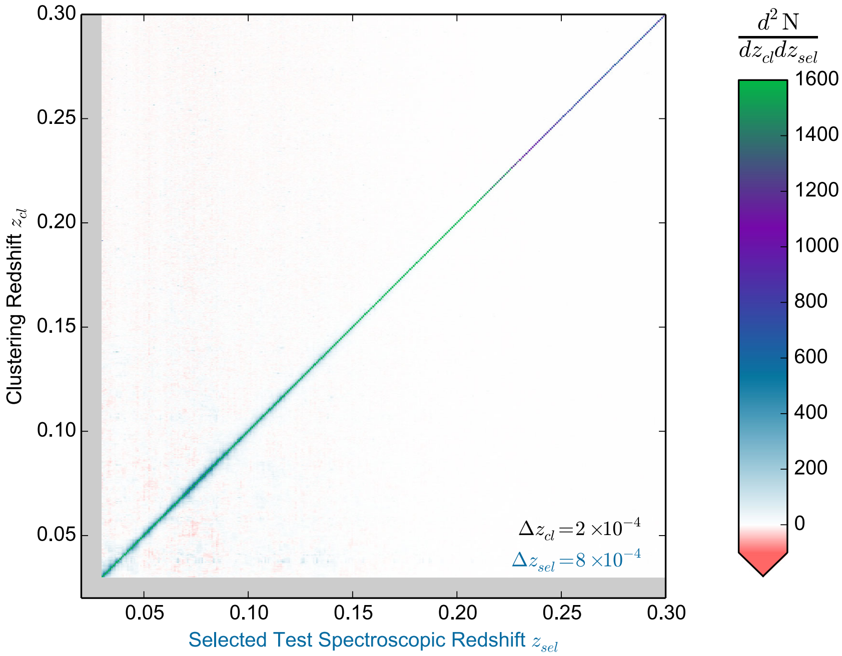

Figure 2. The comparison between clustering redshifts $z_{\rm cl}$ and spectroscopic redshifts $z_{\rm sel}$ for samples selected in narrow redshift bins. The figure shows the density distribution $\mathrm{d}^2N/\mathrm{d}z_{\rm cl}\mathrm{d}z_{\rm sel}$ sampled with $\Delta z_{\rm cl} = 2\times10^{-4}$ and $\Delta z_{\rm sel} = 8\times10^{-4}$, derived from about half a million cross-correlation measurements over the Northern Galactic Cap of the SDSS. A given column corresponds to the redshift distribution inferred for galaxies selected with a given spectroscopic redshift $z_{\rm sel}$.

present at $z_{\rm cl} = 0.037$. It extends from $z_{\rm sel} \simeq 0.05$ to about 0.2. This is caused by the presence of the Hercules Supercluster located at this redshift, the largest cluster in the local universe ($M \sim 10^{16}\,\mathrm{M}_{\odot}$; Barmby & Huchra 1998).

(ii) The radial velocity of an extragalactic object has contributions from both the Hubble flow and its own peculiar velocity. Since only cosmological redshifts correlate with distance, regimes in which peculiar velocities are high will create spurious signals in the $\mathrm{d}^2N/\mathrm{d}z_{\rm cl}\mathrm{d}z_{\rm sel}$ estimates. This effect is strongest when massive galaxy clusters are present. For galaxy clusters with a gravitational potential of the order of $\Delta v = 10^3\,\mathrm{km\,s^{-1}}$, this amounts to $\Delta z \sim 10^{-3}$. The contribution of peculiar velocities affects our ability to properly infer redshifts solely due to the Hubble flow. We note that this degeneracy cannot be removed based on velocity information alone. This leads to a spurious correlation signal perpendicular to and symmetric about the $z_{\rm cl} = z_{\rm sel}$ line. Objects moving around the cluster with a negative line-of-sight velocity, inferred to have a higher redshift, will correlate with cluster objects with a positive velocity, inferred to have a lower redshift. This can be seen in the region $z_{\rm sel} \sim z_{\rm cl} \sim 0.037$. The peculiar velocity effect of the cluster is illustrated as a perpendicular spread of correlation signal (bottom-left of Fig. 3). The velocity spread of the signal ($\sim$2000 $\mathrm{km\,s^{-1}}$) is consistent with the mass of the supercluster.

(iii) Chance superpositions of large-scale structure from two different redshifts, when projected on to the sky, produce an artificial correlation signal: if two structures well-separated in redshift overlap on the sky, the reference galaxies at one redshift will measure an overdensity in the selected sample of the second redshift. Similarly, the reference galaxies at the second redshift will measure an overdensity in the first redshift. Consequently, these spurious correlations appear symmetric about the $z_{\rm cl} = z_{\rm sel}$ line. An example of the first effect is seen in Fig. 3 as structure at $(z_{\rm sel}, z_{\rm cl}) = (0.08, 0.11)$ and symmetrically at $(z_{\rm sel}, z_{\rm cl}) = (0.11, 0.08)$.

These three effects explain the origin of virtually all the structures appearing in the left panel of Fig. 3. The larger volumes sampled at higher redshift minimize the effect of cosmic variance, thereby decreasing the amplitude of these artificial signals. Since the origin of the spurious correlations is primarily due to spatial inhomogeneity of the reference population, we can filter the reference sample to minimize these effects. We describe this procedure below.

3.4 Cleaning the reference sample

The spurious correlations described above are mainly due to the inhomogeneous sampling arising from the clustered distribution of objects in the reference sample. The issues from clustering can be addressed by homogenizing its spatial distribution. This is done by optimally weighting each reference galaxy based on its local density and propagating these weights when characterizing the spatial correlations. Applying this procedure in angular space would maximize the number of usable galaxies in the reference sample but it would not remove the spurious correlation due to the peculiar velocity effects. As the main goal of the present analysis is not to optimize the statistical power of clustering-based redshift inference but simply test its accuracy, we homogenize the sample through a simple selection of the data, keeping only regions of the sky for which the galaxy density does not strongly depart from its mean value and applying an appropriate masking to ensure that effects due to peculiar velocities are negligible. To do so, we split the reference sample into 60 equal-area regions on-sky, each covering $\sim$90 deg^2, and select redshift bins with $\Delta z = 10^{-3}$ in the range $0.03 < z < 0.45$, corresponding to 420 sequential bins. We measure the densities of the cells and keep only those for which the value is within $2\,\sigma$ of the mean. For each region exceeding it, we remove all galaxies from the reference sample within $\Delta z = \pm1.5\times10^{-3}$. We refer to the remaining galaxies as the *cleaned reference sample*.

