Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **452,** 686–700 (2015) doi:10.1093/mnras/stv1289

nIFTy cosmology: Galaxy/halo mock catalogue comparison project on clustering statistics

Chia-Hsun Chuang,[1★†] Cheng Zhao,[2] Francisco Prada,[1,3,4] Emiliano Munari,[5,6] Santiago Avila,[1,7] Albert Izard,[8] Francisco-Shu Kitaura,[9] Marc Manera,[10] Pierluigi Monaco,[5,6] Steven Murray,[11,12] Alexander Knebe,[7] Claudia G. Scóccola,[13,14] Gustavo Yepes,[7] Juan Garcia-Bellido,[1] Felipe A. Marín,[15,16] Volker Müller,[9] Ramin Skibba,[17] Martin Crocce,[8] Pablo Fosalba,[8] Stefan Gottlöber,[9] Anatoly A. Klypin,[18] Chris Power,[11,12] Charling Tao[2,19] and Victor Turchaninov[20]

Affiliations are listed at the end of the paper

Accepted 2015 June 8. Received 2015 May 13; in original form 2014 December 30

ABSTRACT

We present a comparison of major methodologies of fast generating mock halo or galaxy catalogues. The comparison is done for two-point (power spectrum and two-point correlation function in real and redshift space), and the three-point clustering statistics (bispectrum and three-point correlation function). The reference catalogues are drawn from the BigMultiDark *N*-body simulation. Both friend-of-friends (including distinct haloes only) and spherical overdensity (including distinct haloes and subhalos) catalogues have been used with the typical number density of a large volume galaxy surveys. We demonstrate that a proper biasing model is essential for reproducing the power spectrum at quasi-linear and even smaller scales. With respect to various clustering statistics, a methodology based on perturbation theory and a realistic biasing model leads to very good agreement with *N*-body simulations. However, for the quadrupole of the correlation function or the power spectrum, only the method based on semi-*N*-body simulation could reach high accuracy (1 per cent level) at small scales, i.e. $r < 25\,h^{-1}$ Mpc or $k > 0.15\,h\,\mathrm{Mpc}^{-1}$. Full *N*-body solutions will remain indispensable to produce reference catalogues. Nevertheless, we have demonstrated that the more efficient approximate solvers can reach a few per cent accuracy in terms of clustering statistics at the scales interesting for the large-scale structure analysis. This makes them useful for massive production aimed at covariance studies, to scan large parameter spaces, and to estimate uncertainties in data analysis techniques, such as baryon acoustic oscillation reconstruction, redshift distortion measurements, etc.

Key words: cosmology: observations – distance scale – large-scale structure of Universe.

1 INTRODUCTION

The scope of galaxy redshift surveys has dramatically increased in the last years. The 2dF Galaxy Redshift Survey[1] (2dFGRS) obtained 221 414 galaxy redshifts at $z < 0.3$ (Colless et al. 2001, 2003), and the Sloan Digital Sky Survey[2] (SDSS; York et al. 2000) collected 930 000 galaxy spectra in the Seventh Data Release at $z < 0.5$ (Abazajian et al. 2009). WiggleZ[3] collected spectra of 240 000 emission-line galaxies at $0.5 < z < 1$ over 1000 deg^2 (Drinkwater et al. 2010; Parkinson et al. 2012), and the Baryon Oscillation Spectroscopic Survey[4] (BOSS; Dawson et al. 2013) of the SDSS-III project (Eisenstein et al. 2011) has surveyed 1.5 million luminous red galaxies at $0.1 < z < 0.7$ over 10 000 deg^2. There are new upcoming ground-based and space experiments, such as 4MOST[5] (4-metre Multi-Object Spectroscopic Telescope; de Jong et al. 2012),

★E-mail: chuang@nhn.ou.edu
†MultiDark Fellow.
[1] http://www2.aao.gov.au/2dfgrs/
[2] http://www.sdss.org
[3] http://wigglez.swin.edu.au/site/
[4] https://www.sdss3.org/surveys/boss.php
[5] http://www.4most.eu/

Table 1. The methodologies of generating mock halo/galaxy catalogues developed in the last years. The methodologies included in this study are highlighted using bold font.

Methodology	Reference
Log-Normal	Coles & Jones (1991)
PTHALOS	Manera et al. (2012, 2015)
PINOCCHIO (PINpointing Orbit-Crossing Collapsed Hierarchical Objects)	Monaco, Theuns & Taffoni (2002), Monaco et al. 2013
COLA (COmoving Lagrangian Acceleration simulation)	Tassev, Zaldarriaga & Eisenstein (2013)
PATCHY (PerturbAtion Theory Catalog generator of Halo and galaxY distributions)	Kitaura et al. (2014, 2015)
QPM (quick particle mesh)	White, Tinker & McBride (2014)
EZMOCK (Effective Zel'dovich approximation mock catalogue)	Chuang et al. (2015)
HXALOGEN	Avila et al. (2015)

DES[6] (Dark Energy Survey; Frieman & Dark Energy Survey Collaboration 2013), DESI[7] (Dark Energy Spectroscopic Instrument; Schlegel et al. 2011; Levi et al. 2013), eBOSS[8] (Extended Baryon Oscillation Spectroscopic Survey), HETDEX[9] (Hobby-Eberly Telescope Dark Energy Experiment; Hill et al. 2008), J-PAS[10] (Javalambre Physics of accelerating universe Astrophysical Survey; Benitez et al. 2015), LSST[11] (Large Synoptic Survey Telescope; Abell et al. 2009), *Euclid*[12] (Laureijs et al. 2011), and *WFIRST*[13] (*Wide-Field Infrared Survey Telescope*; Green et al. 2012), which would observe even larger galaxy samples.

Mock galaxy catalogues are essential for analysing the clustering signal drawn from these surveys. Tight constraints on cosmological models can be determined provided that the covariances of the clustering measurements are reliably estimated. For such purpose, we need a large number of realizations of a simulation designed to reproduce the volume of the Universe observed in a given survey. *N*-body simulations are an ideal tool for reproducing cosmological structures, e.g. LasDamas[14] (Large Suite of Dark Matter Simulations), which has been used to analyse the SDSS-II galaxy sample (e.g. Chuang, Wang & Hemantha 2012; Samushia, Percival & Raccanelli 2012), although running many realizations is expensive, or even unfeasible if such number has to be very large (e.g. we might need $\sim 10^3$ or even more.). In order to circumvent this problem, some alternatives have been proposed. In the last decades, many new tools (see Table 1) have been developed for reconstructing in an approximate way the large-scale structures down to the mildly non-linear scales, allowing a fast generation of simulated volumes of the Universe. In this way, a direct computation of the covariance matrices by means of large numbers of realizations is possible.

In this paper, we compare these different methods, including COLA, EZMOCK, HALOGEN, Log-Normal, PATCHY, PINOCCHIO, and PTHALOS. We generate the halo mock catalogues using the same initial power spectrum (except lognormal model since it uses the observed correlation function as the input) and compare with the *N*-body simulation which also used the same initial power spectrum. This comparison is meant to investigate the performances of the different methods for computing the clustering properties (power spectrum, correlation function, bispectrum and three-point correlation function) in real and redshift space, leading to considerations on the capabilities of recovering the properties of the baryonic acoustic oscillations (BAO) and redshift space distortion. We do not include the comparison of the positions of individual haloes which can be provided by COLA, PINOCCHIO, and PTHALOS. The other methods, i.e. EZMOCK, HALOGEN, Log-Normal, and PATCHY, generate haloes with some biasing models calibrated with the *N*-body simulations.

This paper – emerging out of the 'nIFTy cosmology' workshop[15] – is organized as follows. In Section 2, we describe the reference *N*-body simulation catalogues used for our study. In Section 3, we present a quick description of the main characteristics of the different codes used in this comparison work, highlighting their similarities and the differences. The results are presented in Section 4, first for the main haloes and then also including the presence of substructures. We discuss the results of the previous section, and finally conclude in Section 5.

2 REFERENCE *N*-BODY HALO CATALOGUES

To test the different methods, we use a reference halo catalogue at redshift $z = 0.5618$ extracted from the BigMultiDark (BigMD) simulation[16] (Klypin et al. 2014), which was performed using GADGET-2 Springel 2005 with 3840^3 particles on a volume of $(2500\ h^{-1}\,\mathrm{Mpc})^3$ assuming ΛCDM Planck cosmology with $\{\Omega_\mathrm{M} = 0.307115, \Omega_\mathrm{b} = 0.048206, \sigma_8 = 0.8288, n_\mathrm{s} = 0.96\}$, and a Hubble constant ($H_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$) given by $h = 0.6777$. Within the MultiDark project, a series of dark matter (DM) only simulations in different cosmologies and with different box sizes and resolutions have been performed (see Klypin et al. 2014 for an overview). The MultiDark simulations have been used already to interpret the clustering of the BOSS galaxy sample (Nuza et al. 2013).

