

Halo occupation distribution estimation performance for LSST data

P. Cataldi^{1,2}, V. Cristiani^{3,4,5}, F. Rodriguez^{3,4}, A. Taverna^{3,6,7}, M.C. Artale⁸, B. Levine⁹, H. Wu¹⁰ & the LSST Dark Energy Science Collaboration

¹ Instituto de Astronomía y Física del Espacio, CONICET-UBA, Casilla de Correos 67, Suc. 28, 1428, Buenos Aires, Argentina e-mail: pcataldi@iafe.uba.ar

² Universidad Nacional Autónoma de México. Instituto de Astronomía. A.P. 70-264, 04510. Ciudad de México, México

³ Instituto de Astronomía Teórica y Experimental - Observatorio Astronómico de Córdoba (IATE, UNC-CONICET), Córdoba, Argentina

⁴ Observatorio Astronómico de Córdoba, Observatorio Astronómico de Córdoba, 854, X5000BGR, Córdoba, Argentina

⁵ Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba (FAMAF-UNC). Bvd. Medina Allende s/n, Ciudad Universitaria, X5000HUA, Córdoba, Argentina

⁶ Instituto de Astronomía, Universidad Nacional Autónoma de México, Apdo. Postal 106, Ensenada 22800, B.C., Mexico

⁷ National Centre for Nuclear Research, Pasteura 7, 02-093, Warsaw, Poland

⁸ Universidad Andres Bello, Facultad de Ciencias Exactas, Departamento de Física y Astronomía, Instituto de Astrofísica, Fernandez Concha 700, Las Condes, Santiago RM, Chile

⁹ Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA

¹⁰ Department of Physics, Southern Methodist University, Dallas, TX 75205, USA

Received XX, XXXX; accepted XX, XXXX

ABSTRACT

Context. Upcoming imaging surveys, such as the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST), will enable high signal-to-noise measurements of galaxy clustering. The halo occupation distribution (HOD) is a widely used framework for describing the connection between galaxies and dark matter haloes, playing a key role in evaluating models of galaxy formation and constraining cosmological parameters. Consequently, developing robust methods for estimating this statistic is crucial for fully exploiting data from current and future galaxy surveys.

Aims. The main goal of this project is to extend a background subtraction method to estimate the HOD with more photometry-based information in preparation for the clustering analysis of the upcoming LSST data and to enable the study of the HOD with significantly improved statistical power. We evaluated the performance of the method using a mock galaxy redshift survey constructed from the COSMOC2 catalogue.

Methods. We implemented an extension of the background subtraction technique to utilize information from photometric galaxy surveys. To identify the centres of galaxy groups, we implemented an iterative centroiding approach (Central Galaxy Finder). We evaluated the impact of each step in our pipeline, including group size estimation from luminosity, purity, and completeness on group identification, as well as the influence of observational systematics such as the use of photometric redshifts and halo mass uncertainties.

Results. We demonstrate the validity of the proposed method using a mock galaxy catalogue, recovering the HOD from COSMOC2 over the absolute magnitude range $M_r = -20.0$ to -17.0 and halo masses up to $10^{15} M_\odot$. We present key performance metrics to quantify the precision and reliability of the group finder and the resulting HOD measurements.

Key words. dark matter – galaxies: halos – large-scale structure of universe – methods: statistical

1. Introduction

Understanding how galaxies trace the underlying dark matter distribution is a central goal of modern astrophysics. In the Λ cold dark matter (Λ CDM) paradigm, galaxies form within the gravitational potential wells of dark matter haloes where structure formation is hierarchical: smaller haloes collapse first and subsequently merge to build larger systems (White & Rees 1978). However, the diversity and complexity of the astrophysical processes involved in galaxy formation and evolution pose significant challenges in determining exactly how galaxies occupy and populate dark matter haloes.

The link between galaxies and haloes is fundamental for several reasons. First, it provides a way to infer cosmological constraints. Galaxy surveys allow us to study the spatial distribution of haloes through measures such as galaxy clustering, and infer key cosmological parameters (van den Bosch et al. 2003; Zheng et al. 2005). Furthermore, they provide a framework for explaining the physics behind galaxy formation, allowing the contribution of central and satellite galaxies to be studied separately (Kravtsov et al. 2004; White et al. 2007; Yang et al. 2008). Observationally, central galaxies are identified as those located near the centres of galaxy groups or clusters, or as galaxies lacking comparably bright neighbours. Theoretically, they are expected to

reside close to the minimum of the gravitational potential of their host dark matter haloes. Satellite galaxies correspond to the remaining group members.

Two commonly used statistical approaches to describe the galaxy–halo connection are the conditional luminosity function (CLF) and the halo occupation distribution (HOD). The CLF (e.g. van den Bosch et al. 2003; Yang et al. 2005; van den Bosch et al. 2007) describes the average number of galaxies with luminosity $L \pm dL/2$ that reside in haloes of mass M_h . The HOD (e.g. Peacock & Smith 2000; Zheng et al. 2005; Tinker et al. 2005) defines the probability distribution $P(N|M_h)$, which represents the likelihood that a halo of mass M_h hosts N galaxies with certain properties, typically distinguishing between central and satellite populations. While both approaches are closely related, the CLF is typically framed in terms of luminosity, whereas the HOD focuses on galaxy number counts.

The development of HOD models in the early 2000s (e.g. Seljak 2000; Wechsler et al. 2001; Scoccimarro et al. 2001; Berlind & Weinberg 2002) represented a major advancement in linking galaxy clustering with halo statistics. These models have since been used to constrain both galaxy formation scenarios and cosmological parameters. Subsequent studies have refined and applied the HOD formalism across a wide range of galaxy samples (e.g. More et al. 2009; Cacciato et al. 2009; Neistein et al. 2011; Avila-Reese et al. 2011; Leauthaud et al. 2012), confirming its effectiveness in connecting observed galaxies to the underlying dark matter structure. The HOD of dark matter haloes typically follows a simple form: satellite occupation scales with halo mass via a power-law relation ($N \propto M_h^\alpha$), and each halo typically hosts a central galaxy (Kravtsov et al. 2004).

A key challenge in studying satellite galaxies is the limited availability of spectroscopic data. Any satellite-counting exercise is complicated by the faintness of the satellites, which makes obtaining redshift information difficult. To address this, Liu et al. (2011) proposed a method based on photometric data to identify candidate satellites surrounding central galaxies, followed by statistical corrections for foreground and background contamination. This approach can be further improved using a background subtraction technique (hereafter BST), which is effective for analysing galaxy populations statistically (Lares et al. 2011). Several studies, from early (Holmberg 1969; Phillipps & Shanks 1987) to more recent efforts (Rodríguez et al. 2015; Duckworth et al. 2019), have demonstrated the value of combining spectroscopic and photometric surveys to detect satellite galaxies at fainter magnitudes.

Rodríguez et al. (2015) validated the BST for estimating HODs using mock catalogues constructed from semi-analytic galaxy formation models within the Millennium simulation (Springel et al. 2005). They found strong overall agreement between the HODs obtained via BST and those derived from direct galaxy counts in volume-limited samples across different absolute magnitude thresholds. Despite a slight overestimation at the faint end ($M_{\text{lim}} \approx -16.0$), the method was shown to be reliable across the full magnitude range ($-21.5 < M_r < -17.0$) (see Fig. 1 in Rodríguez et al. 2015).

In this work, we seek to extend the BST for estimating the HOD using galaxy catalogues primarily based on photometric data. Previous implementations of the BST, such as Rodríguez et al. (2015), relied on spectroscopically confirmed group positions and masses. Our extension pro-

poses and assesses a technique for determining the group central galaxy (and centre) and group size, which defines the background subtraction region, and further incorporates photometric group redshifts, enabling the application of the method to larger galaxy surveys. For satellite galaxies, the method retains the main feature of the classical BST, namely that the only required information for the satellite candidates is their angular positions and apparent magnitudes.

Successfully adapting BST to photometric data would make it applicable to current and forthcoming wide-field surveys such as the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al. 2009; LSST Dark Energy Science Collaboration 2012), Roman (Spergel et al. 2015; Akeson et al. 2019), and Euclid (Laureijs et al. 2011). These surveys will deliver unprecedented sky coverage, depth, and galaxy number densities—LSST alone will observe billions of galaxies out to $z \sim 3$ and identify around 100,000 galaxy clusters up to $z \sim 1$ (LSST Science Collaboration et al. 2009). A BST optimised for photometric data would therefore provide a timely and powerful tool to exploit these datasets, significantly enhancing the statistical power of HOD studies.

This paper is structured as follows. In Sect. 2 we present the mock catalogue and the object sample used in this work. Sections 3.1 and 4 describe the background subtraction and central galaxy identification techniques. Section 5 outlines our testing of the method and details the methods applied for halo mass assignment and photometric redshift estimation. Finally, we summarise our conclusions in Sect. 6.

2. Data

2.1. The DESC Data Challenge 2 dataset

This work is based on the Data Challenge 2 (DC2) dataset (Korytov et al. 2019; Kovacs et al. 2022) developed by the LSST Dark Energy Science Collaboration (LSST Dark Energy Science Collaboration (LSST DESC) et al. 2021), which covers 440 deg² and is designed for testing and developing DESC analysis pipelines. The DC2 dataset adopts a fiducial cosmology close to the WMAP7 best-fit Λ CDM model (Komatsu et al. 2011), with $h = 0.71$, $\Omega_{\text{cdm}}h^2 = 0.1109$, $\Omega_b h^2 = 0.02258$, $n_s = 0.963$, $\sigma_8 = 0.8$, and $w = -1$. The core of DC2 is the COSMOC2 galaxy catalogue, built from the Outer Rim N-body simulation (Heitmann et al. 2019). Dark matter haloes were identified using a friends-of-friends (FoF) algorithm (Davis et al. 1985) (linking length of $b = 0.168$), with two mass definitions provided: the total FoF mass M_{FoF} and the spherical overdensity (SO) mass M_{200} . The SO of the haloes is defined as the mass within a sphere of radius (r_{200}), whose mean internal density is 200 times the critical density of the Universe at redshift z .

Galaxies were populated using UniverseMachine (Behroozi et al. 2019) and GalSampler (Hearin et al. 2020), with their physical and photometric properties generated through the Galactica semi-analytic model (Benson 2010). Spectral energy distributions (SEDs) were built from stellar population models using FSPS (Conroy 2013). The final catalogue includes approximately 2.26 billion galaxies, each with ~ 550 properties—such as stellar mass, morphology, magnitudes in six LSST bands, redshift, and lensing quantities. The simulation assumed an idealised setting, free of dust extinction or stellar contamination.

Although stellar masses were taken directly from the simulation and did not enter the BST—contributing only indirectly to the validation tests—luminosities were used to estimate the group radius, as described in Subsection 5.2. These luminosities were computed in the observer frame by blue-shifting the filter transmission functions according to the cosmological redshift (with peculiar velocities neglected), while k -corrections were interpolated to the light cone redshifts to minimise discreteness effects in the observed colour space.

To account for photometric redshift uncertainties, we used the COSMOC2 photo- z add-on catalogues produced by DESC: FlexZBoost (Izbiicki & Lee 2017) and Bayesian Photometric Redshifts (BPZ; Benítez 2011). FlexZBoost is based on an empirical machine learning method that estimates conditional redshift distributions $P(z|m)$ from photometry in the LSST bands. BPZ is a template-fitting approach that incorporates SED models and galaxy-type priors. FlexZBoost and BPZ have demonstrated high accuracy in previous studies (e.g. Schmidt et al. 2020; Payerne et al. 2025) and are suited for statistical applications such as HOD estimation.

2.2. Group and galaxy catalogues

Group members were defined as galaxies with the same FoF dark matter halo according to the halo IDs provided by the COSMOC2 catalogue. Importantly, selecting groups directly from the COSMOC2 galaxy catalogue ensured that the resulting group catalogue remained unaffected by atmospheric or instrumental effects. The COSMOC2 catalogue includes information on galaxy group positions, redshifts, and halo masses.