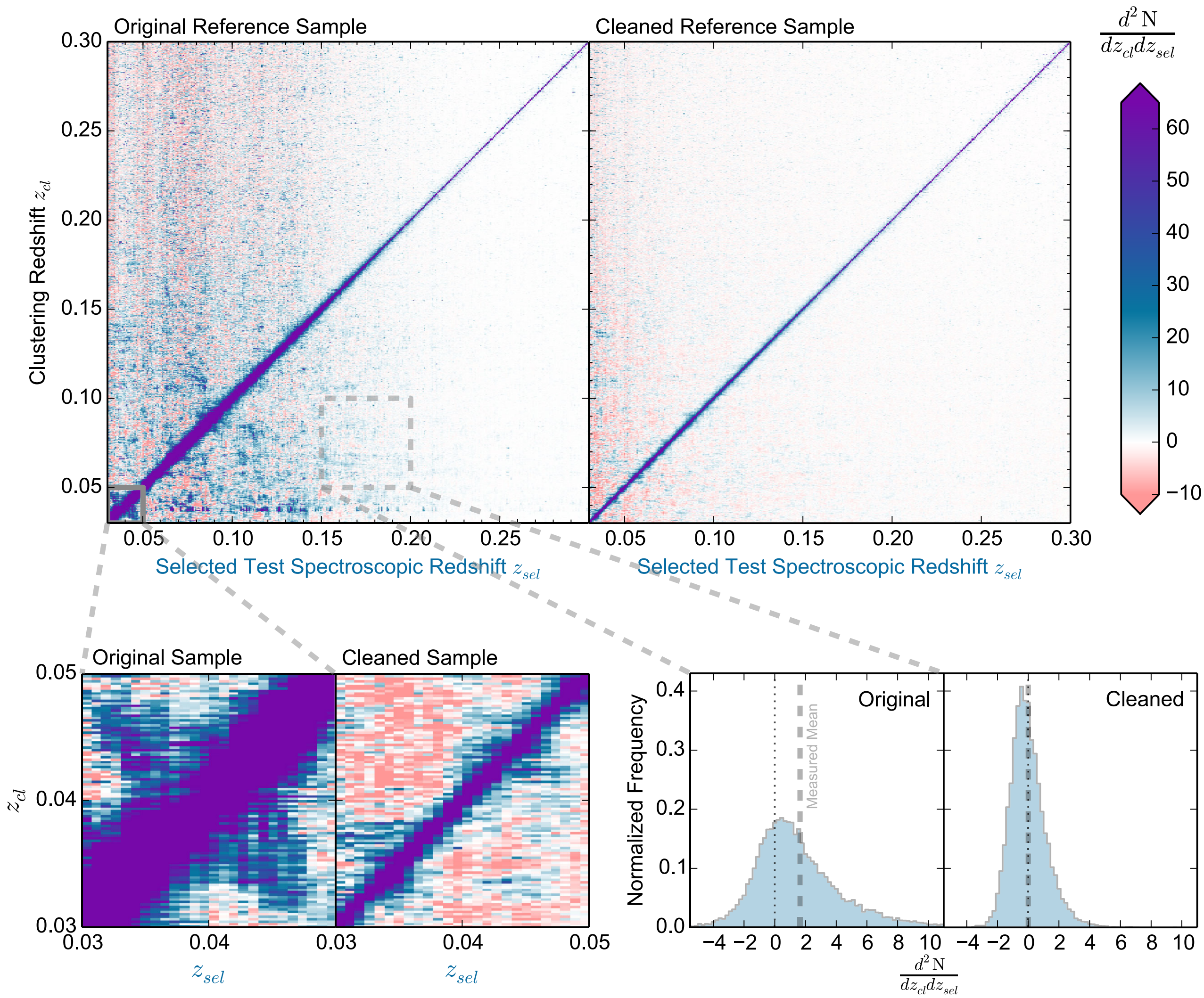

Figure 3. Top: the density distribution $d^2N/dz_{cl}dz_{sel}$ as shown in Fig. 2 but showing only the first few per cent of its dynamical range ($d^2N/dz_{cl}dz_{sel} < 60$). The intrinsic clustering of galaxies from the reference sample imprints correlated patterns which can contaminate clustering-redshift estimation, as shown in the left panel. A spatial homogenization of the density distribution of the reference sample can be applied to suppress these spurious effects, as shown in the right panel. Bottom-left: a zoom-in at the redshift range where the Hercules Supercluster is located. The peculiar velocities of galaxies within the cluster give rise to a signal perpendicular to the diagonal. This effect is almost entirely removed when using the cleaned reference sample. Bottom-right: a histogram comparing the distribution of redshift densities where no signal is expected. The cleaned reference sample allows a reduction in the scatter as well as a more robust zero-point estimate as indicated by the dashed line.

The fraction of regions masked is about 30 per cent at $z = 0.03$ and decreases to less than 5 per cent at $z = 0.3$. Since these regions are selected based on the excess number of galaxies within them, a greater fraction of the galaxies will be masked than the fraction of area spanned on sky. Consequently, this aggressive homogenization strategy masks about 45 per cent of the galaxies, leaving a final cleaned reference sample of 442 000 objects. While this procedure removes a fair fraction of the reference galaxies, the cleaned reference sample is robust to all three spurious correlation effects mentioned above, thereby decreasing the overall noise level in spite of the substantially decreased reference sample size.

The right side of Fig. 3 shows the resulting clipped redshift distribution estimates when using the cleaned reference sample defined above. As compared to the original density plot, the vast majority of the spurious signal disappears. The overdensity originating from the Hercules Supercluster has been suppressed without any a priori knowledge of its location. This is highlighted in the insets shown on the bottom left. The related horizontal track described above, as well as almost all the large-scale overdensities, is no longer detectable. The overall $d^2N/dz_{cl}dz_{sel}$ signal is also more concentrated on to the $z_{cl} = z_{sel}$ line. This leads to a reduction in the overall noise level, despite the use of fewer reference galaxies from the aggressive masking strategy. This can be seen in the subpanel shown in the bottom right of the figure, presenting the distribution of $d^2N/dz_{cl}dz_{sel}$ values. The use of the cleaned reference sample leads to better control of the zero-point, i.e. the estimate of the mean galaxy density. In addition, we observe a significant reduction in the width of the distribution, corresponding to a noise level below 0.3 per cent. We note that the cleaning process is only applied to the reference sample. No operation is applied to the set of selected test galaxy subsamples.