Haloes were defined based on two different algorithms. A friends-of-friends based code (called FOF; e.g. see Riebe et al. 2011) and a spherical overdensity (SO) based code (called BDM; e.g. see Klypin & Holtzman 1997; Riebe et al. 2011). The former code does not ab initio give subhaloes whereas the latter does, and haloes that are not subhaloes are also referred to as 'distinct haloes'. Note that we use 'BDM haloes' and 'SO haloes' interchangeably. In this work, we use the FOF catalogue (linking length = 0.2) as our reference to compare between the different approximate methods; and also use the SO catalogues (obtained with BDM code) to discuss the effect of substructures. From the halo catalogue, we select a complete sample, selected by mass, with number density $3.5 \times 10^{-4}\ h^3\,\mathrm{Mpc}^{-3}$, which is similar to that of the BOSS galaxy sample at $z \sim 0.5$. This

[6] http://www.darkenergysurvey.org
[7] http://desi.lbl.gov/
[8] http://www.sdss.org/sdss-surveys/eboss/
[9] http://hetdex.org
[10] http://j-pas.org
[11] http://www.lsst.org/lsst/
[12] http://www.euclid-ec.org
[13] http://wfirst.gsfc.nasa.gov
[14] http://lss.phy.vanderbilt.edu/lasdamas/
[15] http://popia.ft.uam.es/nIFTyCosmology
[16] http://www.multidark.org/

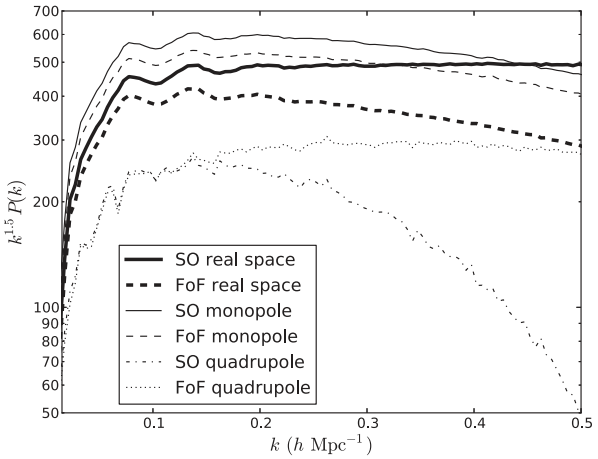


Figure 1. Clustering statistics in real and redshift space for the reference BigMD SO and FOF catalogues, both with the same number density. Monopole of the power spectrum in real space for BigMD SO catalogue (thick solid line) and FOF catalogue (thick dashed line); monopole of power spectrum in redshift space for SO (thin solid line) and FOF (thin-dashed line); and quadrupole of power spectrum in redshift space for SO (dash–dotted line) and FOF (dotted line). In real space, SO monopole has higher amplitude due the clustering signal of subhaloes at small scales; but, in redshift space, the signal in the monopole is compensated out by the local motions. On the other hand, the quadrupole of the SO catalogue has much less signal due to the local motions.

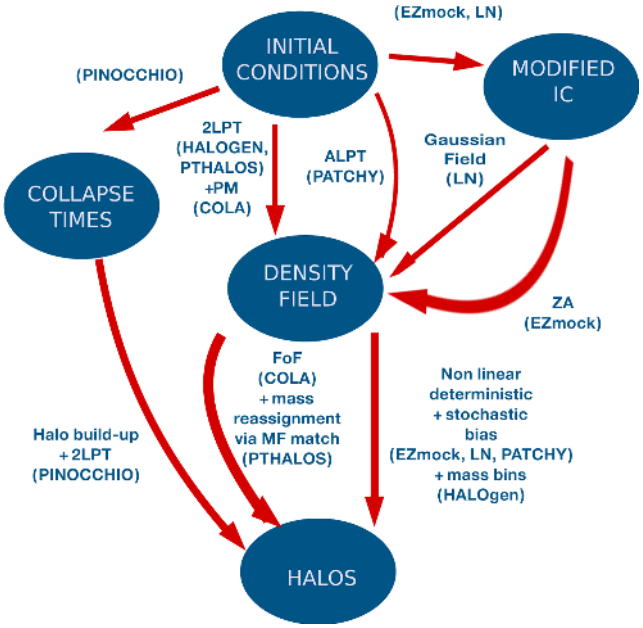


Figure 2. A summary plot of different methodologies to generate mock halo catalogue. See the context for detail description.

abundance is equivalent to a mass cut of $\sim 1 \times 10^{13}\ \mathrm{M}_{\odot}\,h^{-1}$ for the FOF catalogue and $\sim 8.5 \times 10^{12}\ \mathrm{M}_{\odot}\,h^{-1}$ for the SO catalogue. Note that the BigMD simulation is designed to have the proper box size and mass resolution for constructing the mock galaxy catalogues for the BOSS survey which has collected the largest spectroscopic galaxy sample to date. While it would be interesting to go past these limits both in box size and mass resolution, we nevertheless leave this for future studies.

Fig. 1 displays the impact of substructures on the large-scale clustering statistics. Specifically, we want to show how the power spectrum at wavenumbers $k \lesssim 1\,h\,\mathrm{Mpc}^{-1}$ is affected by the one-halo term of the correlation function. Naively, one does not expect that there is such an effect. After all, why should clustering at $\lambda > 2\pi/k \sim 6\,h^{-1}\,\mathrm{Mpc}$ be affected by inclusion of subhaloes at much smaller scales? However, there are two effects. The first one is rather simple. There are more subhaloes of a given mass (or circular velocity) in each massive distinct halo as compared with less massive halo. When subhaloes are included, larger haloes give proportionally larger contribution to the estimate of the power spectrum. Because larger haloes are more biased, the power spectrum (and the correlation function) is larger on all scales (see Fig. 1). In practice, this effect results in an almost scale-independent bias. The second effect is more subtle: there is a change – a boost due to subhaloes – in the power spectrum even when there is no change in the large-scale correlation function. This happens because the power spectrum and the correlation function are connected through an integral relation. This effect results in a scale-dependent bias and its effect gets progressively small for small wavenumbers k. In redshift space, this effect on the monopole is compensated due to the peculiar velocities, which yield to much smaller differences between both BigMD catalogues: SO, including substructures, and FOF, which only contains distinct haloes (see Fig. 1). On the other hand, the quadrupole of the SO catalogue has much less signal due to those peculiar velocities.

3 APPROXIMATE METHODS FOR MOCK COMPARISON

The methods used for this comparison project start from a set of initial conditions (ICs hereafter) with the aim of generating catalogues of DM haloes. The way the different methods reach this goal can be divided into three logical branches, as sketched in Fig. 2. PINOCCHIO reaches the first step by predicting the collapse times of the particles from the ICs. The others instead construct the density field before the identification or population of the haloes. While most of them compute the density field directly from the ICs, EZMOCK and Log-Normal perform a modification of the ICs (see Chuang et al. 2015 and Coles & Jones 1991 for more details).

In Table 2, we compare the main technical features of the methods. Below, we summarize the main ideas and ingredients behind each method. For a detailed description of the methods, we refer the reader to the cited papers.

3.1 COLA

COLA (COmoving Lagrangian Acceleration; Tassev et al. 2013) is a method to produce cheaper N-body simulations for large-scale structure. It uses a particle-mesh (PM) code with few timesteps to solve for the residual displacements of particles with respect to their trajectories calculated in lagrangian perturbation theory (LPT). Large-scale dynamics is exactly solved since the second-order LPT (2LPT) evolution allows us to recover the correct growth factor of fluctuations at such scales. At small scales, the accuracy is controlled by the number of timesteps (in Tassev et al. 2013, they propose 10 timesteps starting at redshift 9).

The key point of this method is how the equation of motion is rewritten. The displacement field is decomposed in two terms, one describing the 2LPT trajectory and another one for the residual displacement:

$$x_{\mathrm{res}} \equiv x - x_{\mathrm{LPT}}, \qquad (1)$$

Table 2. Main technical features of the methodologies: COLA, PINOCCHIO, and PTHALOS resolve haloes with some halo finders which can also provide the estimation of halo mass. EZMOCK and PATCHY provide halo catalogues with mass by applying a post-processing procedure (see Zhao et al. 2015). The post-processing can be used to assign mass and other mass related quantities, e.g. circular velocty. HALOGEN constructs halo catalogues in mass bins; different IC codes are used to construct the DM density field for different methodologies; all the codes are using parallelization techniques to speed up the computation. The methods using halo finders do not use bias models; EZMOCK, Log-Normal, and PATCHY construct the catalogues with substructures without post-processing; PINOCCHIO provides the merger histories. We also list the number of parameters used in each method.

	COLA	EZMOCK	HALOGEN	Log-Normal	PATCHY	PINOCCHIO	PTHALOS
Mass, Vel	M + V	M(post-process) + V	M(binned) + V	–	M(post-process) + V	M+V	M +V
Need of resolving the haloes	YES	NO	NO	NO	NO	YES	YES
ICs	2LPTic	ZA	2LPTic	Gaussian	ALPT	N-GenIC; can read in graphic2	2LPTic
Parallel	MPI + openMP	openMP	openMP	openMP	openMP	MPI + openMP	MPI
Assumed MF	NO	YES	YES	–	YES	NO	YES
Assumed bias model	NO	YES	YES	NO	YES	NO	NO
Substructures	Post-process	YES	Post-process	Yes	YES	Post-process	Post-process
Merger histories	NO	NO	NO	NO	NO	YES	NO
No. of free params	0	7	1 (each mass bin)	–	7	5	1
No. free params for z-space dist.	0	1	1	–	2	0	0
No. free params for MF	0	–	Adopt MF	–	–	5	Adopt MF
No. free params for bias	0	6	1	–	5	0	0

so that the equation of motion schematically reads

$$\partial_t^2 x_{\rm res} = -\nabla\Phi - \partial_t^2 x_{\rm LPT}. \quad (2)$$

COLA discretizes the time derivatives only on the left-hand side, while uses the LPT expression at the right-hand side.

In Tassev et al. (2013), they developed a serial code for the demonstration of the method. Afterwards, J. Koda parallelized it and made it suitable for running large ensembles of simulations, as done in Kazin et al. (2014). For an optimized and parallel version of COLA, including lightcone outputs, see Izard et al. (in preparation).