We constructed a mock catalogue using synthetic galaxies extracted from a semi-analytic model of galaxy formation applied to the COSMOC2 simulation. We built this catalogue by collecting a comprehensive set of galaxy properties, including angular positions on the sky (RA, DEC), halo id (HALO_ID), magnitude in the LSST r -band (MAG_R_LSST), stellar and halo masses (STELLAR_MASS, HALO_MASS), and a flag indicating whether each galaxy is central (IS_CENTRAL). We also included the true redshift (REDSHIFT), the corresponding photometric redshift estimates from FlexZBoost and BPZ (PHOTO_Z_MEAN), and the derived redshift quality metric (PHOTO_Z_ODDS). From this full dataset, we selected galaxies with a true redshift $z < 0.4$ and apparent r -band magnitude $m_r < 21$, resulting in a sample of approximately 1.15×10^6 galaxies. The magnitude cut ensured high completeness, as $m_r < 21$ can be easily achieved with the LSST single-visit depth, while the redshift limit guaranteed that we retained all potential satellite galaxies associated with the highest redshift volume-limited sample used in this work ($z_{\max} = 0.304$).

Constructing volume-limited samples is crucial to mitigate redshift-dependent incompleteness resulting from observational limits on apparent magnitude. At higher redshifts, fainter galaxies fall below the detection threshold, making it increasingly difficult to identify low-luminosity systems. This incompleteness can indirectly impact the halo mass range being probed, since lower-mass haloes tend to contain lower-mass, fainter galaxies, which are preferentially lost at higher redshift.

To ensure completeness, we defined several absolute magnitude limits, $M_r < -17$, -18 , -19 , and -20 , each cor-

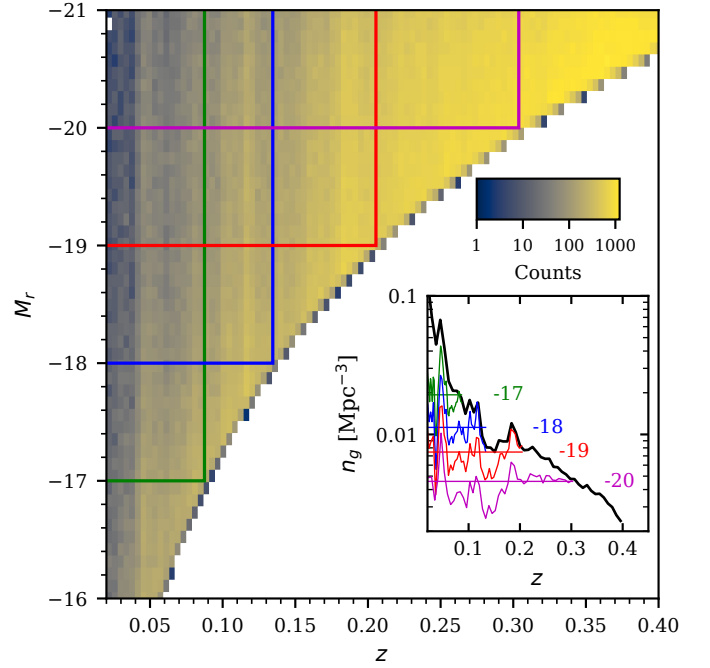


Fig. 1. Construction of volume-limited galaxy samples with varying luminosity thresholds. The colour map shows the distribution of COSMOC2 galaxies as a function of true redshift z and r -band absolute magnitudes M_r . The different coloured lines delineate the selection boundaries for the volume-limited samples. Inset panel: Their mean galaxy number density n_g (colour lines) and total sample density (black line). The solid horizontal lines show the average number density for each sample (the values are given in Table 1).

Table 1. Galaxy samples used in this work.

M_{lim}	z_{max}	N_{gal}	N_{cen}	n_g
-17	0.088	$4.1 \cdot 10^4$	$8.8 \cdot 10^3$	19.3
-18	0.135	$8.5 \cdot 10^4$	$2.9 \cdot 10^4$	11.2
-19	0.206	$1.9 \cdot 10^5$	$1.1 \cdot 10^5$	7.5
-20	0.304	$3.5 \cdot 10^5$	$3.2 \cdot 10^5$	4.6

Notes. Galaxy samples used to test the method described in subsection 5.1. Group centres correspond to central galaxies with absolute magnitudes $M_r < -19.5$. Each row represents a different selection based on magnitude and completeness cuts. The mean galaxy number density n_g is in units of 10^{-3}Mpc^{-3} .

responding to a higher redshift cut and a brighter selection threshold. For each magnitude cut, we applied a redshift limit to guarantee that the brightest galaxy in each group remains observable across the entire volume. The selection criteria applied to the galaxy catalogue is illustrated in Fig. 1, while the resulting subsample properties—including the number of galaxies, number of central galaxies, and their number densities—are summarised in Tab 1.

We restricted the luminosity of group members: they required at least one bright galaxy to be considered a group. We defined a bright galaxy as one with an absolute r -band magnitude $M_r < -19.5$ ¹, following Rodríguez & Merchán

¹ This particular choice was motivated by consistency with the methodology adopted in previous work (Yang et al. 2008; Tempel et al. 2009; Rodríguez & Merchán 2020; Yang et al. 2021)

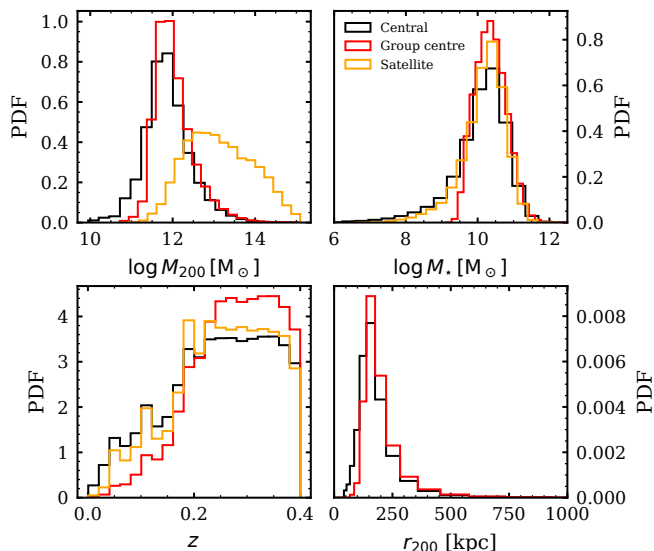


Fig. 2. Normalized distributions for our galaxy catalogue, shown separately for centrals (black), group centres (red), and satellites (orange) galaxies. Top left: Distributions of halo mass M_{200} . Top right: Stellar mass M_* . Bottom left: Galaxy redshift z . Bottom right: Halo radius r_{200} , derived from the halo mass. The halo radius distribution includes only the central and group centre samples. All distributions are shown for the combined sample used in our analysis, which includes galaxies across the four absolute magnitude cuts.

(2020) and Zandivarez et al. (2024). Thus, in our analysis, a group of galaxies is a gravitationally bound system that contains at least one bright (central) galaxy. The central galaxies in this magnitude-limited subset were used to define the group centres.

Central and satellite galaxies exhibit different distributions of their properties, as shown in Fig. 2. The black histogram represents all true central galaxies labelled as centrals by COSMOC2 (without a magnitude cut threshold), while the red histogram shows the group centres defined above. The yellow lines indicate satellite galaxies, defined as those not labelled as centrals in the simulation.

As expected, the halo mass distribution for central galaxies presents a peak at a lower mass than that of satellites. This arises from the fact that low-mass haloes tend to have one (central) galaxy, while massive haloes are likely to host satellite galaxies. Both true centrals and our group centres show similar stellar mass peaks; however, the latter tend to be slightly more massive on average. This reflects the fact that their selection is based on brightness, which naturally favours galaxies with higher stellar and halo masses, excluding fainter, low-mass centrals.

In this work, we aimed to investigate the impact of photometric redshifts on the performance of the BST at determining group centres. To that end, we used for group centres the photometric redshift add-on COSMOC2 catalogues derived from FlexZBoost and BPZ. Fig. 3 presents a comparison between true redshifts (z_{true}) and photometric redshifts (z_{photo}). The dashed lines represent the one-to-one relation, highlighting where photometric redshifts perfectly match the true values. Despite being unbiased, there is considerable scatter between the photometric and true redshifts. We quantified this using the standard deviation

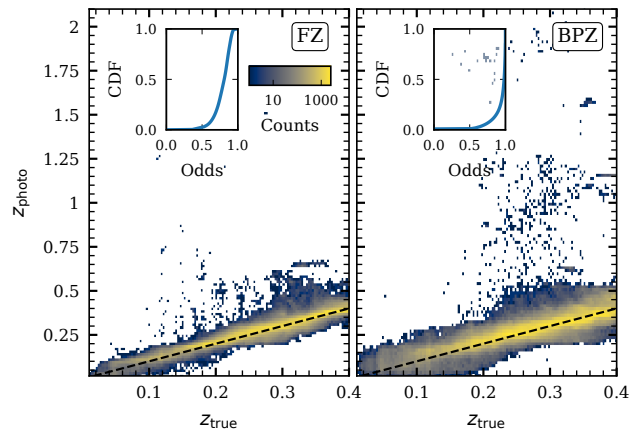


Fig. 3. Left: Comparison between central galaxies' true redshifts (z_{true}) and photometric redshifts using FlexZBoost (z_{photo} , mean value). Right: Same as in the left panel but for BPZ redshifts. The dashed line represents the one-to-one relation ($z_{\text{photo}} = z_{\text{true}}$). Inset panels: Cumulative distribution function of Odds parameters.

of the normalised redshift residuals, $(z_{\text{photo}} - z_{\text{true}})/(1 + z_{\text{true}})$, obtaining $\sigma = 0.02$ for FlexZBoost and $\sigma = 0.04$ for BPZ. Nevertheless, FlexZBoost results should be considered optimistic due to the deep selection of the training complete sample up to $i < 25.3$, based on expected spectroscopic sample available by the 10-year data release (and thus unrealistic for real data), while the ‘discreteness’ in the colour-redshift of the modelled galaxies (see Sect. 5.3.2 of Korytov et al. 2019) negatively impacts the performance of BPZ. As pointed out in Payerne et al. (2025): the actual LSST data will most likely fall between the pessimistic BPZ and optimistic FlexZBoost runs we analysed.

This uncertainty in redshift, and therefore in inferred galaxy positions, significantly complicates the recovery of galaxy clustering signals. Nonetheless, photometric surveys offer the advantage of much larger galaxy samples. Large errors in the photometric redshift are especially common for faint blue galaxies, whose colours can appear similar at both low and high redshifts, making them difficult to distinguish in photometric surveys. The quality-derived redshift PHOTO_Z_ODDS parameter (hereafter, Odds) provides additional insight into the reliability of a photometric redshift estimate (see inset panels of Fig. 3). It quantifies how peaked the redshift probability distribution function (PDF) is by integrating the PDF within a defined interval around its mode. A high Odds value indicates a narrow, well-constrained PDF, suggesting that the algorithm is confident in a specific redshift solution. For reference, the percentage of galaxies with Odds > 0.8 in our sample is $\sim 57\%$ for FlexZBoost and $\sim 91\%$ for BPZ. In contrast, a low Odds value implies that the PDF is either broad or multi-modal, reflecting greater uncertainty or multiple plausible redshift solutions. In the context of this work, we tested the performance of photometric redshift estimates and, where appropriate, applied an Odds-based selection to our galaxy catalogue to obtain more constrained subsamples. We note, however, that while an Odds cut can significantly remove a good deal of the outliers, it does not eliminate them entirely and may also remove a substantial number of well-estimated galaxies with point estimates near the true redshifts.