This cleaned reference sample can be used to infer the redshift distribution of any unknown population spanning the redshift range $0.03 < z < 0.45$. It will be used in a companion paper (Rahman

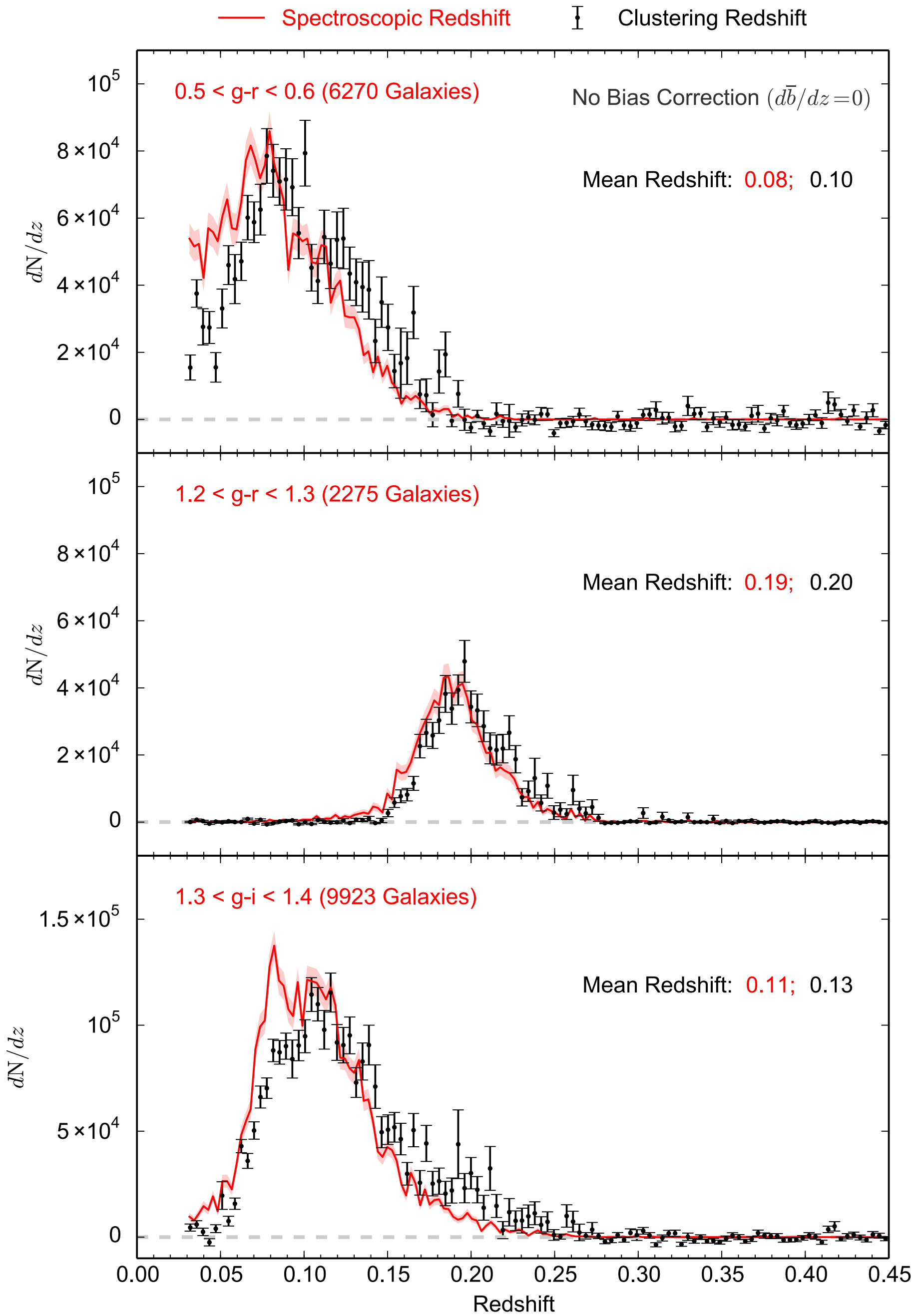

Figure 4. A comparison of the spectroscopic (red line) and clustering redshift distribution (black points) for galaxies selected with limiting magnitude $r < 17.77$ and three different colour cuts: $0.5 < g - r < 0.6$ (top), $1.2 < g - r < 1.3$ (middle), and $1.3 < g - i < 1.4$ (bottom). Here we ignore the possible redshift evolution of the clustering amplitude of the galaxies (i.e. we use $d\bar{b}_{\rm sel}/dz = 0$). The mean clustering-redshifts agree with the spectroscopic ones within $\delta z = 0.02$.

et al., in preparation) to investigate the distribution of clustering redshifts for the entire photometric sample of SDSS galaxies. This analysis demonstrates our ability to robustly estimate redshift for populations selected within a narrow redshift range – the regime in which the method is expected to provide us with nearly exact results.

4 RESULTS WITH PHOTOMETRIC SAMPLES

Having demonstrated the ability to estimate the distributions for narrow redshift samples, we now apply our method to a more generic scenario: galaxies photometrically selected, thus expected to span a finite redshift range. As an illustration, we consider three arbitrary galaxy samples selected by their photometric properties. We use a limiting magnitude of $r < 17.77$ and three colour cuts:

$$\begin{aligned} &0.5 < g - r < 0.6 \ \ (6270 \text{ galaxies}) \\ &1.3 < g - i < 1.4 \ \ (2275 \text{ galaxies}) \\ &1.2 < g - r < 1.3 \ (9923 \text{ galaxies}). \end{aligned} \qquad (6)$$

The corresponding densities on the sky are 0.5–2 sources deg^{-2}. Fig. 4 shows the spectroscopic redshift distribution of these galaxies with the red line. The red contours indicate the estimated Poisson noise of each redshift bin, determined through the variance of the measurement. The selected samples have mean spectroscopic redshifts of 0.08, 0.19 and 0.11, respectively. The width of the distributions is of the order of $\Delta z = 0.05$. We compare these 'true' redshift distributions to those inferred through clustering redshifts using the cleaned reference sample, as defined in Section 3.4. Further, to fully simulate a practical application to real data sets and work with two distinct samples, we exclude the selected galaxies from our cleaned reference sample. We use the parameters listed in Section 3.2 and an estimation of the redshift dependence of the reference sample clustering amplitude $\mathrm{d}\bar{\mathrm{b}}_\mathrm{r}/\mathrm{d}z$ measured from its auto-correlation (see Appendix A for more detail).

We first consider the simplest case for which we ignore the redshift evolution of the clustering amplitude of the unknown population, i.e. we use $\mathrm{d}\bar{\mathrm{b}}_\mathrm{u}/\mathrm{d}z = 0$. The results are shown with the black data points. We estimate the errors from the variance of the mean density measurement. Overall, we observe some level of agreement. In all three examples, the mean clustering redshifts match the spectroscopic ones within $\Delta z = 0.02$. We note that this is in agreement with the predicted error level from M13. We also note that, even if the redshift evolution of the galaxy bias is unknown, a useful property of the clustering redshift technique is the ability to check for the absence of galaxies in a given redshift range. We note that without any assumption, this method enables the inference of the redshift interval over which the photometric sample is distributed, without any assumption on the nature or the spectral energy distribution of the sources.

We now assume that the redshift evolution of the bias of the unknown population is characterized by $\mathrm{d}\bar{\mathrm{b}}_\mathrm{u}/\mathrm{d}z = 1$, i.e. it evolves linearly with redshift. While this is not expected to be exact, it appears to provide a first approximation to the redshift dependence of galaxies selected by stellar mass at $z \lesssim 1$ (see Appendix A). The corresponding results are shown in Fig. 5. The clustering redshift distributions obtained with this simple assumption are found to be in excellent agreement with the spectroscopic redshifts. For each sample considered, the mean clustering redshifts match the spectroscopic one within $\Delta z = 0.01$. The shape of the redshift distributions also appears to be in good agreement with the spectroscopic redshift distributions. We can even observe the bimodality in the distribution of the sample selected with $1.2 < g - r < 1.3$.