3.2 EZMOCK

EZMOCK (Effective Zel'dovich approximation mock catalogue; Chuang et al. 2015) is constructed from the Zel'dovich approximation (ZA) density field. It absorbs the non-linear effect and halo bias (i.e. linear, non-linear, deterministic, and stochastic bias) into some effective modelling with few parameters, which can be efficiently calibrated with *N*-body simulations. The following required steps are recursively applied until convergence:

(I) generation of the DM density field on a grid using the ZA;
(II) mapping the probability distribution function (PDF) of haloes measured in BigMD to the ZA density field;
(III) adding scatter to the PDF mapping scheme by

$$\rho_s(\boldsymbol{r}) = \begin{cases} \rho_o(\boldsymbol{r})(1+G(\lambda)) & \text{if } G(\lambda) \geq 0; \\ \rho_o(\boldsymbol{r})\exp(G(\lambda)) & \text{if } G(\lambda) < 0, \end{cases} \quad (3)$$

where $\rho_s(\boldsymbol{r})$ and $\rho_o(\boldsymbol{r})$ are the ZA density field after and before the scattering respectively. $G(\lambda)$ is a random number drawn from the Gaussian distribution with width λ. The exponential function is used to avoid the negative density;
(IV) fitting the amplitude of the power spectrum and bispectrum with a density threshold and saturation before the scattering scheme by

$$\rho_{o'}(\boldsymbol{r}) = \begin{cases} 0, & \text{if } \rho_o(\boldsymbol{r}) < \rho_{\rm th}^{\rm low}; \\ \rho_{\rm th}^{\rm high}, & \text{if } \rho_o(\boldsymbol{r}) > \rho_{\rm th}^{\rm high}, \end{cases} \quad (4)$$

where $\rho_{o'}(\boldsymbol{r})$ is the modified density, $\rho_o(\boldsymbol{r})$ is the original ZA density, $\rho_{\rm th}^{\rm low}$ and $\rho_{\rm th}^{\rm high}$ are the density threshold and density saturation respectively;
(V) fitting the shape of the final power spectrum by modifying the tilt in the initial input power spectrum with a scale-dependent function by

$$P_{\rm ePK}(k) = P_{\rm eBAO}(k)(1 + Ak), \quad (5)$$

where A is a free parameter;
(VI) fitting BAOs by enhancing the amplitude of BAOs in the initial input power spectrum by

$$P_{\rm eBAO}(k) = (P_{\rm lin}(k) - P_{\rm nw}(k))\exp(k^2/k_*^2) + P_{\rm nw}(k), \quad (6)$$

where $P_{\rm eBAO}(k)$ is the BAO enhanced power spectrum, $P_{\rm lin}(k)$ is the linear power spectrum, $P_{\rm nw}(k)$ is the smoothed no-wiggle power spectrum obtained by applying a cubic spline fit to $P_{\rm lin}(k)$, and k_* is usually known as the damping factor (however, for the damping model, one should use $\exp(-k^2/k_*^2)$ instead);
(VII) computing the peculiar motions v within the ZA for each object by adding to the linear coherent motion, which is proportional to the ZA displacement field, a dispersion term modelled by a random Gaussian distribution, i.e.

$$v_i(\boldsymbol{r}) = B\psi_i(\boldsymbol{r}) + G(\lambda'), \quad (7)$$

where B is a constant corresponding to linear growth; ψ is the displacement field, i denotes the direction x, y, or z; and $G(\lambda)$ is a random number drawn from the Gaussian distribution with width λ'.

3.3 HALOGEN

The aim of HALOGEN (Avila et al. 2015) is to provide a simple and efficient approximate method for generating mock halo catalogues with correct mass-dependent two-point statistics. The basic algorithm is as follows.

(I) Create a cosmological matter field, sampled by N particles using 2LPT.
(II) Sample a number of halo masses corresponding to the desired number density from an appropriate analytical mass function (or reference *N*-body simulation).
(III) Reconstruct the density field on a regular grid of cell size $l_{\rm cell} \approx 2d_{\rm part}$ (twice the mean-interparticle distance, for this comparison we used $l_{\rm cell} = 4\,h^{-1}$ Mpc).

(IV) Distribute haloes into mass bins (for this comparison we use eight bins), and for each bin M_j from highest to lowest mass, place each halo in the following way.

(i) Choose a cell with probability $P(i|M_j) \propto \rho_i^{\alpha(M_j)}$.

(ii) Place the halo on a random 2LPT particle within the cell.

(iii) Ensure that the halo does not overlap previous halo centres (if so, repeat the cell choice).

(iv) Decrease the mass of the cell by the mass of the halo (ensuring mass conservation on scales of $l_{\rm cell}$).

(V) Assign particle velocities to haloes with a factor $\boldsymbol{v}_{\rm h} = f_{\rm vel} \cdot \boldsymbol{v}_{\rm p}$, computed as the ratio of the velocity dispersions of the selected particles to the reference halo catalogue: $f_{\rm vel} = \frac{\sigma(\boldsymbol{v}_{\rm p})}{\sigma(\boldsymbol{v}_{\rm ref})}$

The only free parameter of the placement is $\alpha(M)$, which primarily controls the linear halo bias. It can be fitted once for a given cosmology, redshift and $l_{\rm cell}$, and used for any number of random ICs. An additional parameter controls the velocity bias, and is simply calculated via the ratio of the variance of the N-body velocities to the 2LPT particle velocities. The efficiency of HALOGEN is primarily constrained by the 2LPT step, as the algorithms intrinsic to HALOGEN are very fast.

3.4 Lognormal

The distribution of galaxies on intermediate to large scales ($>10\,h^{-1}$ Mpc) has been found to follow a lognormal distribution (see Hubble 1934; Wild et al. 2005) especially when correcting for shot noise effects (see Kitaura et al. 2009). The physical argument for this behaviour has been found in the continuity equation, as the comoving solution of the evolved density field is related to the linear density field through a logarithmic transformation when shell-crossing is neglected (see Coles & Jones 1991; Kitaura & Angulo 2012). This implies that under the assumption of Gaussian primordial fluctuations the evolved density field is expected to be lognormal distributed on intermediate to large scales. It has the advantage that its two-point statistics can be exactly controlled. Therefore, it has been widely used to study cosmic variance (and covariance matrices) in large-scale structure measurements (e.g. Cole et al. 2005; Percival et al. 2010; Reid et al. 2010; Blake et al. 2011; Beutler et al. 2011). The Log-Normal mock is constructed with the following steps.

(I) Given an input correlation function, $\xi(r)$, the Gaussian field correlation function is obtained by

$$\xi_{\rm G}(r) = \ln[1 + \xi(r)], \tag{8}$$

and this can be Fourier transformed to the power spectrum, $P_{\rm G}(k)$.

(II) A Gaussian density field $\delta_{\rm G}(r)$ is generated on the grid with the power spectrum, $P_{\rm G}(k)$,

(III) A lognormal field is calculated by

$$\delta_{\rm LN}(r) = \exp\left[\delta_{\rm G}(r) - \frac{\sigma_{\rm G}^2}{2}\right] - 1, \tag{9}$$

where $1 + \delta_{\rm LN}(r)$ is the lognormal density field which is always positive by definition and $\sigma_{\rm G}^2$ is the variance of the Gaussian density field which can be calculated by

$$\sigma_{\rm G}^2 = \sum_{i,j,l=1}^{N_{\rm grid}} P_{\rm G}\left[\left(k_{x_i}^2 + k_{y_j}^2 + k_{z_l}^2\right)^{1/2}\right], \tag{10}$$

where $N_{\rm grid}$ is the number of grid points, $k_{m_n} = \frac{2\pi}{L}\left(n - \frac{N_{\rm grid}}{2}\right)$, L is the box length, and $m = x$, y, or z.

(IV) Draw the Poisson random variables with the means given by this lognormal field.

In principle, one could assign the velocity to the Log-Normal mocks (e.g. see White et al. 2014), but it is not done in this study.

3.5 PATCHY

PATCHY (Kitaura, Yepes & Prada 2014) relies on modelling the large-scale structure density field with an efficient approximate gravity solver and populating the density field following a non-linear, scale-dependent, and stochastic biasing description. Below, the main ingredients are listed.

(I) A one-step gravity solver based on augmented Lagrangian perturbation theory (ALPT; Kitaura & Hess 2013), correcting 2LPT in the high- and low-density regimes with a non-linear local term derived from the spherical collapse (SC) model matching N-body simulations. In this approximation, the displacement field $\boldsymbol{\Psi}_{\rm ALPT}(\boldsymbol{q}, z)$, mapping a distribution of DM particles at initial Lagrangian positions q to the final Eulerian positions $\boldsymbol{x}(z)$ at redshift z ($\boldsymbol{x}(z) = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q}, z)$), is split into a long-range and a short-range component, given by 2LPT and SC, respectively:

$$\boldsymbol{\Psi}_{\rm ALPT}(\boldsymbol{q}, z) = \mathcal{K}(\boldsymbol{q}, r_{\rm S}) \circ \boldsymbol{\Psi}_{\rm 2LPT}(\boldsymbol{q}, z) + (1 - \mathcal{K}(\boldsymbol{q}, r_{\rm S})) \circ \boldsymbol{\Psi}_{\rm SC}(\boldsymbol{q}, z). \tag{11}$$

(II) A deterministic bias model relating the expected number density of haloes $\rho_{\rm h}$ to the DM density field $\rho_{\rm M}$ including a thresholding $\rho_{\rm th}$ and (or) an exponential cut-off $\exp\left[-\left(\frac{\rho_{\rm M}}{\rho_\epsilon}\right)^\epsilon\right]$, a power-law density relation $\rho_{\rm M}^\alpha$:

$$\rho_{\rm h} = f_{\rm h}\,\theta(\rho_{\rm M} - \rho_{\rm th})\,\rho_{\rm M}^\alpha\,\exp\left[-\left(\frac{\rho_{\rm M}}{\rho_\epsilon}\right)^\epsilon\right], \tag{12}$$

with

$$f_{\rm h} = \bar{N}_{\rm h}/\langle\theta(\rho_{\rm M} - \rho_{\rm th})\,\rho_{\rm M}^\alpha\,\exp\left[-\left(\frac{\rho_{\rm M}}{\rho_\epsilon}\right)^\epsilon\right]\rangle, \tag{13}$$

and $\{\rho_{\rm th}, \alpha, \epsilon, \rho_\epsilon\}$ the parameters of the model.

(III) A sampling step, which deviates from Poissonity modelling overdispersion and stochasticity in the bias relation, in particular using the negative binomial distribution function:

$$P(N_i \mid \rho_{hi}, \beta) = \frac{\lambda_i^{N_i}}{N_i!}\frac{\Gamma(\beta + N_i)}{\Gamma(\beta)(\beta + \rho_{\rm h})^{N_i}}\frac{1}{(1 + \rho_{\rm h}/\beta)^\beta} \tag{14}$$

with β being the stochastic bias parameter.