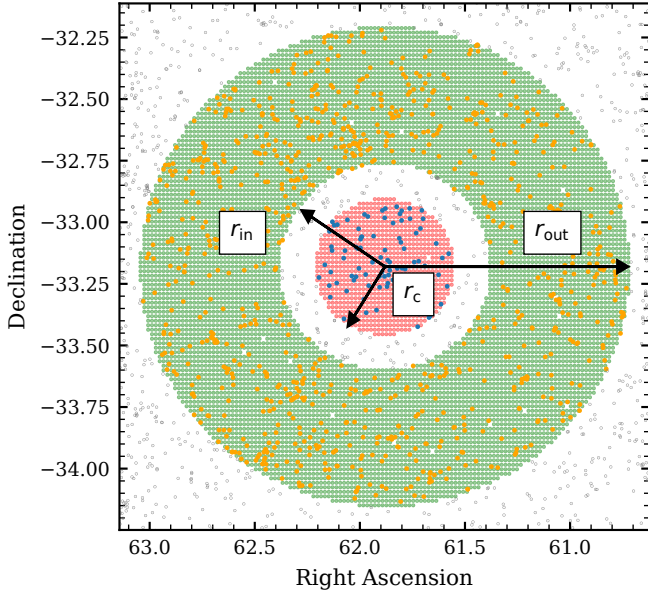


Fig. 4. Illustration of the BST for a single group. The red-shaded area and blue dots indicate the region and galaxies located within the characteristic radius of the host halo ($r < r_c$). The green-shaded area and yellow dots represent the region and galaxies located in the outer annulus ($r_{in} < r < r_{out}$). The grey dots correspond to the galaxies that lie outside the region considered by the method.

3. Background subtraction technique

3.1. The method

The BST has been validated through comparisons with other HOD estimation methods based on observational data. The BST method can yield results consistent with those obtained from spectroscopically identified galaxy groups, while enabling the study of satellite populations at fainter absolute magnitudes than previously accessible (see Fig. 5 in Rodriguez et al. 2015). The method has been successfully applied in various contexts, including the investigation of the environmental dependence of kinematically misaligned galaxies using MaNGA data (Duckworth et al. 2019) and the study of the mass–richness relation using galaxy catalogues from the Sloan Digital Sky Survey (Gonzalez et al. 2016).

The background subtraction procedure assumes that the large-scale distribution of galaxies is isotropic and homogeneous and that galaxy groups represent local overdensities. This technique enables estimating of the number of galaxies that appear to belong to a group in projection, but are in fact located either in front of or behind it (commonly referred to as background galaxies, although some may lie in the foreground), using statistical analysis. Given the sky area and depth of LSST-like photometric catalogues, we can estimate the HOD over a wider magnitude range and improve the overall statistics by combining spectroscopic surveys for central galaxies with photometric galaxy catalogues. The primary advantage of this approach is its ability to determine the HOD of satellite galaxies to fainter magnitudes than previously studied.

In Fig. 4, we show the application of BST to a single group extracted from the COSMOC2 simulation. The figure illustrates the circular aperture used to select can-

didate members, and the surrounding annuli employed to estimate the local background density. To apply the BST, we followed the steps detailed below:

- Redshift assumption: Given that only photometric information is available, we assumed that all galaxies are located at the same redshift as the central galaxy of the group, z_{gr} . In this work, we first adopted the true redshift for the group centres and subsequently extended the analysis to the case where photometric redshifts were used. The absolute magnitude was then computed as:

$$M = m - 25 - 5 \log(d_L(z_{gr})) \quad (1)$$

where $d_L(z_{gr})$ is the luminosity distance in megaparsec at z_{gr} .

- Galaxy counting: We defined n_i as the number of galaxies with absolute magnitudes $M_r < M_{lim}$ within a circle centred in each group with a radius determined by a given projected characteristic radius of the halo (r_c), typically taken to be r_{200} . The M_{lim} value set the upper limit for the range of absolute magnitudes where the HOD was estimated.
- Background estimation: A statistical method is necessary, as the background contribution to n_i is challenging to determine directly. Given the hierarchical nature of large-scale structure, it is known that a given overdensity is always embedded in a larger density mode, which may cause the matter density nearby to be greater or less than the cosmic mean. To estimate the number of background galaxies (n_0), we counted those that satisfy the same magnitude criterion in an annulus ($r_{in} < r < r_{out}$) around each group, rather than using a global average.
- Background subtraction: To discount the contribution of background galaxies, which cannot be distinguished from the group galaxies, we needed to determine their number density in the region surrounding the group. We computed the corrected number of galaxies N in each group using:

$$N = n_i - \frac{n_0}{A_0} A_i \quad (2)$$

where n_0 is the galaxy count in the annulus, and A_i and A_0 are the projected areas of the group circle and background annulus, respectively.

- Sample selection: Since the method has no restrictions on the choice of groups, we carefully selected the samples used in each measurement. Specifically, low-mass groups cannot be detected at high redshifts because they typically contain faint galaxies, while high-mass groups can be observed across the entire redshift range. This was addressed by setting a redshift limit z_{max} for each mass bin, such that the brightest galaxy in each group would be detectable.
- HOD estimation: We binned the groups by halo mass and computed the HOD as the average number of galaxies per group: $\langle N | M_h \rangle$, considering only those groups for which galaxies with M_{lim} , located at z_{gr} , would have an apparent magnitude brighter than the limiting apparent magnitude of the photometric catalogue.

3.2. Extension to photometric data

Traditionally, implementing BST requires detailed information for each group, including halo mass, redshift, angular position, and a characteristic radius, alongside a galaxy catalogue containing angular positions and magnitudes. We adapted this framework to operate solely on photometric catalogues, removing the reliance on spectroscopically confirmed group properties. To achieve this, we incorporated a complementary method to identify potential central galaxies directly from photometric data. This approach is referred to as the Central Galaxy Finder (CGF) and is outlined in Sect. 4. We also leveraged available photometric redshift estimates to statistically constrain galaxy membership and estimate the HOD.

Our approach is more challenging due to the inherent uncertainty in photometric redshifts, but it enables the application of the method to the full depth and area of upcoming wide-field photometric surveys. Moreover, this implementation lays the groundwork for a wide range of future studies on the galaxy–halo connection, including halo assembly bias, galactic conformity, and secondary clustering dependence on properties such as colour and environment density.

4. Central Galaxy Finder

In this work, we extended the BST algorithm to enable its application to galaxies with only photometric information. To make this extension feasible, we first needed to identify potential galaxy groups—or at least their central galaxies—using photometric data alone, since the BST method requires accurate group positions.

In this context, we introduce a method to identify central galaxy candidates based on their brightness relative to the local galaxy density. This step provides the necessary input catalogue of group positions and properties to which the BST can be applied. Below, we outline the main steps employed in our CGF (see Fig. 5 for a schematic illustration.)

Step 1: Initial Selection. We began by selecting all galaxies with an absolute magnitude in the r band of $M_r < -19.5$. These galaxies were considered potential group candidates and were ordered in decreasing luminosity. Step 1 is illustrated in panel 1 of Fig. 5, showing the full sample of luminous galaxies.

Step 2: Discard-radius estimation. Starting with the brightest galaxy in the list, we estimated its discard group radius using the relation $r_{\text{dsc}} = f_r \cdot r_c$, where f_r is a free scaling parameter and r_c is the previously mentioned characteristic radius of the halo. The radius r_{dsc} defines the projected search region for potential group members and is referred to as the ‘discard radius’. This projected volume for the brightest galaxy is shown in panel 2 of Fig. 5.

Step 3: Member Selection. We selected all galaxies that lie within a projected distance $r < r_{\text{dsc}}$ and have a redshift difference $\Delta z < f_z$, where $\Delta z = |z_{\text{can}} - z_{\text{gal}}|$. Here, z_{can} is the redshift of the brightest galaxy, and z_{gal} is the redshift of the other candidates. The parameter f_z is another free variable of the method. All selected galaxies were marked as members of the group, and they were then removed from the catalogue for subsequent steps. Panel 3 of Fig. 5 shows the catalogue after discarding the identified group members.

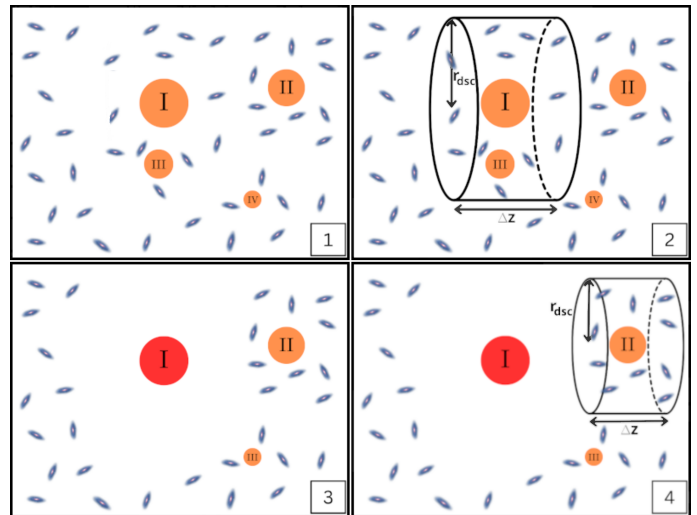


Fig. 5. Scheme of the CGF method applied to a galaxy sample. Top left: Selection of candidate group centres, ranked by brightness (orange dots). Top right: Definition of a discard region centred on the brightest galaxy (I), determined by its luminosity. Bottom left: Removal of all galaxies within the discard region from the candidate list, except for galaxy (I). Bottom right: Continue with the next remaining brightest galaxy. In all panels, the line of sight lies along the direction perpendicular to the plane of the plot (i.e., viewed from the side of the plot).

Step 4: Iteration. We then selected the next brightest galaxy that had not been previously discarded and repeated Steps 2 and 3. Panel 4 in Fig. 5 illustrates this step, where the selection moves to the next remaining bright galaxy. This iterative process continued until the entire catalogue was processed.

For reliable HOD estimation, high central galaxy purity across the full halo mass range is essential—and also the most challenging to achieve. Purity is defined as the percentage of selected candidates that are truly central galaxies, and completeness is defined as the fraction of true centrals that are correctly identified.

In Fig. 6, we present the resulting purity and completeness of the identified central galaxy candidates for a fixed set of parameters: $(f_r, f_z) = (2.5, 0.25)$, shown as a function of absolute magnitude (left panel), stellar mass (middle), and halo mass (right), with different coloured lines indicating median values under various magnitude cuts representing different observational selections. The choice of this fixed set of parameters improves the purity of true centrals (above $\sim 70\%$ over halo mass) by excluding interlopers over a larger projected and redshift volume. Although this configuration results in lower completeness, the large statistical samples expected from future photometric surveys can partially compensate for this reduction. However, this may introduce a selection bias that must be accounted for, since completeness depends strongly on the properties of the central galaxies and their host haloes, as discussed below.

The U-shaped trend in purity across M_r , $M_\star$, and most prominently M_{200} , arises naturally from the way the CGF operates. Because the CGF initially selects the brightest galaxies in the catalogue, these early candidates are, in the vast majority of cases, true centrals. They are also less likely to be removed in the iterative discard steps of the algorithm. This leads to both high completeness and high purity at

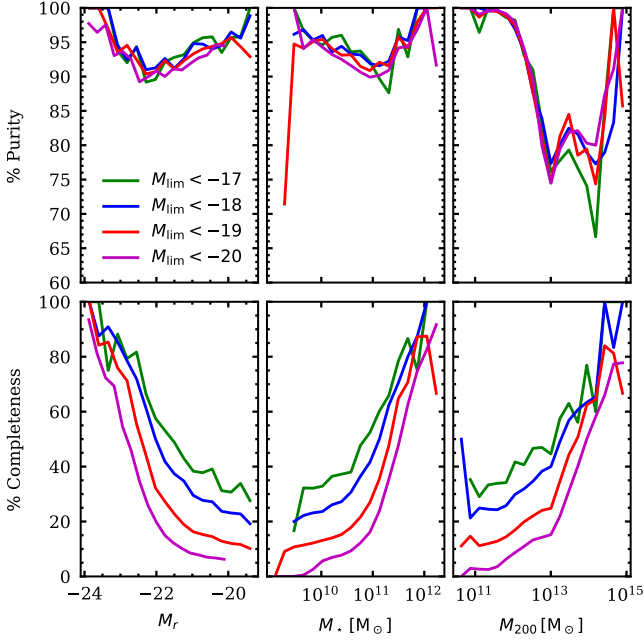


Fig. 6. Top panels: Purity of the resulting catalogue as a function of absolute magnitude (left), stellar mass (middle), and halo mass (right). The galaxies were selected from the CGF method with a fixed $f_r = 2.5$ and a redshift bin width of $f_z = 0.25$. Bottom panels: Corresponding completeness shown as a function of the same variables. The different coloured lines represent the median values for the volume-limited samples.

the bright or massive end. As the CGF continues through the candidate list, progressively fainter galaxies are examined. At this stage, completeness decreases because more galaxies are discarded during iteration, including some genuine centrals. Purity also declines because the removal of a true central can result in an ‘orphan’ group, for which the CGF assigns the brightest remaining galaxy in the local region—often a satellite. This effect drives the drop in purity at intermediate magnitudes, stellar masses, and halo masses.