Given that the redshift evolution of the bias is the primary uncertainty in this technique, we refer to the discussion in M13 regarding the error introduced by incorrect assumptions of the form of the bias. We note that the error in the redshift determination is most strongly correlated with the width of the redshift distribution. This is a parameter that we derive accurately regardless of bias assumption, thus putting a strong constraint on the accuracy of the resultant distribution.

4.1 Generalization to the full colour space

Having illustrated the potential of the clustering-redshift technique with three arbitrary colour-selected samples, we generalize our analysis to the entire colour space spanned by the SDSS legacy spectroscopic galaxies. To do so, we select galaxies as a function of their $g - r$ and $g - i$ colours. We divide this colour space into square cells with width $\Delta\mathrm{colour} = 0.05$ mag, and for each of them we measure the *mean* spectroscopic redshift. We limit this analysis to cells containing a minimum of 1000 sources. The results are shown in the left panel of Fig. 6. The solid contours indicate the number density of galaxies. Most of the SDSS spectroscopic galaxies have colours $g - r \sim 0.9$ and $g - i \sim 1.2$ mag. The colour scale indicates the mean redshift in each galaxy colour cell. As expected, higher redshift galaxies appear redder. The distribution indicates that the $g - r$ and $g - i$ colours are highly correlated with each other and with mean redshift. One can, however, observe that, at a fixed $g - r$ colour, galaxies redder in $g - i$ appear to be at higher redshift.

We apply the clustering redshift technique to the same sample, using the same colour sampling. We do so using the linear bias correction as described above ($\mathrm{d}\bar{\mathrm{b}}_\mathrm{u}/\mathrm{d}z = 1$). The results are presented in the middle panel of the figure. As can be seen, the agreement between the mean spectroscopic redshifts and clustering redshifts is remarkable over the entire colour space. The differences between the two panels are nearly indistinguishable, demonstrating the generalization of the results found in the previous section. In order to reveal these differences, we show the residuals in the right panel, i.e. $\langle z_s\rangle - \langle z_{\mathrm{cl}}\rangle$. These residuals are all positive, showing that the clustering redshifts tend to be slightly lower than the spectroscopic determination. This reflects that fact that $\mathrm{d}\bar{\mathrm{b}}_\mathrm{u}/\mathrm{d}z = 1$ corresponds to a redshift dependence that is slightly too shallow. Nevertheless, the amplitude of the residuals indicates that the difference in the mean redshift between the two estimators is smaller than 10^{-2}, reaching 10^{-3} in some regions of the colour space. We can recover the redshift distribution of the whole input catalog by summing the redshift distributions from each of the colour samples.

We present the full distribution in Fig. 7, compared to the spectroscopic redshifts, demonstrating that the reconstructed distribution is in agreement with the true distribution.

This application to the SDSS legacy spectroscopic galaxies demonstrates the potential of the data-driven technique of clustering-based redshift estimation. We note that the accuracy we found is in agreement with the theoretical expectations presented in M13. For samples with a redshift distributions characterized by $\Delta z \sim 0.05$, assuming $\mathrm{d}\bar{\mathrm{b}}_\mathrm{u}/\mathrm{d}z = 0$ or 1, we expect mean clustering redshifts to be accurate within 0.02. Our application to the entire colour space of SDSS spectroscopic galaxies has demonstrated this property.

4.2 Inferring the redshift PDF of a single galaxy

Redshift estimation based on photometric information can be described as the mapping that connects volume elements in the space of photometric observables to redshift space. Photometric redshifts determine this mapping with a calibration based on theoretical or observed sets of spectral energy distributions. Our clustering-based estimation aims at determining the same mapping but using spatial correlations.

Fig. 6 is an example of such a mapping, connecting galaxy colour space to redshift space. In each cell of colour space, the application of the clustering-redshift technique produces an estimate of the corresponding redshift distribution. The mapping of the entire space can be used to infer the redshift probability distribution function (PDF) of a single galaxy, either by using the redshift distribution of the corresponding colour cell, or by using more advanced interpolation techniques and making use of information including nearby cells to increase the accuracy of the estimate. This is analogous to the process used to define the redshift PDF of a single galaxy with classical photometric redshifts: using the comparison between observed colours and the spectral energy distributions of modelled or observed galaxies. With clustering redshifts, it is determined using a set of spatial correlation functions.