(IV) The parameters are constrained to efficiently match the halo (or galaxy) PDF and the power spectrum for a given number density. In this way, we can match the three-point statistics.

(V) Peculiar velocities are split into a coherent and a quasi-virialized component. The coherent flow is obtained from ALPT and the dispersion term is sampled from a Gaussian distribution assuming a power-law relation with the local density.

3.6 PINOCCHIO

PINOCCHIO[17] (Monaco et al. 2002, 2013) is based on the ellipsoidal collapse, solved with the aid of third-order LPT, to compute the time at which mass elements collapse (in the orbit-crossing sense), and Extended Press & Schecther to deal with multiple smoothing radii.

[17] http://adlibitum.oats.inaf.it/monaco/Homepage/Pinocchio/index.html

Table 3. This table lists the particle mesh sizes adopted by the different approximate methods presented in this comparison project; whether the reduced white noise is used, and the computational requirements including CPU hours and memory used for the mocks provided in the study. Although using the BigMD white noise is not required for mock generation, it will have an effect on the performances at large scales. Note that the computational requirements might depend on the machines used which could be a factor of 2 or even more.

	BigMD	COLA	EZMOCK	HALOGEN	Log-Normal	PATCHY	PINOCCHIO	PTHALOS
Particle mesh size	3840^3	1280^3 (3840^3 for force)	960^3	1280^3	1280^3	960^3	1920^3	1280^3
Using white noise	YES	NO	YES	YES	NO	YES	YES	NO
CPU hour	800,000	130	1.3	6.7	0.5	8	440	45
Memory	8Tb	550Gb	28Gb	130Gb	15Gb	24Gb	890Gb	112Gb

It starts from the generation of a linear density field on a regular grid in Lagrangian space, in the same way as ICs are generated for an *N*-body simulation. The density field is smoothed on a set of scales, and the collapse time is computed for each particle and at each smoothing radius. The earliest time is recorded as the bona-fide estimate of collapse time.

The collapsed medium is then fragmented into disjoint haloes by applying an algorithm that mimics the hierarchical formation and merging of haloes. This works as follows: particles are sorted in order of increasing collapse times. When a particle collapses, the fate of its six Lagrangian neighbours is checked. If all neighbours have not collapsed, then a new group with one particle is formed. If one neighbour already belongs to a group, then the particle and the group are displaced from the Lagrangian to the Eulerian space using Zel'dovich or 2LPT displacements computed at the same time of collapse of the particle. If the particle gets within the 'virial radius' of the group, then it is accreted to the group, otherwise it is tagged as a 'filament' particle. Filaments are later accreted on a group each time a neighbouring particle is accreted on the same group. If the Lagrangian neighbours of the collapsing particle belong to more groups, then the groups are displaced to check whether the centre of mass of one group gets within the 'virial radius' of the other. If this takes place, the two groups are merged. The estimate of the 'virial radius' implies the use of parameters, as fully explained in Monaco et al. (2002). These parameters are chosen requiring to reproduce a given (universal) mass function. Their values are independent of redshift, mass resolution, and cosmology, so once they are fixed the code can be applied to any configuration.

Because of the algorithm used to create haloes, PINOCCHIO can also generate accurate merger histories of haloes with continuous time sampling.

In this paper, we use a new version of the code, with 2LPT displacements and a better tuning of the mass function, that will be presented in a forthcoming paper. To compute the covariance of two-point correlation function for the VIPERS survey (de la Torre et al. 2013) used, a limited set of lightcones drawn from one of the MultiDark simulations and 200 mocks constructed with the PINOCCHIO code described above, using the Shrinkage technique (Pope & Szapudi 2008) to deal with the bias introduced by the approximate code.

3.7 PTHALOS

The basic steps in this method, inspired by Scoccimarro & Sheth 2002, can be summarized as follows (Manera et al. 2012, 2015).

(I) Create a DM particle field-based 2LPT.

(II) Identify haloes using an FOF (Davis et al. 1985) halo finder with an appropriately chosen linking length. Alternatively, one can identify haloes with SO with an equivalent density threshold.

(III) The haloes can be later populated with galaxies.

Because the 2LPT dynamics is an approximation to the true dynamics of the DM field, it yields halo densities that consistently differ from the *N*-body densities. Consequently, the FOF linking length of the 2LPT matter field, $b_{\rm 2LPT}$, needs to be rescaled from the value used in *N*-body simulations, $b_{\rm sim}$. The rescaling is given by

$$b_{\rm 2LPT} = b_{\rm sim}\left(\frac{\Delta_{\rm vir}^{\rm sim}}{\Delta_{\rm vir}^{\rm 2LPT}}\right)^{(1/3)} . \quad (15)$$

Both the halo virial overdensity in *N*-body simulations, $\Delta_{\rm vir}^{\rm sim}$, and its corresponding value in the 2LPT field, $\Delta_{\rm vir}^{\rm 2LPT}$ are easy compute. For the *N*-body case, we take the value of Bryan & Norman 1998,

$$\Delta_{\rm vir}^{\rm sim} = (18\pi^2 + 82(\Omega_{\rm m}(z) - 1) - 39(\Omega_{\rm m}(z) - 1)^2)/\,\Omega_{\rm m}(z)\ , \quad (16)$$

where

$$\Omega_{\rm m}(z) = \Omega_{\rm m}(1+z)^3/H^2(z). \quad (17)$$

For the Lagrangian case, $\Delta_{\rm vir}^{\rm 2LPT}$ can be obtained from the relation between the Lagrangian and Eulerian coordinates, giving a value, within the SC approximation, of 35.4 times the mean background density (Manera et al. 2012).

Scoccimarro & Sheth 2002 originally constructed a merger tree to assign haloes masses in cells. This method adopts a mass function and imparts it to the rank-ordered haloes found by the halo finder. PTHALOS has been used for BOSS galaxy clustering analysis (Manera et al. 2012, 2015).

4 RESULTS

In this section, we present and compare the performance of all the methodologies to generate halo catalogues including FOF catalogue (distinct haloes only) and SO catalog (distinct and subhaloes) described in the previous sections. Table 3 lists the particle mesh sizes adopted by the different methodologies, and also shows whether the reduced white noise is used. Note that the mesh sizes used by these methodologies are different from the BigMD simulation (3840^3), so that we cannot use the white noise used by the BigMD as IC directly. We compute the reduced white noise by averaging and rescaling the noise on the neighbour grid points to have the white noise on the smaller mesh size. The reduced white noise will share part but not the whole of the noise with the BigMD simulation. One should keep in mind that the adopted mesh serves different purposes for the different codes and also affects the timing and required resources. For some methodologies, the mesh size influences the scales on which haloes are resolved whereas other methodologies use the reference catalogue to calibrate their specific biasing model to arrive at the final mock halo catalogue.

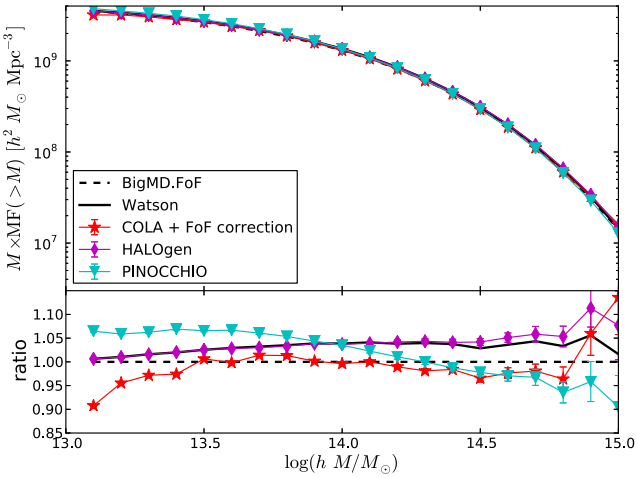


Figure 3. Cumulative mass functions comparing with the BigMD FOF reference catalogue. The error bars were estimated using Jack-knife resampling using 64 different subvolumes. All the methods reproduce the numerical mass function to 5 per cent accuracy.

4.1 Mocks for FOF CATALOGUES

Here, we compare the different mocks with the BigMD FOF reference catalogue (see Section 2). The mesh size used for computing the statistics is 960^3 if applicable.

Some of the methods provide the masses for the halo catalogue. Fig. 3 shows the mass functions provided by COLA, HALOGEN, and PINOCCHIO, compared with that from the BigMD FOF catalogue. COLA FOF masses include the Warren correction due to discrete halo sampling (Warren et al. 2006):

$$M = Nm_{\rm p}(1 - N^{-0.6}), \tag{18}$$

where N is the number of particles in the halo and $m_{\rm p}$ is the particle mass. HALOGEN uses a theoretical mass function from Watson et al. 2013 as an input. All the methods reproduce the numerical mass function to 5 per cent accuracy. The other mocks which do not provide masses could be assigned with a post-processing based on the particle density field (see Zhao et al. 2015).

4.1.1 *Two-point clustering statistics of FOF CATALOGUES*

Two-point clustering statistics is one of the most useful measurements in the clustering analysis of the galaxy surveys. Fig. 4 shows the monopole of the correlation function in real space. Besides PTHALOS, all the mocks agree with the simulation within 5 per cent at the scales between 10 and 50 h^{-1} Mpc. At larger scales, the deviations are basically due to noise. Fig. 5 shows the monopole and quadrupole of the correlation function in redshift space. The comparison of the monopole in redshift space is basically the same as in real space. We have checked that the deviations that COLA have at large scales are mainly due to sample variance (COLA did not use the BigMD white noise). For the quadrupole, COLA agrees with the BigMD within 5 per cent down to the minimum scale we measured (10 h^{-1} Mpc); PINOCCHIO agrees within 10 per cent; EZMOCK and PATCHY are within 15 per cent.