At the faint and/or low-mass end, the completeness is indeed lower (these galaxies are considered last in the CGF iteration cycle), but most of the corresponding haloes host only one or two members. In such cases, the CGF classification is typically correct, which produces the recovery in purity at the faint end. Finally, the more pronounced U-shape observed when purity is plotted against M_{200} reflects the fact that the CGF selection is based on brightness (and therefore indirectly on $M_\star$), so purity is expected to be systematically lower and to show stronger variations across halo mass bins.

To further understand the trade-offs involved, Fig. A.1 (Appendix A) presents the purity and completeness of the CGF as a function of halo mass (M_{200}) in the range 10^{12} – $10^{15} M_\odot$, focusing on this high-mass regime where achieving high purity for central galaxies is particularly challenging. We find a clear tension between purity and completeness across different combinations of f_r and f_z : high values of both parameters increase the exclusion volume in each iteration of the CGF, resulting in high purity but reduced completeness. Conversely, lower values of f_r and f_z improve completeness by retaining more candidates, but at

the cost of lower purity due to increased contamination. Given that future surveys such as LSST will provide excellent statistics through large galaxy samples, we opted for a parameter set that prioritises higher purity, ensuring more accurate central galaxy identification and thus more reliable HOD recovery.

The CGF naturally favours isolated, bright galaxies, leading to high purity when selections are based on stellar mass or magnitude. Alternative algorithms can also be used to identify galaxy associations in large photometric surveys, including group finders such as that of Yang et al. (2021) and cluster-finding algorithms such as REDMAPPER (Rykoff et al. 2014), CFSFDP (Zou et al. 2021), and WAZP (Aguena et al. 2021). Each method introduces selection biases arising from the trade-off between purity and completeness. Our CGF-derived central-galaxy catalogue remains broadly consistent with the true centrals from the simulation (see Fig. A.2). For comparison, Yang et al. (2021) report $\geq 80\%$ completeness in member detection, but only $\geq 70\%$ of groups reached high purity, and their method is reliable only for haloes above $10^{12.5} M_\odot$. Likewise, REDMAPPER and WAZP achieve an almost unbiased cluster catalogues but operate at even higher masses ($M \gtrsim 10^{14} M_\odot$), well above the regime relevant for our HOD analysis, where lower-mass systems are more prone to misclassification and mass overestimation.

5. Results

In the following sections, we present a step-by-step evaluation of our method. In Subsection 5.1, we validate the BST using simulation-based quantities such as true redshifts, angular positions, halo masses, and central galaxy identifiers to define group centres. Subsection 5.2 explores an alternative approach to estimating group sizes by substituting halo size with luminosity derived from absolute r -band magnitudes. In Subsection 5.3, we investigate the impact of FlexZBoost and BPZ photometric redshifts on the inference of the HOD. Subsection 5.4 assesses the impact of uncertainties in halo mass estimation. Finally, in Subsection 5.5, we evaluate the performance of our method on the CGF outputs applied to photometric data.

5.1. Evaluation of the method

This section aims to evaluate the performance of BST. To this end, we use the mock catalogue introduced earlier, which provides full knowledge of galaxy populations within dark matter haloes, enabling a direct comparison between the true group properties and those recovered by our method. As a first test, we applied the technique under idealised conditions: using the true angular positions, redshifts, and halo masses of central galaxies to define group locations, along with the true angular positions of satellite galaxies. This setup isolates the intrinsic accuracy of the method prior to introducing observational uncertainties.

We adopted a characteristic radius of $r_c = r_{200}$, as it serves as a good upper limit for group sizes, defined as the maximum Cartesian distance between the central galaxy and its most distant satellite. The COSMOC2 intra-halo distribution of satellites follows a truncated NFW profile (Korytov et al. 2019), so we expect that the satellite population to drop sharply after $r > r_{200}$.

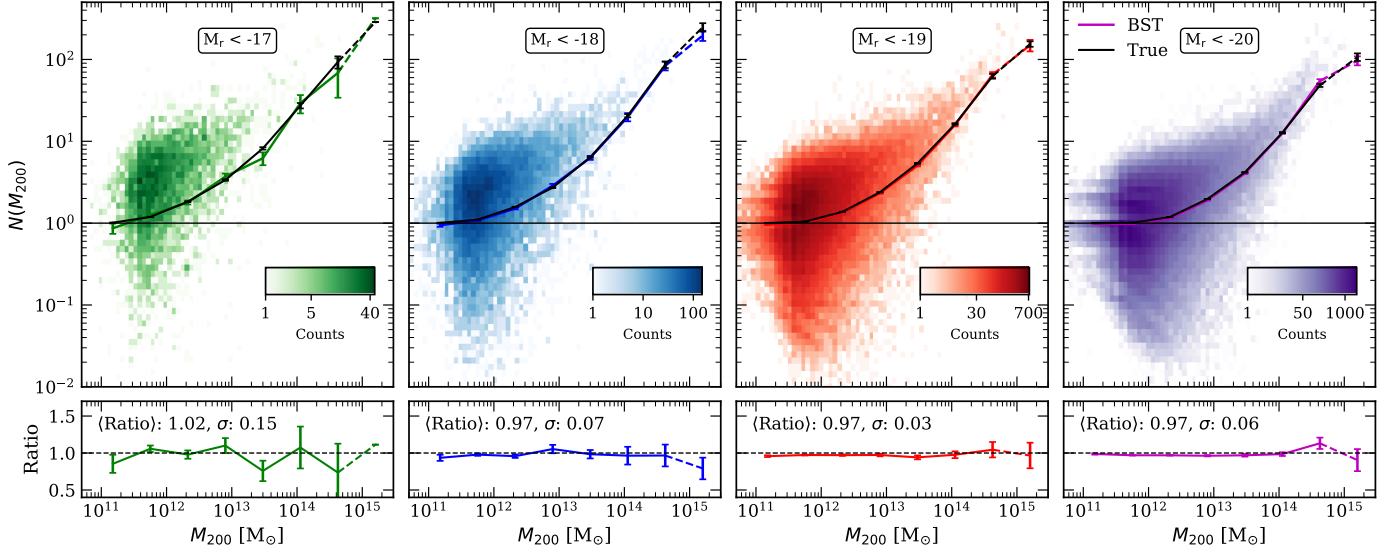


Fig. 7. Comparison between the BST results (colour lines) and the true HOD obtained by direct counts in volume-limited samples (black lines). Each panel shows results for $M_{\text{lim}} = -17, -18, -19$ and -20 . The colour maps represent the occupation number obtained for each volume-limited sample according to the BST, while the coloured lines indicate the median values of the corresponding distributions. The dashed lines correspond to the highest halo-mass bins ($M_{200} > 5 \times 10^{14} M_{\odot}$), which are affected by low-number statistics. For a better quantification of the differences, the lower panels indicate the ratio between the BST-recovered and true HOD, where we show the median and dispersion values. All uncertainties were calculated by the standard jackknife procedure.

We imposed a maximum radius of $r_c < 1$ Mpc to avoid excessively broad annuli. This ensures that the computed local galaxy density is not diluted by including regions representative of the cosmic mean density, which would otherwise lead to an overestimation of group membership in massive clusters. Only a small fraction of groups exceed this upper limit: about 0.2% for $M_{\text{lim}} = -17$, and even lower fractions for the brighter samples. Therefore, this cut-off has a minimal impact on the overall catalogue.

To compute the background contribution as described in Sect. 3.1, we defined an annular region of $1.5 r_c < r < 3.5 r_c$. We tested different combinations of inner and outer radii of the annular region, specifically $r_{\text{in}} = \{1.5, 2.0, 2.5\}$ and $r_{\text{out}} = \{3.5, 4.0, 4.5, 5.0\}$. We find that the BST results remain stable across all choices, as long as the annular region excludes actual group members, but is not so large that it samples environments unrepresentative of the local density near the group.

Figure 7 shows the HOD across a broad range of halo masses. The true HOD was computed from the known galaxy group memberships in the mock catalogue by binning in halo mass and averaging the number of galaxies per bin. The black lines represent the median true HOD. The background colour maps display the 2D distribution of galaxy counts recovered using the BST, while the coloured lines show the median HOD for different galaxy samples defined by varying magnitude limits. Uncertainties were estimated using a random jackknife sample, in which the BST output is randomly divided into 50 equal subsets and the HOD is recomputed while omitting each subset in turn. The lower panels show the ratio between the BST and the true HOD, for which we computed the median and dispersion as a quantitative measure of the BST performance.

We find that the agreement between the BST-recovered and true HOD generally improves as the magnitude limit becomes brighter (i.e. lower M_{lim}), which increases the effective redshift range (up to $z < 0.3$) and enhances sample

completeness. This improvement is also due to the increased sample size in these brightness bins (see Table 1), leading to lower statistical scatter. We note that the median and dispersion of the ratio between the BST-recovered and true HOD are typically below unity; however, in all cases, they remain consistent with unity within the quoted uncertainties. Across the halo mass range, we observe both underestimation and overestimation of the HOD, indicating no systematic bias of the BST-recovered HOD. The offset between the median and the mode of the BST distribution arises from the way the BST computes the number of galaxies in each group through Eq. 2, which can yield $N(M_{200}) \leq 0$ for some groups, particularly in low-density environments. In Fig. 7, we display the results on a logarithmic scale, excluding these non-positive values. Nevertheless, all $N(M_{200})$ are included when computing the median relation shown in the HOD analysis.

The highest halo mass bins ($M_{200} > 5 \times 10^{14} M_{\odot}$), indicated with a dashed line, suffer from poor statistics due to the limited number of massive groups in the COSMOS2 simulation. Consequently, the median values in this regime should be interpreted with caution. We include these results for completeness, noting their statistical limitations, and because the recovered HOD still shows reasonable agreement. This issue is expected to be significantly alleviated with future datasets from wide-area surveys such as LSST, which will cover $\sim 18,000 \text{ deg}^2$ and provide a vastly larger sample of both galaxies and massive clusters.

Interestingly, the mock catalogue itself exhibits flattening of the HOD at the high-mass end, with the number of satellites not increasing as expected. This deviates from the typical HOD behaviour, where the satellite population continues to grow with halo mass, following a power-law slope. Notably, the BST-recovered HOD reproduces this same trend, demonstrating its ability to capture the underlying distribution even in regimes affected by modelling.

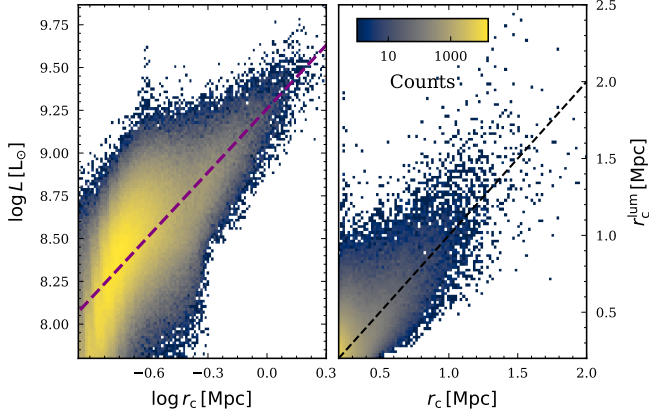


Fig. 8. Left panel: Relation between r_c and luminosity for central galaxies in our catalogue, with the best linear fit shown as dashed purple lines. Right panel: Relation between r_c and the estimated r_c^{lum} , with the dashed line indicating the one-to-one relation.