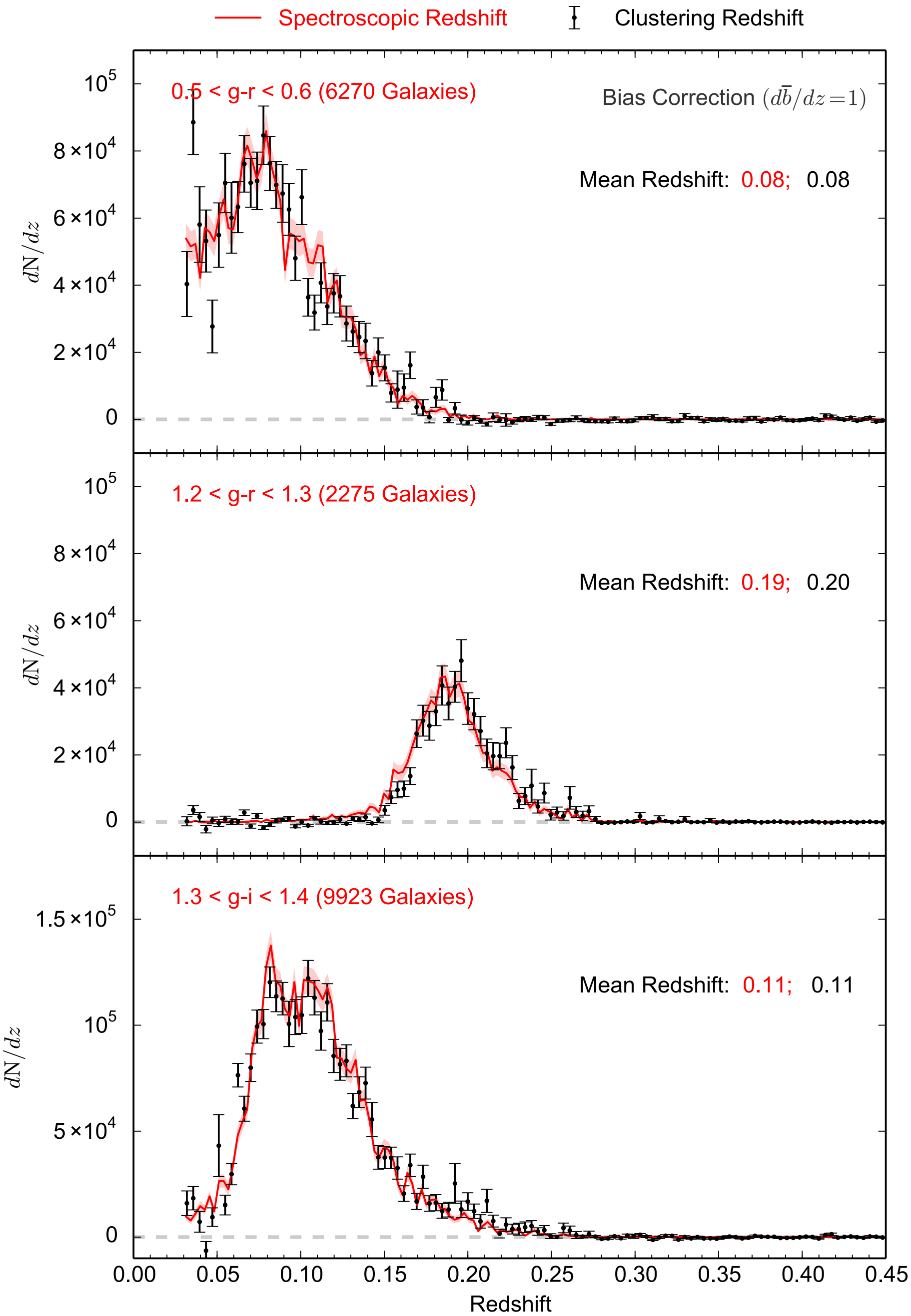

Figure 5. Same as Fig. 4, but using a linear redshift evolution correction, i.e. $d\bar{b}/dz = 1$. The mean clustering-redshifts agree with the spectroscopic ones within a δz better than 0.01.

There exists a unique mapping between a given space of photometric observables and redshift space. Every photometric cell j maps on to a redshift distribution of finite extent Δz_j. Certain regions of this space may map on to multimodal regions of redshift space due to *intrinsic* degeneracies in the mapping itself. There is a limit to how much redshift information can be extracted from the photometry and it applies to both photometric redshifts and clustering-based redshifts in the same way. In the case of a photometric cell mapping on to a multimodal redshift distribution, if subsampling in the other photometric dimensions does not break the redshift degeneracy, all the information used to map the photometric space on to the redshift space has been exhausted.

4.3 Comparison to photometric redshifts

To compare our clustering redshifts with photometric redshifts, we use the KD-Tree nearest neighbour redshifts from Csabai et al. (2007), determining their total redshift probability distribution. This method has been demonstrated to have similar error properties to other empirical and template-based photometric redshift algorithms (Hildebrandt et al. 2010). The individual probability distribution of a given galaxy was modelled as Gaussian using the mean redshift and quoted uncertainty, producing the total redshift distribution by summing over all galaxies in the sample. We note that this comparison represents the ideal case for photometric redshift

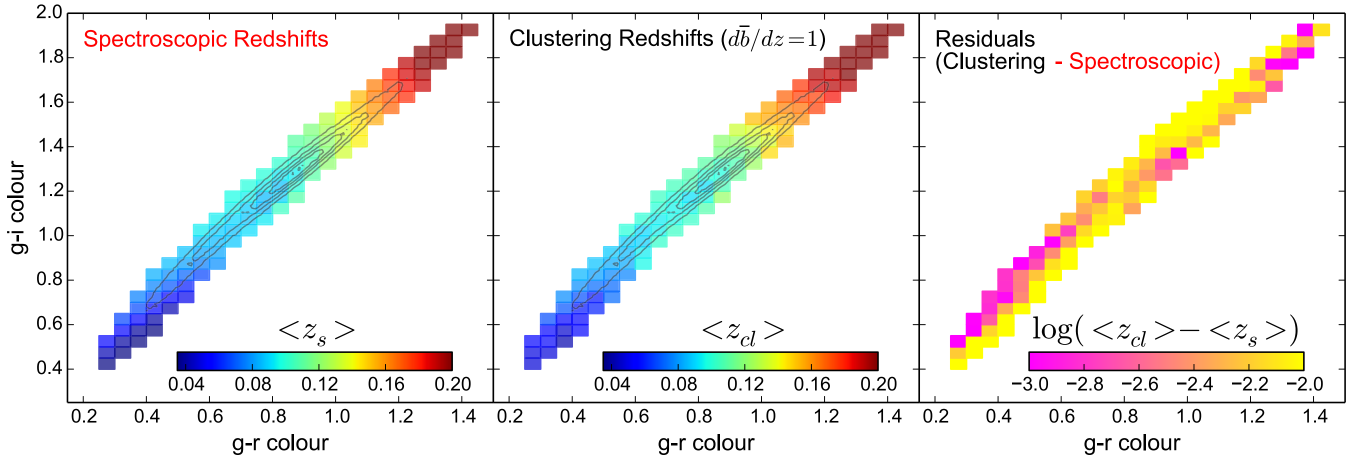

Figure 6. The distribution of the mean redshifts of SDSS spectroscopic galaxies selected as a function of their $g-r$ and $g-i$ colours using spectroscopic redshifts (left) and clustering redshifts with $\mathrm{d}\bar{b}/\mathrm{d}z = 1$ (centre). The solid lines represent the galaxy number count density. We have limited our mapping to colour cells with more 1000 sources. The right panel shows the difference between the two distributions and shows that clustering redshifts are slightly overestimated by δz ranging from 0.001 to 0.01.

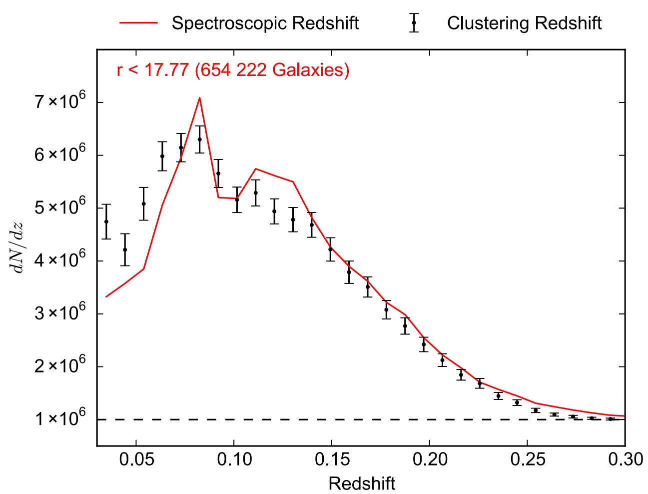

Figure 7. The clustering redshift distribution of the SDSS spectroscopic galaxies with r < 17.77 (black points), produced by summing the individual redshift distributions of the colour-selected samples defined in Section 4.1. The spectroscopic redshift distribution is presented in the red curve. The distribution produced by the clustering redshift agree strongly with the spectroscopic distribution.

inference, since all the galaxies being compared have spectroscopic redshifts and are a part of the algorithm's training sample.