Although, theoretically, the power spectrum is simply a Fourier transform of the two-point correlation function, the performance can be very different. The uncertainties at small scales in the configuration space will propagate to the relative large scales in Fourier space. Fig. 6 shows the monopole of the power spectrum

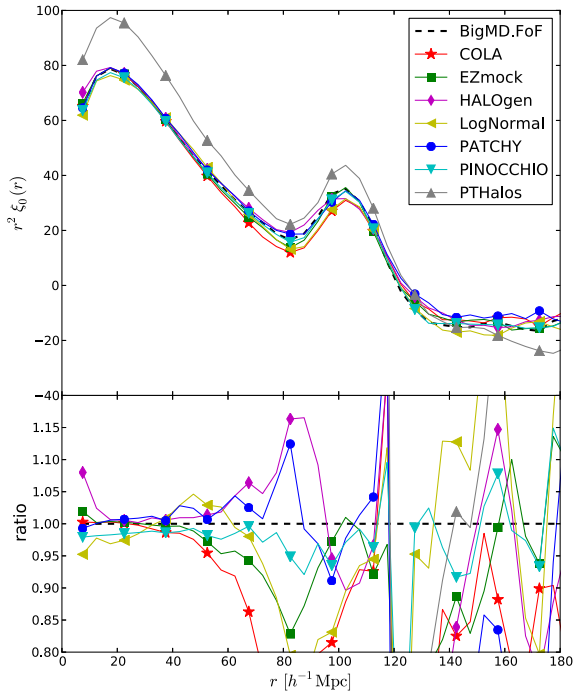


Figure 4. Comparison of the monopole of the correlation function in real space. Dashed line corresponds to the BigMD FOF reference catalogue. COLA FOF masses include the correction due to discrete halo sampling (Warren et al. 2006).

in real space. COLA, EZMOCK, and PATCHY agree with the simulation within 5 per cent for all the scales. HALOGEN, Log-Normal, and PINOCCHIO agree with the simulation within 10 per cent up to $k = 0.2$–$0.25\,h\,{\rm Mpc}^{-1}$. PTHALOS has ∼20 per cent deviation on the linear bias and we have checked that the deviation of PTHALOS can be much smaller if we use lower number density (i.e. massive haloes). In this run, the smaller haloes have mass equivalent to ∼10 particles and some spurious haloes are assigned around large overdensities thus increasing the clustering. Note that the Log-Normal mock is constructed with a input correlation function which is adjusted to be close to that from the simulation. The power spectrum should be better restored if one use a proper input power spectrum. Fig. 7 shows the monopole and quadrupole of the power spectrum in redshift space. For the monopole, COLA, EZMOCK, and PATCHY agree with the simulation within 5 per cent for all the scales shown in the plot; for the quadrupole, COLA agrees with the simulation within 5 per cent for all the scales; PINOCCHIO agrees within 10 per cent; EZMOCK and PATCHY agree with the simulation within 15–20 per cent We find that only the semi-N-body simulation, i.e. COLA, could reach high accuracy at small scales, i.e. $r < 25\,h^{-1}$ Mpc or $k > 0.15\,h\,{\rm Mpc}^{-1}$, on the quadrupole of the correlation function or the power spectrum. The methods based on perturbation theory seem to have some difficulty improving the precision of quadrupole at small scales.

4.1.2 *Three-point clustering statistics of FOF CATALOGUES*

Fig. 8 shows the bispectrum and three-point correlation function in real space. To compute 3PCF, we use the NTROPY-NPOINT software, an exact n-point calculator which uses a kd-tree framework with true parallel capability and enhanced routine performance (Gardner, Connolly & McBride 2007; McBride et al. 2011). We compute the three-point correlation functions with the configuration of the

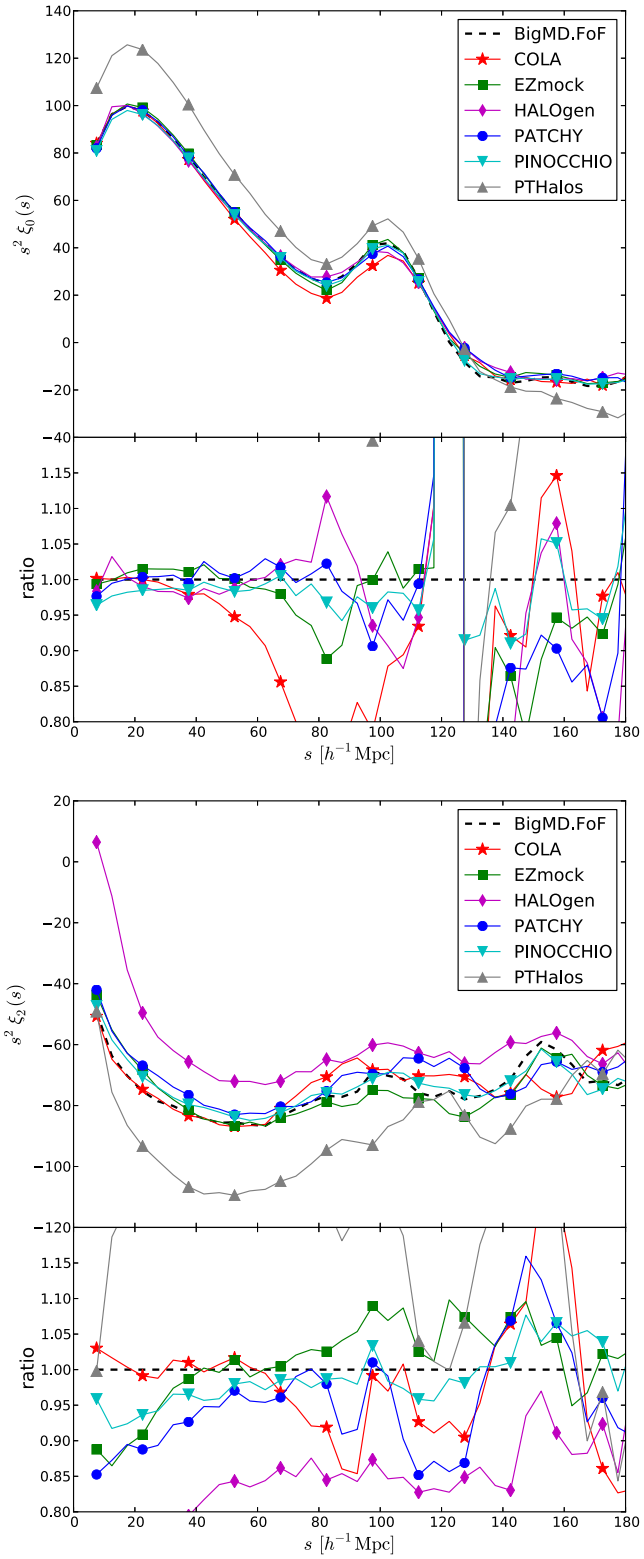


Figure 5. Top panel: comparison of the monopole of the correlation function in redshift space. Bottom panel: performance results for the quadrupole of the correlation function in redshift space. Dashed lines correspond to the BigMD FOF reference catalogue.

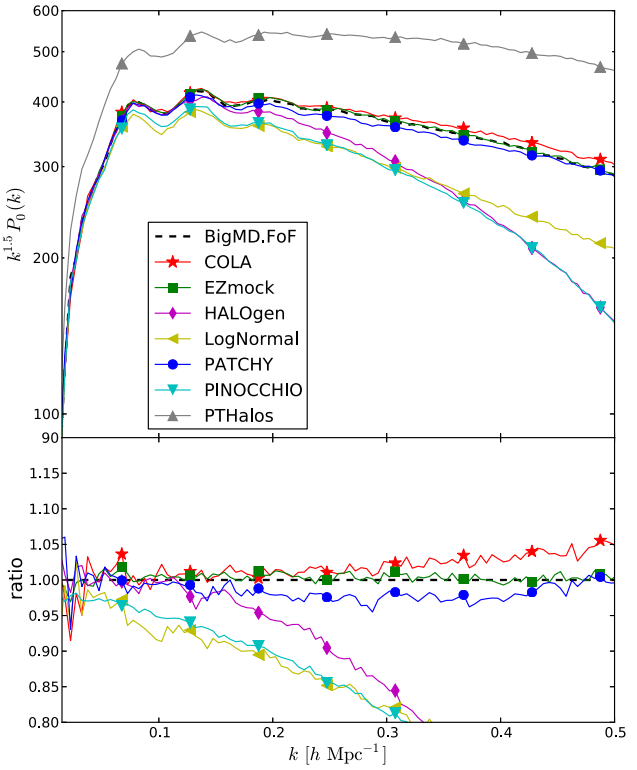


Figure 6. FOF power spectrum comparison, in real space, between the different approximate methods and BigMD.

triangles with two fixed sides, $r_1 = 10\,h^{-1}$ Mpc and $r_2 = 20\,h^{-1}$ Mpc, and varying the third side, r_3. COLA, EZMOCK, PATCHY, PINOCCHIO, and PTHALOS agree with the simulation within the level of noise. We compute the bispectrum with the configuration of the triangles given two fixed sides, $k_1 = 0.1\,h\,\mathrm{Mpc}^{-1}$ and $k_2 = 0.2\,h\,\mathrm{Mpc}^{-1}$, and a varying angle θ_{12}. COLA, EZMOCK, and PATCHY agree very well with the reference simulation catalogue. We conclude that an appropriate bias model is the key to reach high accuracy for the power spectrum and three-point clustering statistics.

4.2 Mocks for SO/BDM CATALOGUES

Here, we discuss the performance of the different approximate methods when we compare with the SO catalogue (obtained using BDM code) from BigMD with the same halo number density. Note that this catalogue includes both distinct haloes and subhaloes (see Section 2). The mesh size used for computing the statistics is 960^3 if applicable. Note that while EZMOCK, Log-Normal, and PATCHY mocks for the SO catalogue are generated with the same procedures as that for the FOF catalogue, COLA, HALOGEN, and PINOCCHIO are including subhaloes following a halo occupation distribution (HOD) scheme described in the appendix. In addition, while COLA and PINOCCHIO are using the FOF mocks as the distinct haloes to assign the subhaloes around them, HALOGEN constructs a new catalogue matching the SO distinct haloes before the HOD process. PTHALOS is not included in this section.