5.2. Luminosity-based group size

Reliable determination of group sizes in flux-limited samples is challenging because the number of detected members decreases with distance, introducing strong selection effects (Yang et al. 2008). We aimed to estimate the characteristic radius of the host halo used in BST from the observed luminosity of the central galaxy of the group (r_c^{lum}) to avoid relying on parameters from the simulation, which we would not have access to in real data.

We computed the linear fit between host halo characteristic radius (r_c) and luminosity for central galaxies, as shown in the left panel of Fig. 8. This relation enables an interpolation scheme to infer r_c , serving as a proxy for group size, directly from the total luminosity of a system. This approach is particularly useful when only photometric data are available.

We propose the relation

$$r_c^{\text{lum}} = C_1 \cdot L + C_2 \quad (3)$$

where the best-fitting parameters are $C_1 = 6.1 \times 10^{-10} \text{ Mpc}/L_\odot$ and $C_2 = -0.03 \text{ Mpc}$. We show the comparison of r_c^{lum} with r_c in the right panel of Fig. 8. Although both characteristic halo radii can be approximately related through a linear relation, the comparison reveals a significant dispersion, which is expected to impact the HOD estimation.

The results when using r_c^{lum} instead of r_c show good agreement between the recovered HOD and the true HOD across the full halo mass range, presented in Fig. 9. While adopting a more flexible relation between central galaxy luminosity and size—such as a broken power law or a mass-dependent slope—could potentially yield even better results, our current linear parametrisation proves sufficiently accurate and robust for this analysis. This simple model is particularly valuable given the challenge of estimating the group radius, which typically requires a prior determination of its halo mass. In contrast, our approach allows a practical and effective estimation of group size directly from the luminosity of the central galaxy, enabling a reliable recovery of the HOD without detailed mass modelling. In the following sections, we therefore discuss adopting r_c^{lum} , computed

from the luminosity of the central galaxies, to validate the method.

5.3. Photometric redshift

In this subsection, we investigate the impact of using photometric redshifts by adopting redshift estimates from FlexZBoost and BPZ. Specifically, we replaced the true redshift from the group centres, z_{gr} , with the photometric redshift, z_{photo} , for estimating r_c^{lum} using BST via Eq. 1. This substitution introduces uncertainty in the absolute magnitude of central galaxies and their associated groups, significantly complicating the recovery of the HOD. Moreover, some migration of galaxies between the volume-limited subsamples is expected once photometric redshifts are applied, as is the case with real observational data.

In this test, we did not apply any quality cut on the photometric redshifts, allowing us to test the robustness of the BST with respect to photometric redshift in the ‘worst’ case scenario—i.e. with no mitigation of the impact of outliers on the final results. The resulting HODs obtained using the BST under these conditions are presented in Fig. 10. Both photometric redshift estimators are generally in good agreement with the truth across all magnitude cuts. BPZ yields lower performance than FlexZBoost, especially in the faintest cut ($M_{\text{lim}} = -17$), due to more dispersed and biased photometric redshifts with respect to the true redshifts.

Uncertainties in the photometric redshifts of group centres do not contribute significantly to the background subtraction estimation of the total HOD. We note that, as shown in Fig. 10, the HOD inferred for groups identified using photometric redshifts exhibits only a mild underestimation at the high-mass end, while otherwise accurately reproducing the true HOD.

5.4. Halo mass estimation

In photometric galaxy surveys, estimating halo masses is inherently challenging due to projection and miscentering effects, limited group membership, and uncertainties in galaxy properties. For instance, Yang et al. (2021) obtained halo mass estimates for each group by applying an abundance matching method using total group luminosity ranking, which links stellar mass to halo mass based on cumulative number density matching (e.g. Moster et al. 2018; Behroozi et al. 2019). Although this empirical approach provides a physically motivated alternative to group-based dynamical mass estimators, accurate stellar masses, star-formation rates, and luminosities are required to obtain reliable halo mass estimates, typically with a mean scatter of $\sigma \sim 0.20 \text{ dex}$ (Van Kampen et al. 2026).

Relying solely on apparent magnitudes of central-galaxy candidates (without confirmed group membership) substantially increases the uncertainties, making it challenging to robustly recover the HOD. An additional source of error arises from redshift uncertainties inherent in the use of photometric catalogues. Our goal here is not to reproduce a specific observational halo-mass estimation pipeline but rather to evaluate whether the HOD recovery remains stable under levels of halo-mass uncertainty comparable to those expected in realistic photometric group catalogues, where both accurate and inaccurate halo-mass estimates coexist. To assess the impact of these uncertainties on the inferred

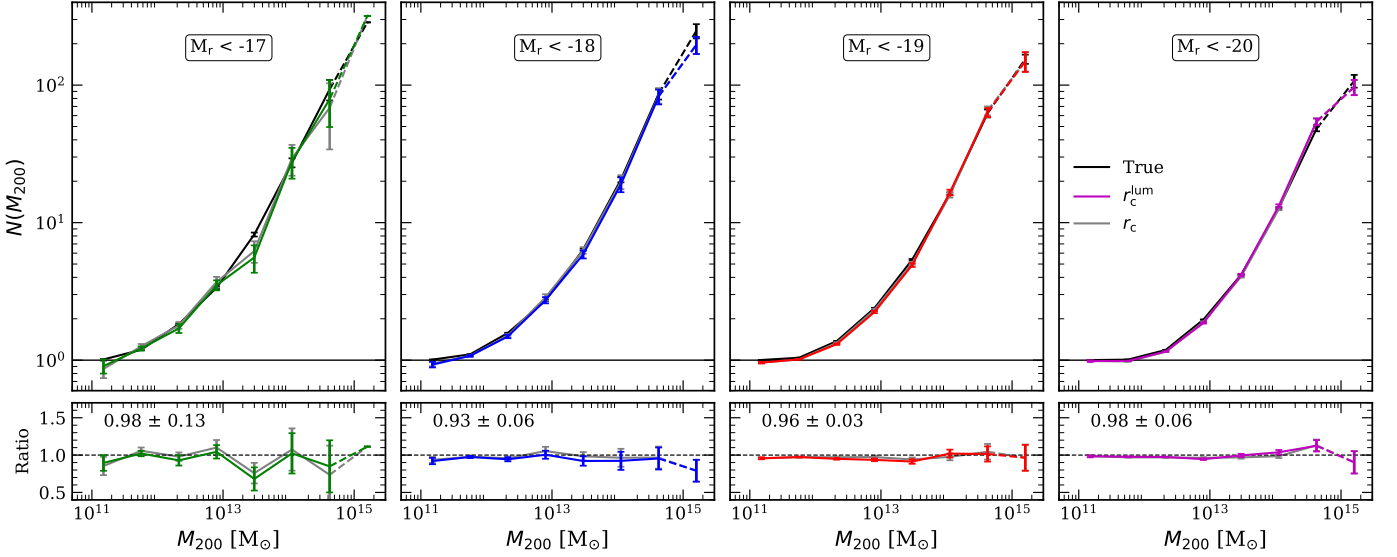


Fig. 9. BST results using r_c^{lum} derived from the central galaxy luminosity using Eq. 3. The grey lines indicate the recovered HOD when using r_c , while the black lines indicate the true HOD. Bottom panels: Ratio between the BST-recovered and true HOD, together with the corresponding median and dispersion.

HOD, we performed a controlled test in which we added a Gaussian dispersion of 0.1 and 0.2 dex to $\log M_{200}$. These values reflect the typical maximum scatter observed in the empirical relation between the halo mass and the number of group members. Figure B.1 (Appendix B) shows the resulting redistribution of halo masses produced by these uncertainties. The associated uncertainty is the largest at the low- and high-mass ends and decreases towards the intermediate mass range.

Figure 11 presents the results of these tests. As shown, the added halo mass uncertainties do not produce a large impact on the recovered HOD, although small deviations become apparent at the high-mass end for the largest mass perturbation considered. BPZ performance weakens, particularly for the fainter magnitude cuts. The overall good agreement between the HODs obtained under different levels of halo mass scatter indicates that the method is robust against typical observational errors in halo mass estimation.

5.5. Application of the CGF

In this final subsection, we discuss incorporating the CGF technique into the previously BST recovered HOD to evaluate the impact of group-centre misidentification in combination with all systematic effects considered so far (luminosity-based group size, photometric redshift errors, and a halo mass uncertainty of 0.1 dex). We used CGF fixed parameters $(f_r, f_z) = (2.5, 0.25)$ —corresponding to $r_{\text{dsc}} = 2.5 \cdot r_c^{\text{lum}}$ and $\Delta z = 0.25$ —motivated by previous analysis (see Sect. 4). Consequently, the total number of galaxies identified as central decreases significantly, primarily due to the lower completeness associated with these conservative selection parameters. This reduction is quantified in Table 2, where the decrease in the number of selected centrals reflects the trade-off between completeness and purity. Nonetheless, the gain in purity justifies this choice, as it enables a more reliable reconstruction of the true HOD.

In Fig. 12, we present the results of applying the BST to this refined sample of central galaxies. We further re-

Table 2. Number of central galaxies identified with the CGF method.

M_{lim}	z_{max}	$N_{\text{cen}}^{\text{CGF}}$	$N_{\text{cen}}^{\text{Odds}}$
-17	0.088	$2.9 \cdot 10^3$	$2.3 \cdot 10^3$
-18	0.135	$6.8 \cdot 10^3$	$5.8 \cdot 10^3$
-19	0.206	$1.3 \cdot 10^4$	$1.2 \cdot 10^4$
-20	0.304	$2.0 \cdot 10^4$	$1.9 \cdot 10^4$

Notes. Rows correspond to different absolute magnitude limits, M_{lim} , and the corresponding maximum redshift, z_{max} . $N_{\text{cen}}^{\text{CGF}}$ gives the number of central galaxies identified with the CGF method using $(f_r, f_z) = (2.5, 0.25)$, while $N_{\text{cen}}^{\text{Odds}}$ lists the subset of these galaxies with photometric redshift quality metric Odds > 0.8.

stricted the sample to objects with Odds > 0.8, to reduce the contribution of photometric redshift outliers while preserving a high fraction of the original CGF-identified central galaxies. The HOD recovery shows good agreement with the true underlying distribution for the brighter magnitude limits ($M_{\text{lim}} = -19, -20$). However, performance degrades for the other subsamples, which can be attributed to the low number of galaxies in this subsample (fewer than 10^4). Due to reduced statistics from the CGF application, achieving good agreement at the faint end remains challenging for FlexZBoost and BPZ, but both perform well in the brighter cuts. Larger simulations with broader sky coverage—such as a potential extension with SKYSIM5000—would provide improved sampling and enable more robust constraints on the HOD in this regime. Additionally, we note that more sophisticated group-centre identification techniques may yield further improvements; however, the approach adopted here provides satisfactory performance while maintaining a relatively simple methodology.