In Fig. 8 we show the spectroscopic, photometric and clustering redshift estimates for two of the three galaxy samples introduced above, i.e. $1.3 < g-i < 1.4$ and $1.2 < g-r < 1.3$. We find that the mean redshift of the three estimators is, in those two cases, relatively similar. However, the overall shape of the redshift distribution indicated by the photometric redshifts significantly departs from the spectroscopic distribution. The peaks of the photometric distributions do not match those of the two other estimators. We note that in each case the photometric redshift distribution displays a high-redshift tail extending beyond the maximum redshift shown in the figure. For the sample selected with $1.2 < g-r < 1.3$, photometric redshifts indicate the presence of a substantial amount of galaxies at $z < 0.12$ while the two other estimates indicate virtually no objects in that range. The bimodality of the distribution shown in both the spectroscopic and clustering redshift distributions for the sample selected with $1.3 < g-i < 1.4$ is not reproduced by the photometric redshifts. This comparison highlights the ability of the clustering-based redshifts to determine the existence of a population at any given redshift, and its robustness to catastrophic failures that plague photometric redshifts.

5 CONCLUSIONS

We have investigated the potential and accuracy of clustering-based redshift estimation, following the method proposed by M13. This technique allows us to infer redshift distributions from the spatial clustering of arbitrary sources with a set of reference objects for which redshift information is available. We applied it to the Main Spectroscopic galaxy sample from the SDSS and after homogenizing the spatial distribution of the reference population over the sky, we find the following.

(i) We can characterize, with a high S/N, the redshift distributions of galaxy samples selected in narrow colour bins corresponding surface densities lower than 1 source deg^{-2}. Ignoring the redshift evolution of the clustering amplitude of the sample, i.e. using $\mathrm{d}\bar{b}_{\mathrm{u}}/\mathrm{d}z = 0$ in our formalism, the clustering-redshift technique provides us with an estimate of mean redshifts with an error $\delta z \sim 0.02$.

(ii) Assuming that the galaxy bias evolve roughly linearly with redshift, i.e. $\mathrm{d}\bar{b}_{\mathrm{u}}/\mathrm{d}z = 1$, the clustering-redshift technique provides us with an estimate of mean redshifts with an error $\delta z \simeq 0.001$–0.01 over the entire colour space of SDSS galaxies.

(iii) The characterization of redshift distributions as a function of galaxy colours provides us with a mapping which can be used to infer the redshift PDF of a single galaxy. This mapping is generic and can be used anywhere on the sky.

(iv) We find our clustering redshift estimates to provide more reliable results than the (KD-Tree) photometric redshifts (Csabai et al. 2007) for the galaxy sample considered.

This analysis demonstrates, using real data, that clustering-based redshift inference provides us with a powerful data-driven technique to explore the redshift distribution of arbitrary data sets, without any prior knowledge of the spectral energy distribution of the sources.

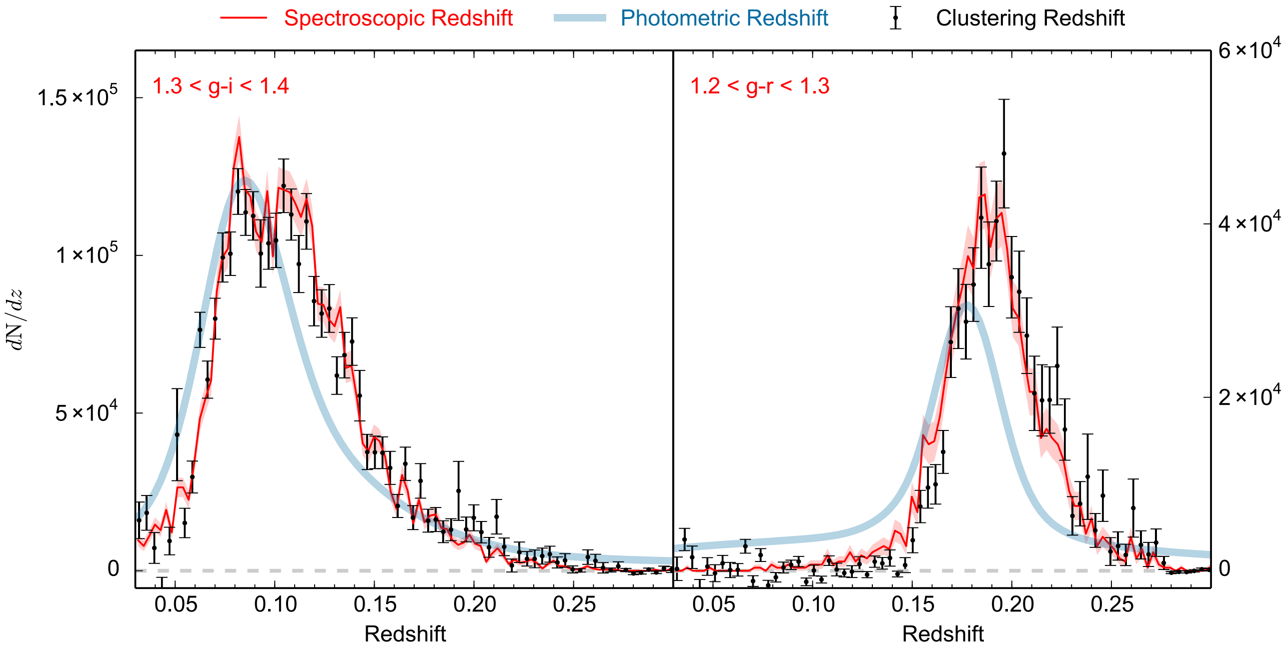

Figure 8. A comparison of the spectroscopic (red), clustering (black) and photometric (blue) redshifts for two colour cut samples from Fig. 4. The photometric redshifts are taken from the KD-Tree nearest neighbour technique (Csabai et al. 2007) using the full probability distributions for each source. The accuracy of the clustering redshift distribution is greater than that from photometric redshifts. We point out the ability for clustering-redshifts to estimate small-scale structure in the distribution, such as the bimodality seen in the left panel – and not captured by photometric redshifts.

ACKNOWLEDGEMENTS

This work is supported by NASA grant 12-ADAP12-0270 and National Science Foundation grant AST-1313302. RS and SJS were supported by National Science Foundation Grant AST-1009514 and Department of Energy Grant DESC0009999.