4.2.1 Two-point clustering statistics of SO catalogues

Fig. 9 shows the performance of the different methods on the monopole of correlation function in real space. All the mocks agree with the simulation very well. Fig. 10 shows the comparison

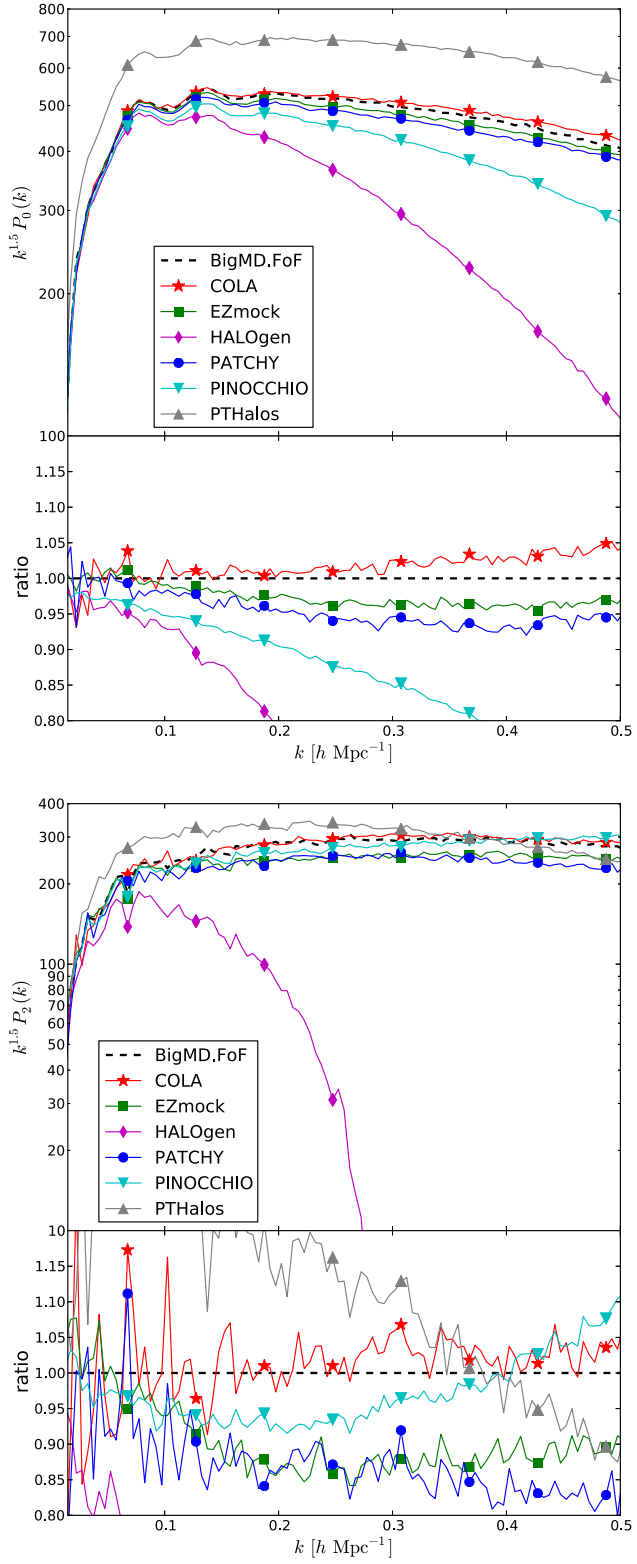


Figure 7. Top panel: performance results for the monopole of the power spectrum in redshift space. Bottom panel: comparison of the quadrupole of the power spectrum in redshift space. Dashed lines correspond to the BigMD FOF reference catalogue.

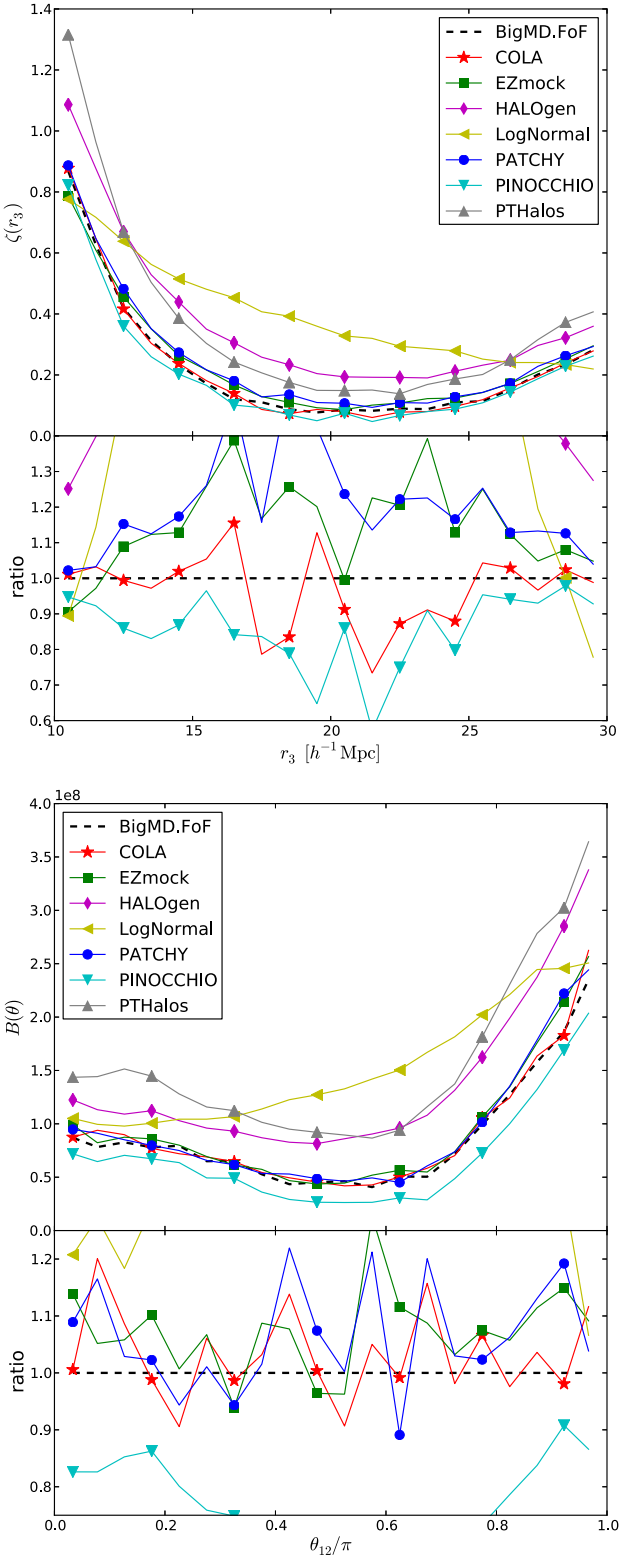


Figure 8. Top panel: performance results for the three-point correlation function in real space. Bottom panel: bispectrum in real space. Dashed lines correspond to the BigMD FOF reference catalogue.

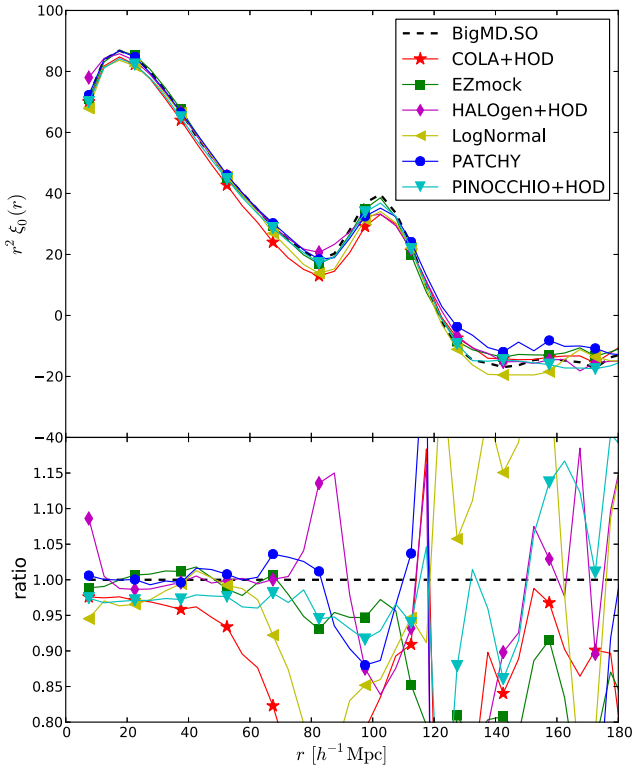
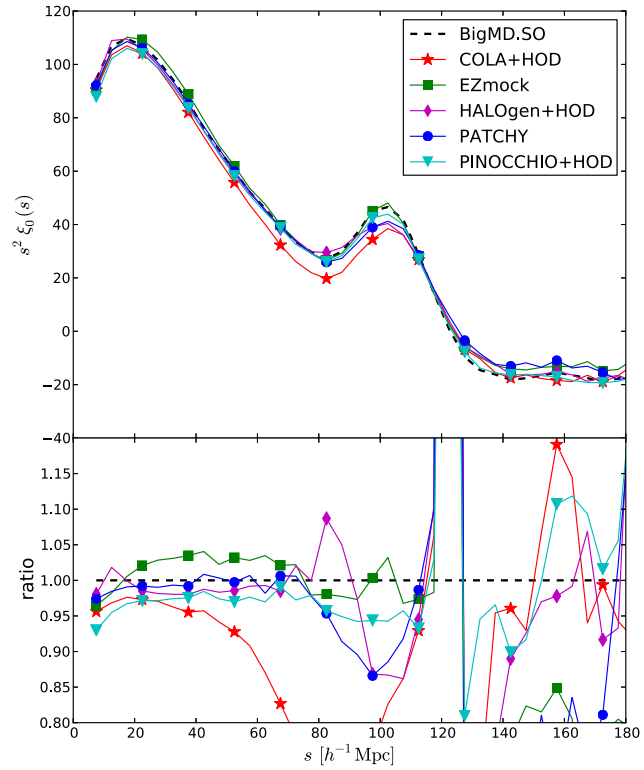


Figure 9. Comparison of the monopole of the correlation function in real space. Dashed line corresponds to the BigMD SO reference catalogue.

for the monopole and quadrupole of the correlation function in redshift space. For the monopole, COLA+HOD shows some deviation at scales $>40\,h^{-1}$ Mpc, which may be due to not using the BigMD white noise. For the quadrupole, EZMOCK, PATCHY, and PINOCCHIO+HOD agree with the simulation catalogue within 10 per cent for all the scales considered. COLA+HOD agrees within 10 per cent down to $r = 15\,h^{-1}$ Mpc.