We further investigated the impact of varying the CGF parameters, f_r and f_z , within the BST framework. Fig. A.3 (Appendix A) shows the effect of changing these parameters on the recovered HOD. We find that increasing f_r leads to better agreement between the recovered and true HOD,

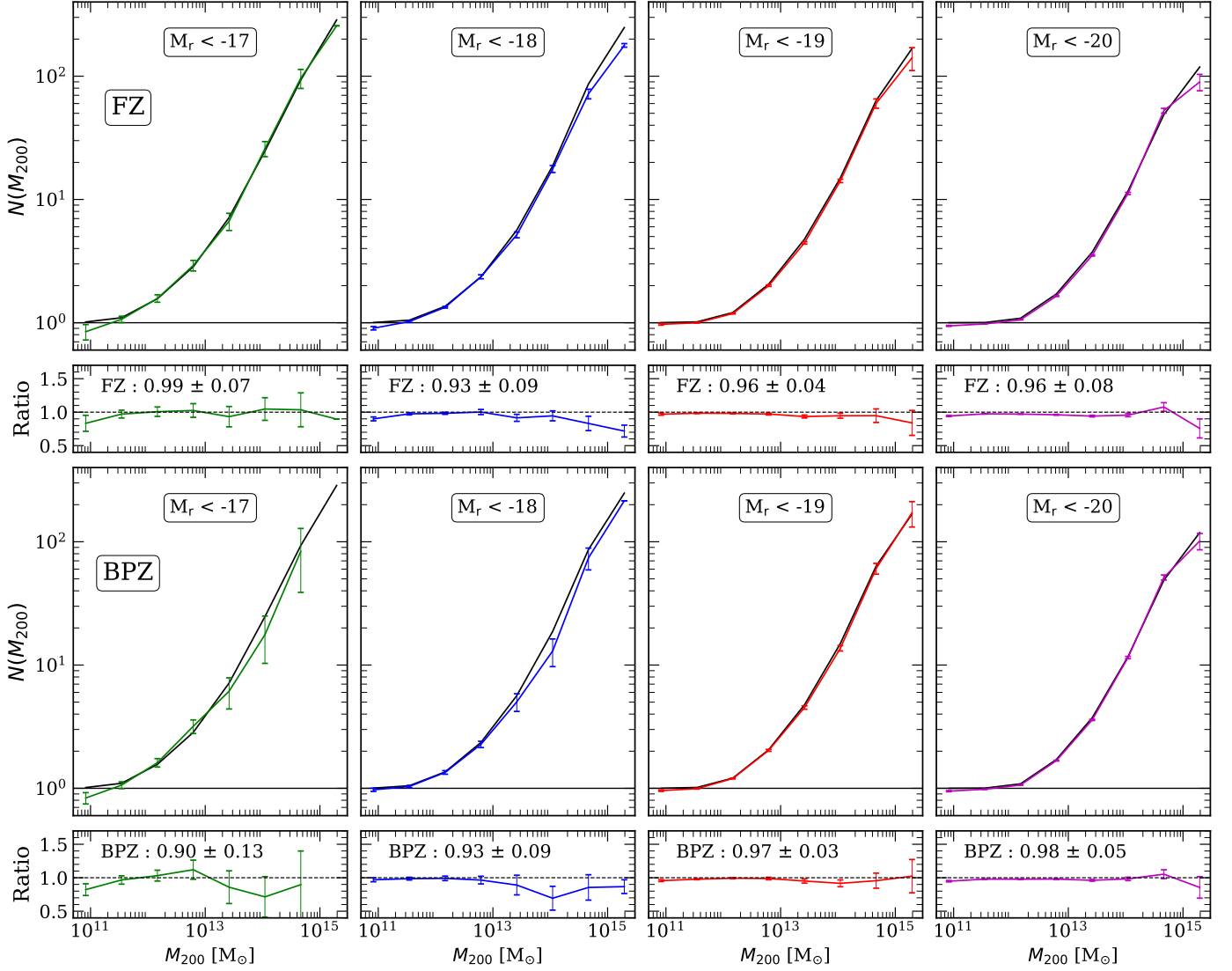


Fig. 10. Top panels: Recovered BST HODs using FlexZBoost (FZ) photometric redshifts. The black lines indicate the true HOD. Lower small panels: Ratio between the BST-recovered and true HOD, together with the corresponding median and dispersion. Bottom panels: Same as the top panels but for BPZ.

primarily through an improvement in the purity of central galaxy identification. This effect is particularly relevant at high halo masses, where contamination from misidentified centrals can strongly bias the HOD estimate. In contrast, variations in the redshift discard parameter f_z have a comparatively minor impact on the BST results, indicating that the method is less sensitive to this parameter within the explored range.

6. Conclusions

In this paper, we extended the BST developed by [Rodriguez et al. \(2015\)](#) to enable its application to galaxy samples based on both true and photometric redshifts. By incrementally introducing observational uncertainties into the variables involved in BST, we evaluated the ability of the method to recover the HOD. The main modifications and improvements to the original BST are summarised as follows:

- We first evaluated the method by applying BST to the true values of redshift, halo mass, and radii in COSMOC2 and successfully recovered the HOD with excellent agreement. Following the classical BST approach, we used the true central galaxies as group centres with known positions.
- We then replaced the characteristic radius r_c with a more observationally motivated proxy, using the luminosity of the central galaxy (r_c^{lum}), showing that luminosity-based scaling provides a sufficiently reliable proxy for group sizes within the context of our analysis.
- Next, we added uncertainty to the group positions by incorporating photometric redshift errors. Despite a significant loss in statistical power due to the limited sky coverage of the DC2 simulation, we are able to recover the HOD. When introducing typical uncertainties in the halo mass estimation, the HOD recovery remains robust.
- We substituted the known central galaxies from the simulation with an inferred central-galaxy identification method (CGF), based on galaxy luminosity and spatial distributions. We achieve a central-galaxy average

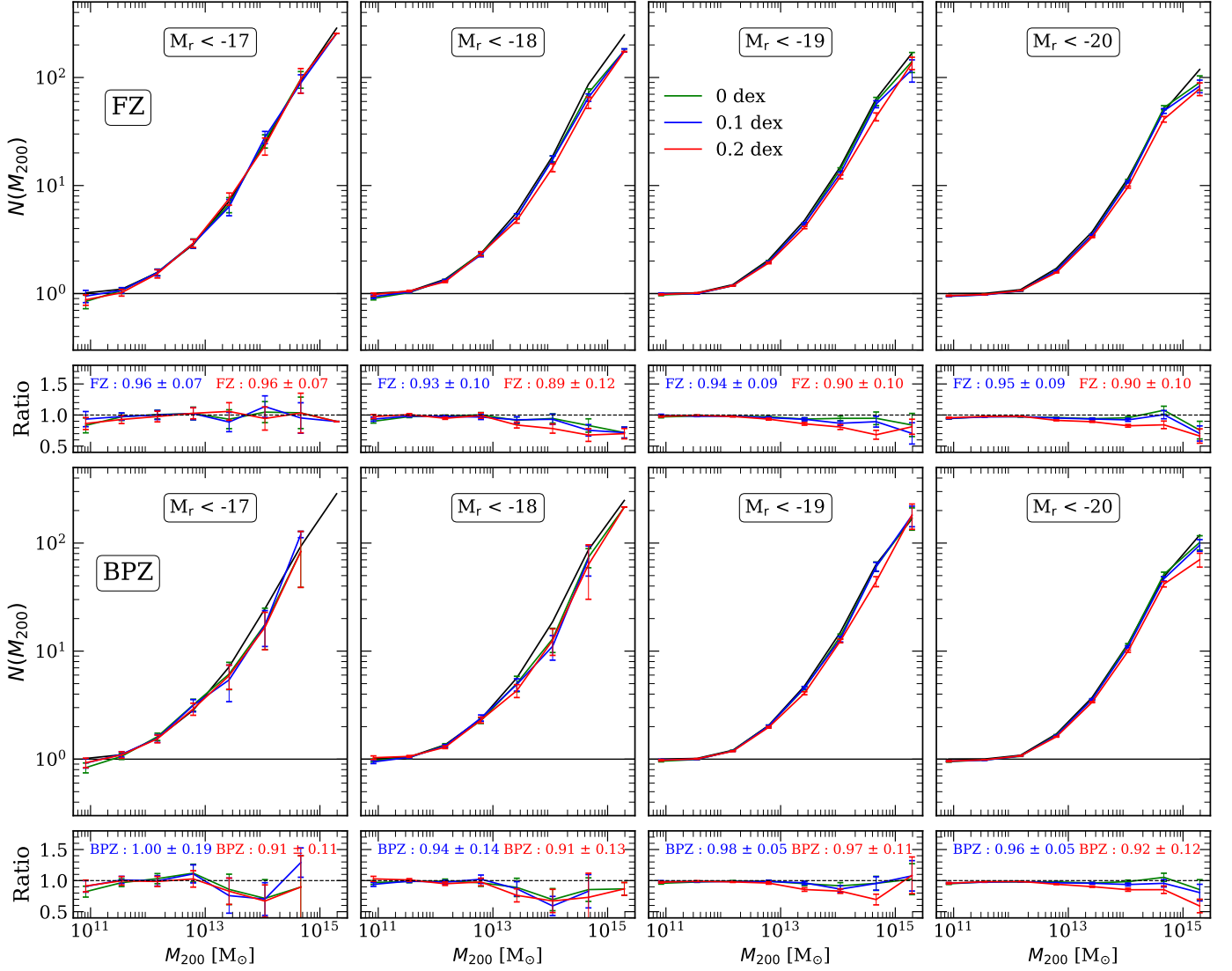


Fig. 11. Top panels: Recovered HODs using the BST under varying levels of Gaussian uncertainty added to the halo masses, expressed in dex for FlexZBoost photometric redshifts. Bottom panels: Same as in the top panels but for BPZ. Increasing the scatter predominantly affects the high-mass end of the HOD, while the low-mass regime remains largely stable. The annotations show the median ratio and dispersion for the FZ and BPZ cases, with blue corresponding to 0.1 dex scatter and red to 0.2 dex scatter.

Table 3. Summary of the BST performance for all tests.

	Radius	Group Centres	Redshift	$\langle \text{HOD}_{\text{BST}} / \text{HOD}_{\text{true}} \rangle$				Sect.
				$M_r < -17$	$M_r < -18$	$M_r < -19$	$M_r < -20$	
Simulation-based (Truth)	r_{200}	True centrals	z_{true}	1.02 ± 0.15	0.97 ± 0.07	0.97 ± 0.03	0.97 ± 0.06	Sect. 5.1
Luminosity-based	r_c^{lum}	True centrals	z_{true}	0.98 ± 0.13	0.93 ± 0.06	0.96 ± 0.03	0.98 ± 0.06	Sect. 5.2
photo-z (FlexZBoost)	r_c^{lum}	True centrals	z_{photo}	0.99 ± 0.07	0.93 ± 0.09	0.96 ± 0.04	0.96 ± 0.08	Sect. 5.3
photo-z (BPZ)	r_c^{lum}	True centrals	z_{photo}	0.90 ± 0.13	0.93 ± 0.09	0.97 ± 0.03	0.98 ± 0.05	Sect. 5.3
Error Mass (FlexZBoost)	r_c^{lum}	True centrals	z_{photo}	0.96 ± 0.07	0.89 ± 0.12	0.90 ± 0.10	0.90 ± 0.10	Sect. 5.4
Error Mass (BPZ)	r_c^{lum}	True centrals	z_{photo}	0.91 ± 0.11	0.91 ± 0.13	0.97 ± 0.11	0.92 ± 0.12	Sect. 5.4
CGF (FlexZBoost)	r_c^{lum}	CGF	z_{photo}	0.74 ± 0.25	0.86 ± 0.15	0.93 ± 0.04	0.94 ± 0.09	Sect. 5.5
CGF (BPZ)	r_c^{lum}	CGF	z_{photo}	0.76 ± 0.25	0.89 ± 0.11	0.95 ± 0.05	0.94 ± 0.09	Sect. 5.5

Notes. Summary of the BST performance across all tests conducted in this work. Each column lists the characteristic halo radius used (r_c), the method adopted to identify group centres, redshift employed for group centres, and the median ratio between the BST-recovered and true HOD values, $\langle \text{HOD}_{\text{BST}} / \text{HOD}_{\text{true}} \rangle$. We only show results for 0.2 dex uncertainty in halo mass. The final column references the corresponding section where each test is discussed.