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

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

Facilities: The Sloan Digital Sky Survey

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Barmby P., Huchra J. P., 1998, AJ, 115, 6
Csabai I., Dobos L., Trencséni M., Herczegh G., Jzsa P., Purger N., Budavri T., Szalay A. S., 2007, Astron. Nachr., 328, 852
Eisenstein D. J. et al., 2001, AJ, 122, 2267
Hildebrandt H. et al., 2010, A&A, 523, A31
Ho S., Hirata C., Padmanabhan N., Seljak U., Bahcall N., 2008, Phys. Rev. D, 78, 043519
Landy S. D., Szalay A. S., Koo D. C., 1996, ApJ, 460, 94
Matthews D. J., Newman J. A., 2010, ApJ, 721, 456
McQuinn M., White M., 2013, MNRAS, 433, 2857
Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, preprint (arXiv:1303.4722) (M13)
Newman J. A., 2008, ApJ, 684, 88
Phillipps S., Shanks T., 1987, MNRAS, 227, 115
Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, MNRAS, 431, 3307
Schneider M., Knox L., Zhan H., Connolly A., 2006, ApJ, 651, 14
Seldner M., Peebles P. J. E., 1979, ApJ, 227, 30
Strauss M. A. et al., 2002, AJ, 124, 1810
York D. G. et al., 2000, AJ, 120, 1579

APPENDIX A: REFERENCE SAMPLE CLUSTERING AMPLITUDE

The accuracy of our clustering-based redshift inference depends on the redshift dependence of the clustering amplitude of both the reference and the unknown samples. Here, we characterize the clustering amplitude $d\bar{b}_r/dz$ of the reference sample we use in our analysis. To do so, we measure the auto-correlation $w_{rr}(z)$ of the reference population as a function of redshift, considering the same range of scales and weighting as used in equation (3) and estimate the clustering amplitude, normalized to an arbitrary redshift z_0, according to

$$\frac{\bar{b}_r(z)}{\bar{b}_r(z_0)} = \sqrt{\frac{\overline{w_{rr}}(z)}{\overline{w_{rr}}(z_0)}} \,. \quad \text{(A1)}$$

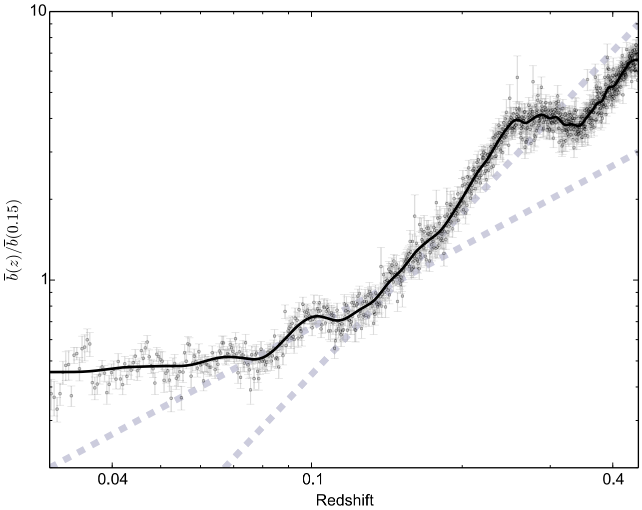

Figure A1. The clustering amplitude redshift evolution as measured from the SDSS Legacy spectroscopic galaxy sample, normalized to 1 at $z = 0.15$. The solid line is the smoothed version of the clustering intensity–evolution curve used to remove this factor from the clustering redshift distributions in this paper. The dashed grey line corresponds to a linearly evolving and quadratically evolving clustering amplitude ($\mathrm{d}\bar{b}/\mathrm{d}z = 1, 2$).

We note that this is different from the classical 'galaxy bias' which is usually defined only on large scales for which the galaxy and dark matter density fields are, on average, linearly related. Our bias definition includes contributions from small scales over which the galaxy and matter fields are non-linearly related. We measure this quantity for our reference sample and we show its redshift dependence in Fig. A1. We also show a smoothed version, for which we convolved the binned measurements with a Hann filter of width $\Delta z = 0.02$, in Fig. A1. For reference, we show two lines indicating the expected evolution for $\mathrm{d}\bar{b}_{\mathrm{r}}/\mathrm{d}z = 1$ and 2, normalized at $z = 0.15$. We find this quantity to be weakly dependent on the maximum scale used in the cross-correlation measurement, measured to an outer radius of 30 Mpc. We note a plateau between $0.25 < z < 0.3$ where the SDSS spectroscopic galaxy selection changes from the main galaxy sample to the luminous red galaxy sample (Eisenstein et al. 2001; Strauss et al. 2002).

APPENDIX B: CLUSTERING REDSHIFT RESPONSE FUNCTION

We investigate the response function of our $\mathrm{d}N/\mathrm{d}z_{\mathrm{cl}}$ estimator, corresponding the intrinsic uncertainty in the technique. In Fig. B1 we present the redshift profiles of the measured redshift density distribution for three different bins of reference redshift. The filled and open symbols show the redshift profiles on either side of the central redshift. The redshift range over which these reference samples are selected is shown by the grey region. Within this range we observe that our technique recovers a roughly flat distribution, as expected. Beyond this limit, our estimator is not consistent with zero (as ideally expected given the absence of reference galaxies in that range) but rather shows a power-law distribution with an index similar to -1 extending over the entire range of redshift intervals considered. This is consistent with the expectation of clustering along the line of sight. This profile drops to below detectable levels beyond $\Delta z \sim 10^{-2}$ or $\Delta d \sim 30$ Mpc. The steep decline of the profile tail enables the clustering redshift technique to be sensitive to small variations in the redshift distribution at a level unreachable by current photometric redshift techniques.

This limit enables precisions previously only possible from spectroscopic redshifts. As has been noted in the text, the vast majority of correlation signal falls within the bounds of the selected sample; the correlation amplitude is a minimum of an order of magnitude greater within the selected sample bounds than in the tail of the response function.