Fig. 11 shows the monopole of the power spectrum in real space. EZMOCK and PATCHY agree with BigMD within 5 per cent for all the scales. COLA+HOD and HALOGEN+HOD are within 10 per cent up to $k \sim 0.35\,h\,\mathrm{Mpc}^{-1}$, and PINOCCHIO+HOD and Log-Normal are within 10 per cent up to $k \sim 0.1\,h\,\mathrm{Mpc}^{-1}$. Note again that the Log-Normal mock should be able to agree better with the simulation if one uses a proper input power spectrum. Fig. 12 shows the performance comparison for the monopole and quadrupole of the power spectrum in redshift space. COLA+HOD, EZMOCK, and PATCHY agree with BigMD monopole within 10 per cent for all the scales; and up to $k \sim 0.1\,h\,\mathrm{Mpc}^{-1}$ for HALOGEN+HOD and PINOCCHIO+HOD. For the quadrupole, EZMOCK and PATCHY agree with the simulation within 10 per cent for all the scales; COLA+HOD and PINOCCHIO+HOD agree up to $k = 0.25\,h\,\mathrm{Mpc}^{-1}$.

As discussed in the appendix, we test our HOD scheme by applying it to the SO distinct haloes from the BigMD simulation, trying to reproduce the clustering of substructures. We also test on the BigMD FOF catalogue. We find that HOD scheme has good performance in real space but the difference between SO distinct halo catalogue and FoF catalogue would introduce some bias. We also find that it is not trivial to correctly model the velocity distribution of the substructure which results the relatively poor performance of the HOD model in redshift space.



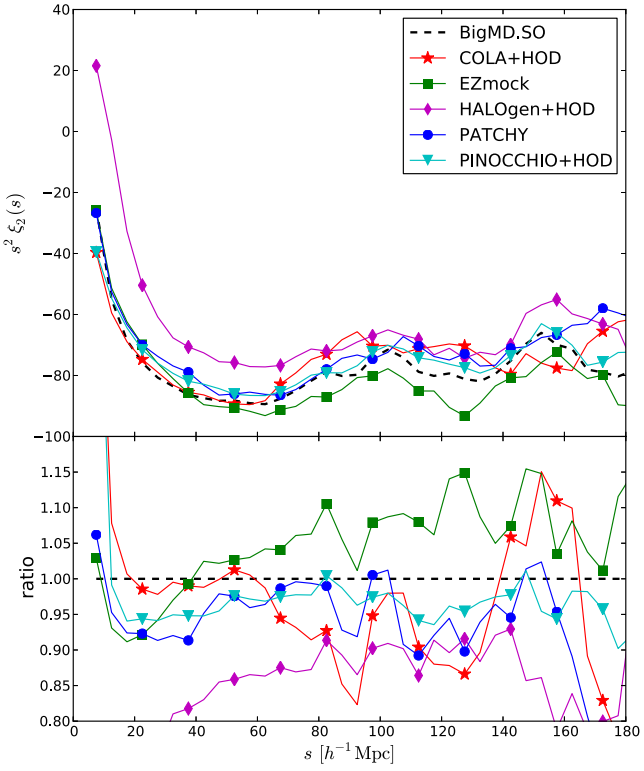


Figure 10. Top panel: comparison of the monopole of the correlation function in redshift space. Bottom panel: performance results for the quadrupole of the correlation function in redshift space. Dashed lines correspond to the BigMD SO reference catalogue.

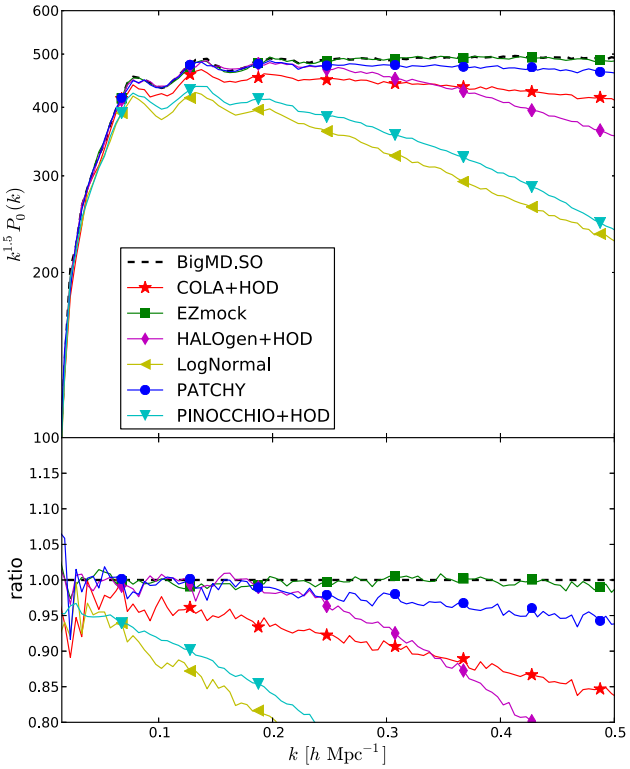


Figure 11. SO power spectrum comparison, in real space, between the different approximate methods and BigMD.

4.2.2 *Three-point clustering statistics of SO catalogues*

Fig. 13 shows the bispectrum and three-point correlation function in real space. The configurations are the same as for FOF catalogues. For the three-point correlation function, EZMOCK and PATCHY agree with the simulation within the level of noise. COLA+HOD and PINOCCHIO+HOD agree with the simulation within 20 per cent. For the bispectrum, COLA+HOD, EZMOCK, and PATCHY agree within 10–20 per cent with the reference simulation catalogue. We conclude that an appropriate bias model is the key to reach high accuracy for the power spectrum and three-point clustering statistics.

5 SUMMARY

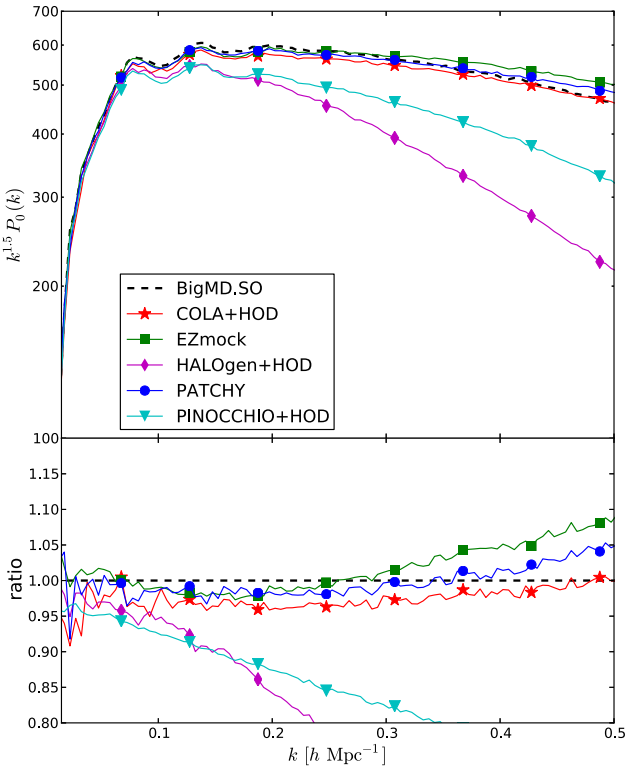
In this paper, we have compared the performance of seven different approximate methods to model the halo/galaxy clustering statistics. The resulting mock catalogues from each method have been compared to a reference FOF and SO halo catalogue drawn from the *Planck* BigMD simulation with similar clustering properties that the BOSS galaxies at $z \sim 0.5$. Note that the methods compared in this study might have different advantages and applications, e.g. merging history, etc., which are not included in this study.

We are listing some items we have learned from this comparison study and have more discussion following the list.

(I) Most of the methods are able to reproduce the two-point statistics in configuration space but not necessary in Fourier space.

(II) An appropriate bias model is the key to reach high accuracy for the power spectrum and three-point statistics, including bispectrum and three-point correlation function.

(III) In redshift space, so far, only the semi-*N*-body simulation, i.e. COLA, could reach high accuracy (1 per cent level) at small scales,



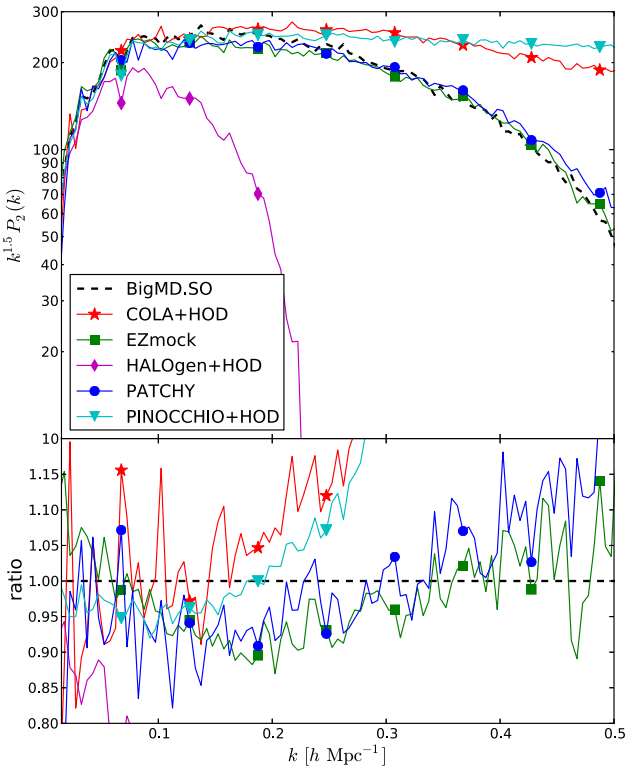


Figure 12. Top panel: performance results for the monopole of the power spectrum in redshift space. Bottom panel: comparison of the quadrupole of the power spectrum in redshift space. Dashed lines correspond to the BigMD SO reference catalogue.