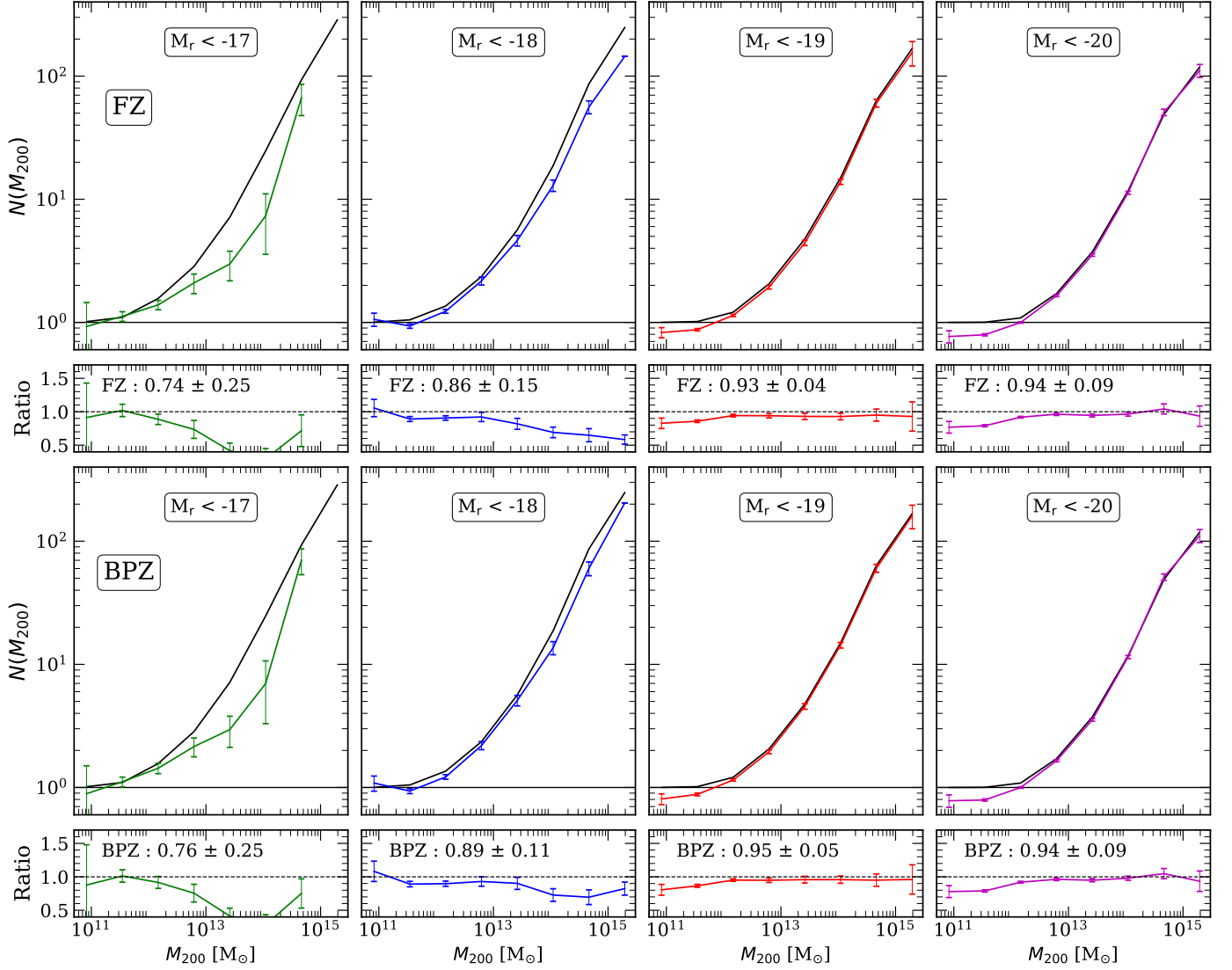


Fig. 12. Top panels: Results of the BST applied to galaxies identified with the CGF, with parameters $(f_r, f_z) = (2.5, 0.25)$ for FlexZBoost photometric redshifts. The black lines show the true HOD. Bottom panels: Same as in the top panel but for BPZ.

purity greater than 70 %, with a completeness average over 60 % over the halo mass range studied. This simple method effectively recovers the HOD for the brighter cuts, bringing our approach closer to observational techniques for constraining halo occupation.

Table 3 summarises the BST performance across all tests. We recover the HOD of the COSMOC2 catalogue in four ranges of limiting absolute magnitudes which extend from -17 to -20 in the LSST r-band and halo masses up to $10^{15} M_{\odot}$.

These results demonstrate that the BST can be effectively extended to a broad range of observational datasets, including those with mixed redshift quality and limited group information. Our updated implementation reduces model dependence by relying on observable-based proxies and inference techniques, rather than requiring full spectroscopic group catalogues. This makes the method particularly well-suited for application to deep, wide-field photometric surveys such as LSST, Euclid, and Roman, which will lack complete spectroscopic coverage but provide unprecedented galaxy statistics.

This analysis is still far from the achieving power of an LSST-like survey: the increase in total area from 440 to 18,000 deg^2 corresponds to an enhancement of the galaxy sample by a factor of ~ 41 , implying a best-case improvement of $\sim \sqrt{41}$ in the statistical constraining power of HOD-based analyses. We note, however, that the galaxy samples used in this work were drawn directly from the COSMOC2 simulation, which represents an idealised dataset unaffected by atmospheric or instrumental effects. Although this work provides a thorough rundown of the major systematics relevant to our methodology, stars and blending were not modelled in COSMOC2; nonetheless, our magnitude range ($m_r < 21$) lies well above the regime where blending significantly affects the detection of ultra-faint galaxies. We did not apply a k-correction when computing galaxy absolute magnitudes. This omission may introduce a systematic bias in the inferred luminosities and in the definition of the volume-limited sample. The required LSST-specific spectral templates are not yet available for these simulations, and implementing an alternative framework is beyond the scope of this work. Although this limitation may affect galaxies

near the luminosity threshold, we do not expect it to qualitatively alter our main conclusions. The catalogue also does not include extinction from Milky Way dust or stellar foregrounds, modelling only the reddening due to dust within the host galaxies.

Importantly, the ability to recover the HOD from more photometry-based data opens the door to a wide range of science applications beyond traditional clustering analyses. By enabling the measurement of satellite occupation as a function of secondary galaxy properties—such as colour, local density, star formation rate, and morphology—this approach can contribute to constraining assembly bias and understanding the connection between galaxy evolution and environment. Furthermore, it provides a valuable framework for investigating phenomena such as galactic conformity, where the properties of satellite galaxies correlate with those of their central galaxy. We plan to explore these aspects further with the upcoming LSST data. Importantly, the present work also helps pave the way for the development of cluster-based analyses by providing a pipeline that is readily adaptable to LSST data releases.

Acknowledgements. PC wrote most of the manuscript and created and co-designed all figures, with guidance and input from the co-authors. VC developed most of the CGF method and provided feedback on the results and the manuscript. FR supervised the lead author and contributed to conceptualisation, methodology, project administration, resources, validation, and writing. AT developed the initial data selection, advised on project progress and presentation, and edited the manuscript. MCA edited the manuscript, provided substantial comments and suggestions during the writing process, contributed to the organisation of the paper, and advised on the presentation of the results. BL and HW provided feedback as internal reviewers, leading to updates to the manuscript structure and content. This paper has undergone internal review in the LSST Dark Energy Science Collaboration. We are very grateful to the internal reviewers: Heidi Wu and Benjamin Levine. PC acknowledges partial support from AN-PCyT through grant PICT 2020-00582. MCA acknowledges support from the ANID BASAL project FB210003. VC, FR and AT thank the support by Agencia Nacional de Promoción Científica y Tecnológica, the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET, Argentina) and the Secretaría de Ciencia y Tecnología de la Universidad Nacional de Córdoba (SeCyT-UNC, Argentina). FR acknowledges support from the ICTP through the Junior Associates Programme 2023-2028. AT acknowledges support from the Dirección General de Asuntos del Personal Académico (DGAPA) of UNAM through the postdoctoral fellowship, and acknowledges support from the PAIRS project via the SONATA grant no.2023/51/D/ST9/02919. DESC acknowledges ongoing support from the IN2P3 (France), the STFC (United Kingdom), and the DOE and LSST Discovery Alliance (United States). DESC uses resources of the IN2P3 Computing Center (CC-IN2P3-Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BEIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515. We thank the developers and maintainers of the following software used in this work: NumPy (van der Walt et al. 2011), SciPy (Jones et al. 2001), Matplotlib (Hunter 2007), Astropy (Astropy Collaboration et al. 2013), and Jupyter (Kluyver et al. 2016).

References

Agüena, M., Benoist, C., da Costa, L. N., et al. 2021, *MNRAS*, 502, 4435
 Akesson, R., Armus, L., Bachelet, E., et al. 2019, arXiv e-prints, arXiv:1902.05569
 Avila-Reese, V., Colín, P., González-Samaniego, A., et al. 2011, *ApJ*, 736, 134

Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019, *MNRAS*, 488, 3143
 Benítez, N. 2011, BPZ: Bayesian Photometric Redshift Code, Astrophysics Source Code Library, record ascl:1108.011
 Benson, A. 2010, *New Astronomy*, 17
 Berlind, A. A. & Weinberg, D. H. 2002, *ApJ*, 575, 587
 Cacciato, M., van den Bosch, F. C., More, S., et al. 2009, *MNRAS*, 394, 929
 Conroy, C. 2013, *ARA&A*, 51, 393
 Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, *ApJ*, 292, 371
 Duckworth, C., Tojeiro, R., Kraljic, K., et al. 2019, *MNRAS*, 483, 172
 Gonzalez, E. J., Rodriguez, F., García Lambas, D., et al. 2016, *Monthly Notices of the Royal Astronomical Society*, stw2803
 Hearin, A., Korytov, D., Kovacs, E., et al. 2020, *MNRAS*, 495, 5040
 Heitmann, K., Finkel, H., Pope, A., et al. 2019, *ApJS*, 245, 16
 Holmberg, E. 1969, *Arkiv for Astronomi*, 5, 305
 Izbicki, R. & Lee, A. B. 2017, arXiv e-prints, arXiv:1704.08095
 Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, *ApJS*, 192, 18
 Korytov, D., Hearin, A., Kovacs, E., et al. 2019, *ApJS*, 245, 26
 Kovacs, E., Mao, Y.-Y., Agüena, M., et al. 2022, *The Open Journal of Astrophysics*, 5, 1
 Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, *ApJ*, 609, 35
 Lares, M., Lambas, D. G., & Domínguez, M. J. 2011, *AJ*, 142, 13
 Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193
 Leauthaud, A., Tinker, J., Bundy, K., et al. 2012, *ApJ*, 744, 159
 Lederman, W. 1984, *Handbook of Applicable Mathematics. Volume 6A-B: Statistics* (Wiley-Blackwell)
 Liu, L., Gerke, B. F., Wechsler, R. H., Behroozi, P. S., & Busha, M. T. 2011, *ApJ*, 733, 62
 LSST Dark Energy Science Collaboration. 2012, arXiv e-prints, arXiv:1211.0310
 LSST Dark Energy Science Collaboration (LSST DESC), Abolfathi, B., Alonso, D., et al. 2021, *ApJS*, 253, 31
 LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv e-prints, arXiv:0912.0201
 More, S., van den Bosch, F. C., Cacciato, M., et al. 2009, *MNRAS*, 392, 801
 Moster, B. P., Naab, T., & White, S. D. M. 2018, *MNRAS*, 477, 1822
 Neistein, E., Weinmann, S. M., Li, C., & Boylan-Kolchin, M. 2011, *Monthly Notices of the Royal Astronomical Society*, 414, 1405
 Payerne, C., Zhang, Z., Agüena, M., et al. 2025, *A&A*, 700, A34
 Peacock, J. A. & Smith, R. E. 2000, *MNRAS*, 318, 1144
 Phillipps, S. & Shanks, T. 1987, *MNRAS*, 229, 621
 Rodríguez, F. & Merchán, M. 2020, *A&A*, 636, A61
 Rodríguez, F., Merchán, M., & Sgró, M. A. 2015, *A&A*, 580, A86
 Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, *ApJ*, 785, 104
 Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, *MNRAS*, 499, 1587
 Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, *ApJ*, 546, 20
 Seljak, U. 2000, *MNRAS*, 318, 203
 Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757
 Springel, V., White, S. D. M., Jenkins, A., et al. 2005, *Nature*, 435, 629
 Tempel, E., Einasto, J., Einasto, M., Saar, E., & Tago, E. 2009, *A&A*, 495, 37
 Tinker, J. L., Weinberg, D. H., Zheng, Z., & Zehavi, I. 2005, *ApJ*, 631, 41
 van den Bosch, F. C., Yang, X., & Mo, H. J. 2003, *MNRAS*, 340, 771
 van den Bosch, F. C., Yang, X., Mo, H. J., et al. 2007, *MNRAS*, 376, 841
 Van Kampen, W., Cluver, M., Taylor, E. N., et al. 2026, *PASA*, 43, e014
 Wechsler, R. H., Somerville, R. S., Bullock, J. S., et al. 2001, *ApJ*, 554, 85
 White, M., Zheng, Z., Brown, M. J. I., Dey, A., & Jannuzi, B. T. 2007, *ApJ*, 655, L69
 White, S. D. & Rees, M. J. 1978, *Monthly Notices of the Royal Astronomical Society*, 183, 341
 Yang, X., Mo, H., Van Den Bosch, F. C., & Jing, Y. 2005, *Monthly Notices of the Royal Astronomical Society*, 356, 1293
 Yang, X., Mo, H. J., & van den Bosch, F. C. 2008, *ApJ*, 676, 248
 Yang, X., Xu, H., He, M., et al. 2021, *ApJ*, 909, 143
 Zandivarez, A., Díaz-Giménez, E., Taverna, A., Rodríguez, F., & Merchán, M. 2024, *A&A*, 691, A6
 Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, *ApJ*, 633, 791
 Zou, H., Gao, J., Xu, X., et al. 2021, *ApJS*, 253, 56