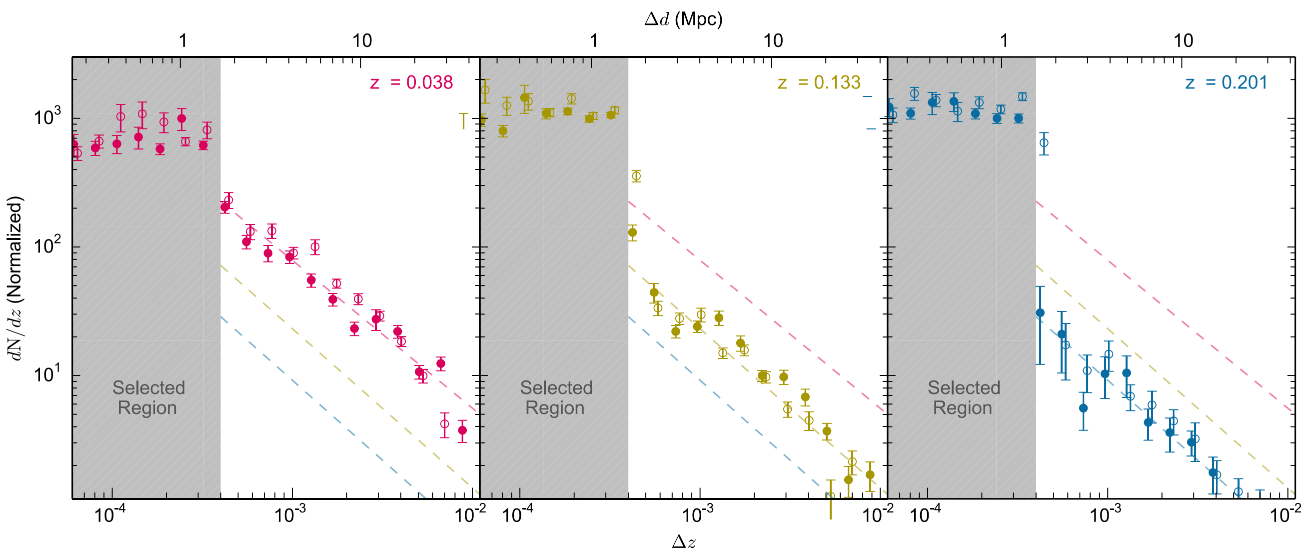

Figure B1. The clustering redshift distribution profiles as a function of distance from the central redshift for three different spectroscopically selected samples. The selected redshift range is labelled in the top-right of each frame. The dashed line represents the power-law fit to the profile between $3 \times 10^{-4} < \Delta z < 3 \times 10^{-3}$. The shaded regions indicate the true width of the selected distribution. In all cases, the tail of the distribution is consistent with the two-point correlation function along the line of sight. This is consistent with the physical scales probed by these redshift scales (indicated on the top axis).

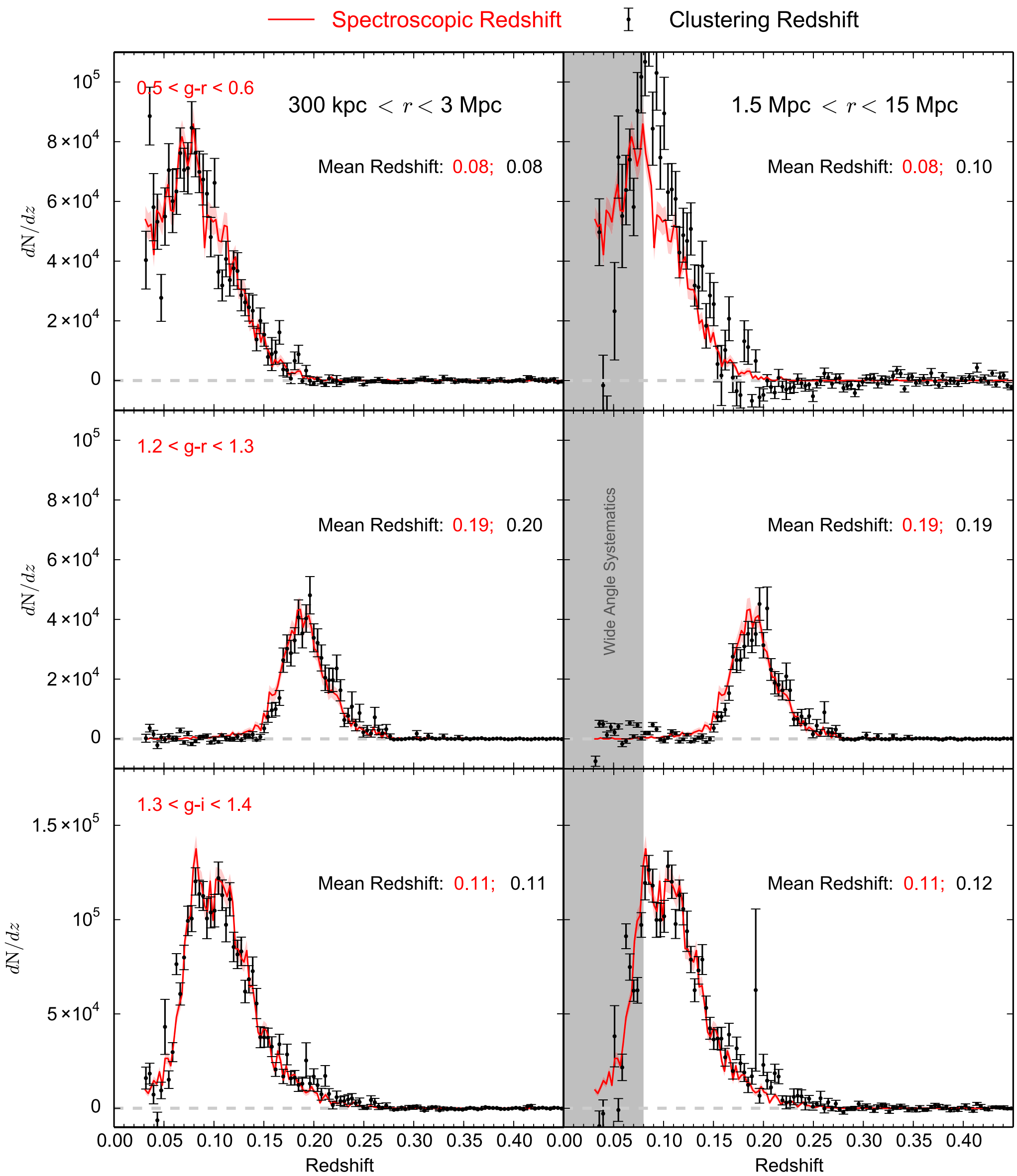

Figure C1. Comparison of the clustering redshift distributions using the smaller cross-correlation annuli (300 kpc $<r<$ 3 Mpc; left column) and larger annuli (1.5 $<r<$ 15 Mpc; right column). Colour cuts and annotations are the same as in Fig. 4. To minimize the effect of cosmic variance, the distributions are normalized to the signal above $z > 0.035$. The distributions are identical using either annuli sizes, demonstrating the scale independence of the technique and the applicability of clustering evolution measurements from larger scales. The cross-correlation measurements for the larger annuli have noise induced by the large angular scale used for measurements at low-z, which we indicate with the grey region; at higher redshift, the measurement is robust.

APPENDIX C: SCALE DEPENDENCE OF CLUSTERING REDSHIFTS

Here we show that our clustering-based redshift inference technique does not strongly depend on the choice of scales used to measure the integrated correlation functions. The method has been shown to be robust to changes in scale through an analysis based on numerical simulations (Schmidt et al. 2013) and we now demonstrate this property from data. We apply the technique to the colour cut samples from Section 4 using the original cross-correlation annulus (300 kpc $< r <$ 3 Mpc) and a much wider cross-correlation annulus (1.5 Mpc $< r <$ 15 Mpc). Ranging a decade of scale in both cases, we expect to extract a similar amount of clustering information from measurements in these two annuli given the slope of the matter correlation function. The results of the comparison are presented in Fig. C1. The measurements from the two different scales are nearly identical, despite covering substantially different areas around each reference source. The differences at low redshift ($z < 0.05$) come primarily from geometric effects of projecting large angles on to a spherical surface, the features present due to cosmic variance (i.e. voids), and the declining accuracy of measuring precise densities decreases as the solid angle increases (i.e. 15 Mpc is equivalent to 7° at $z = 0.03$). The discrepancies between the two scales disappear as redshift increases. The scale independence of this measurement illustrates the applicability of clustering redshifts including small scales where the galaxy and dark matter fields are not linearly related.

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