Appendix A: CGF

We explore the impact of varying the free parameters (f_r, f_z) in the CGF method. Fig. A.1 shows the median purity (top) and completeness (bottom) of the CGF as a function of halo mass (M_{200}) over the range 10^{12} – $10^{15} M_\odot$. The values in each cell of the table correspond to various combinations of the CGF parameters f_r and f_z , which control the projected and redshift-space exclusion regions during iterative centre selection.

As discussed in the main text, a clear trade-off emerges: higher values of f_r and f_z increase the exclusion volume, improving the purity—defined as the fraction of selected central candidates that are truly centrals—while reducing completeness, the fraction of true centrals correctly identified. In contrast, lower parameter values yield better completeness at the expense of purity, as more contaminants are retained.

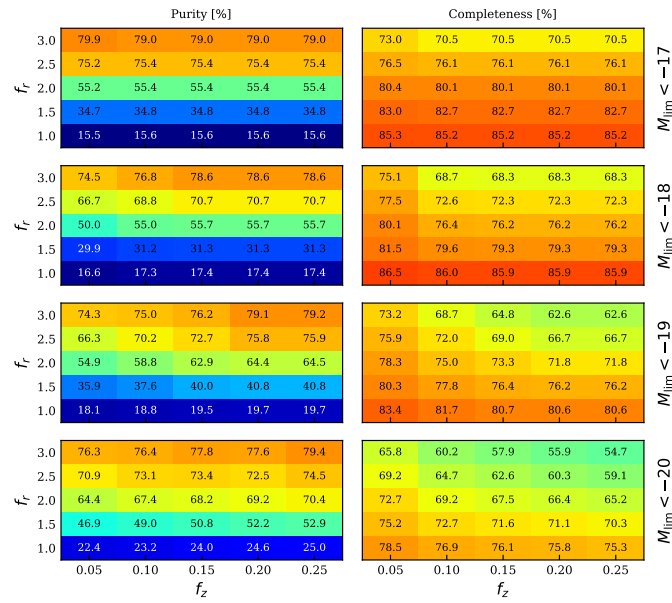


Fig. A.1. Mean purity (left panel) and completeness (right panel) of the CGF, obtained by averaging these quantities over haloes with masses in the range 10^{12} – $10^{15} M_\odot$. Each panel corresponds to a different combination of the CGF parameters f_r and f_z . Increasing f_r and f_z leads to higher mean purity but lower mean completeness, while smaller values improve completeness at the cost of greater contamination.

These results illustrate the motivation behind our adopted parameter choice of $(f_r, f_z) = (2.5, 0.25)$, which maintains purity above $\sim 70\%$ while preserving acceptable completeness, especially important in the context of large upcoming photometric surveys such as LSST. This combination of parameters also yields a distribution of identified groups that is broadly consistent with those found using the original group sample, as shown in Fig. A.2. For this comparison, we compute the Kolmogorov–Smirnov (KS, [Lederer 1984](#)) statistic (D^{KS}) over the full dataset to quantify the maximum deviation between the CGF-identified group centres and those in the original sample. The resulting p -values ($p \gg 0.05$) indicate no significant difference between the two distributions, supporting the accuracy of the CGF in recovering the underlying group-centre distribution. However, it is important to note that this comes at the

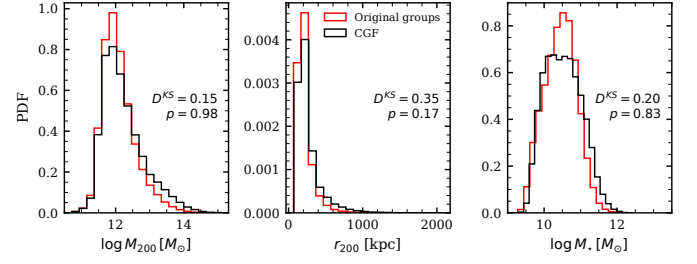


Fig. A.2. Normalized distributions of halo mass M_{200} (left panel), halo radius r_{200} (middle panel), and stellar mass M_* (right panel) of CGF vs the original groups sample. KS statistics and p -values for each comparison are shown in the panels.

cost of a substantial reduction in the number of groups included in the CGF sample, reflecting the trade-off between purity and completeness of the CGF method.

To further evaluate the impact of the CGF parameters on the accuracy of central galaxy identification, we conduct the BST, varying one parameter at a time while fixing the other to a representative pivot value ($f_r = 2.5$ or $f_z = 0.25$). The results, shown in Fig. A.3, highlight how greater discard radii (specifically $r_{\text{dsc}} < 2.5 r_c^{\text{lum}}$) lead to better agreement with the HOD. This improvement is directly related to the resulting purity of the CGF, which is essential to faithfully recover the underlying HOD signal.

Only for $f_r = 2.5$ and $f_r = 3.5$, where the purity exceeds roughly 70% , we obtain good agreement with the true HOD. This is particularly important in the high-mass regime, where the HOD is more sensitive to contamination. This analysis supports our choice of parameters for the main catalogue used in the following sections.

Appendix B: Inference of the halo mass

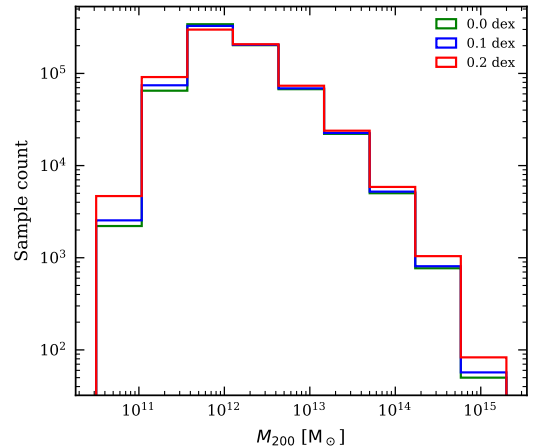


Fig. B.1. Distribution of halo masses in the group sample. The green curve shows the original distribution, while the blue and red curves represent the perturbed distributions obtained by adding Gaussian scatter in M_{200} with different standard deviations (in dex).

To assess the impact of halo mass uncertainties on the overall distribution, we introduce Gaussian perturbations with varying standard deviations (in dex) to the original M_{200} values. As shown in Fig. B.1, increasing the level of scatter

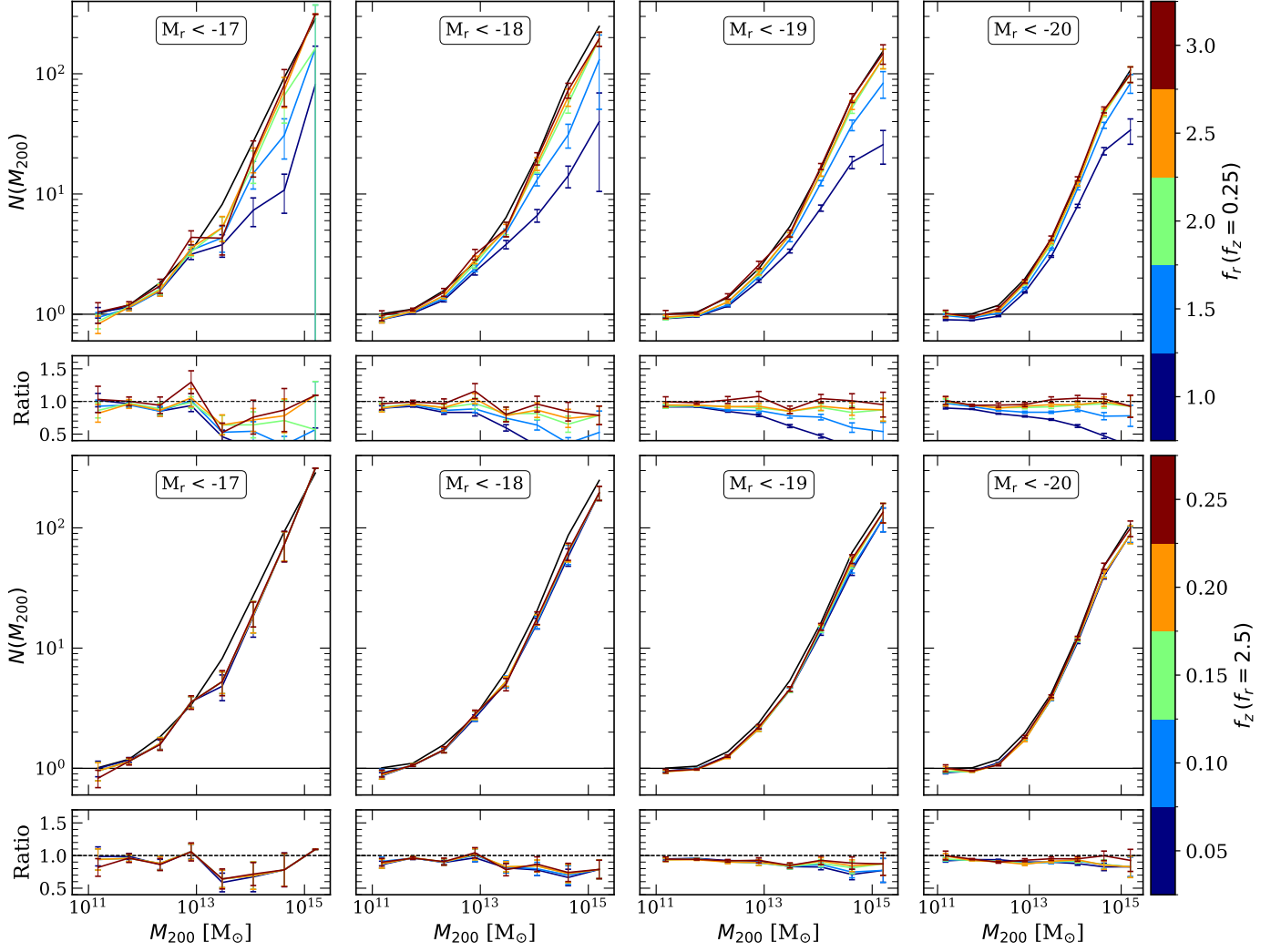


Fig. A.3. Top panels: BST results using a fixed redshift parameter $f_z = 0.25$ and varying f_r parameter. Bottom panels: Results using a fixed $f_r = 2.5$ and varying the redshift parameter f_z . The performance shows better agreement with the true HOD as f_r increases, while it does not change significantly when varying f_z .

leads to a broader redistribution of halo masses. Notably, this results in a net increase in the number of haloes at both the low- and high-mass ends, as objects are scattered into the tails of the distribution from the central mass range.