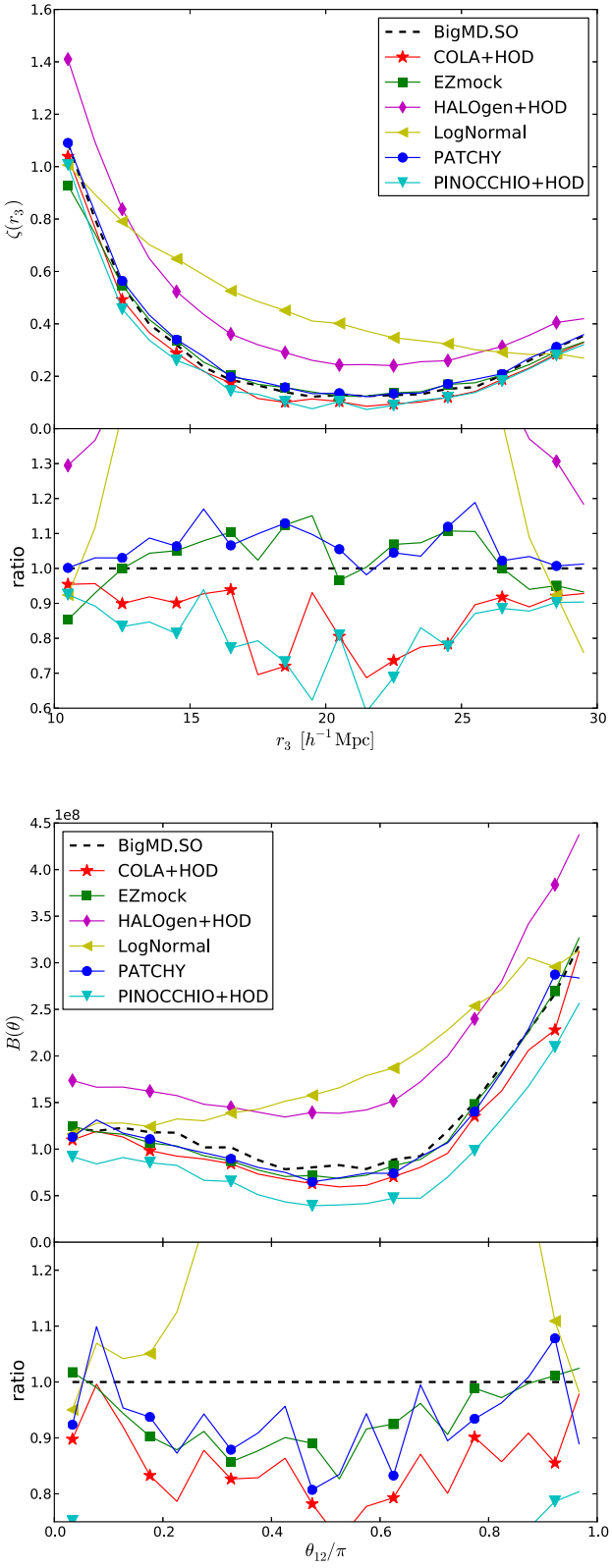


Figure 13. Top panel: performance results for the three-point correlation function in real space. Bottom panel: bispectrum in real space. Dashed lines correspond to the BigMD SO reference catalogue.

i.e. $r < 25\,h^{-1}$ Mpc or $k > 0.15\,h\,\mathrm{Mpc}^{-1}$, on the quadrupole of the correlation function or the power spectrum.

(IV) It is not trivial to fit a catalogue that contains substructures (e.g. SO catalogue) starting from a catalogue with only distinct haloes and applying a HOD scheme on it.

The position of DM particles after cosmic evolution according to perturbative approaches shows a typical uncertainty of roughly a few Mpc, depending on the chosen approximation (e.g. see Kitaura & Hess 2013; Monaco et al. 2013). This does not show up so clearly in the correlation function in configuration space, where the small scales are kept separated from the large ones. However, it does have a very clear impact in the power spectrum, as it does not reproduce the one halo term, and thus lacks the commonly known non-linear bump towards high *k*s. Small-scale uncertainties propagate in Fourier space having the effect of a convolution (see Tassev & Zaldarriaga 2012; Monaco et al. 2013). In this work, we have presented two kinds of approaches based on perturbation theory. Those which rely on the approximate position of the DM particles to find the haloes, and those which just use its large-scale structure density field combined with a statistical population prescription to populate the haloes. We find that the first ones are more sensitive to the uncertainty in the particle positions and thus show a larger deviation in Fourier space than in configuration space. While the second class of methods circumvent the problem, by compensating the deviation with the adopted bias description. It is arguable whether one wants to maintain the analytical models as they are and accept their uncertainties while having a clear understanding of their systematics, or modify them with additional prescriptions to fit the simulations, and introduce more complex relations.

The methods based on perturbation theory seem to have some difficulty improving the precision of quadrupole at small scales. White 2014 built the theoretical model for biased tracers (i.e. haloes or galaxies) in configuration space and also found similar deviations in the quadrupole comparing to the *N*-body simulation at small scales.

An HOD model is typically used to analyse some two-point clustering measurement (e.g. projected correlation function) and therefore the model is consistent with the clustering by construction. However, one could simply adopt an HOD model from a particular halo catalogue, and there is no guarantee that the resulting mock catalogue reproduces the expected clustering signal. In addition, if a model is calibrated only to the clustering length or bias (i.e. the two-halo term), it might not reproduce the small-scale clustering. Also, different types of galaxies (or haloes) may have different spatial clustering and may occupy haloes differently or have different central/satellite fractions, so itis important to note that different HOD models may be required. While our HOD application leads to the results reported in this study, an improved (less standard or less straightforward) application could yield better agreement in terms of two-point statistics. This should further investigated in future works.

ACKNOWLEDGEMENTS

The authors would like to express special thanks to the Instituto de Fisica Teorica (IFT-UAM/CSIC in Madrid) for its hospitality and support, via the Centro de Excelencia Severo Ochoa Program under Grant No. SEV-2012-0249, during the three-week workshop ‘nIFTy Cosmology’ where this work developed. We further acknowledge the financial support of the University of Western 2014 Australia Research Collaboration Award for ‘Fast Approximate Synthetic

Universes for the SKA', the ARC Centre of Excellence for All Sky Astrophysics (CAASTRO) grant number CE110001020, and the two ARC Discovery Projects DP130100117 and DP140100198. We also recognize support from the Universidad Autonoma de Madrid (UAM) for the workshop infrastructure.

We especially like to thank Frazer Pearce for initiating (together with AK) the Mocking Astrophysics[18] programme under whose umbrella the workshop and this work was performed, respectively. We thank the developers of COLA, Tassev, Zaldarriaga, Eisenstein and Koda, for making the code public available. We thank Jeffery Gardner and Cameron McBride for sharing the NTROPY-NPOINT code.

CC and FP were supported by the Spanish MICINNs Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064 and AYA2010-21231-C02-01 grant, and Spanish MINECOs Centro de Excelencia Severo Ochoa Programme under grant SEV-2012-0249. CZ and CT acknowledges support from Tsinghua University, and 973 student programme No. 2013CB834906. CZ also thanks the support from the MultiDark summer student program to visit the Instituto de Física Teórica, (UAM/CSIC), Spain. AK is supported by the *Ministerio de Economía y Competitividad* (MINECO) in Spain through grant AYA2012-31101, as well as the Consolider-Ingenio 2010 Programme of the *Spanish Ministerio de Ciencia e Innovación* (MICINN) under grant MultiDark CSD2009-00064. He also acknowledges support from the *Australian Research Council* (ARC) grants DP130100117 and DP140100198. EM and PM have been supported by a FRA2012 grant of the University of Trieste, PRIN2010-2011 (J91J12000450001) from MIUR, and by Consorzio per la Fisica di Trieste. AI is supported by the JAE program grant from the Spanish National Science Council (CSIC). GY acknowledges support from the Spanish MINECO under research grants AYA2012-31101, FPA2012-34694, AYA2010-21231, Consolider Ingenio SyeC CSD2007-0050 and from Comunidad de Madrid under ASTROMADRID project (S2009/ESP-1496). VT and SG have been supported by the German Science Foundation under DFG grant GO563/23-1. FAM is supported by the Australian Research Council Centre of Excellence for All-Sky Astrophysics (CAASTRO) through project number CE110001020.

The MultiDark Database (www.multidark.org) used in this paper and the web application providing online access to it were constructed as part of the activities of the German Astrophysical Virtual Observatory as result of a collaboration between the Leibniz-Institute for Astrophysics Potsdam (AIP) and the Spanish MultiDark Consolider Project CSD2009-00064. The BigMD simulation suite has been performed in the Supermuc supercomputer at LRZ using time granted by PRACE. The simulation has been performed with an optimized version of GADGET-2 kindly provided by Volker Springel. We also acknowledge PRACE for awarding us access to resource Curie supercomputer based in France (project PA2259). PTHALOS was run in COSMA, the HPC facilities at the ICC in Durham, for a project granted by of the STFC DiRAC HPC Facility (www.dirac.ac.uk). Some other computations were performed on HYDRA, the HPC-cluster of the IFT-UAM/CSIC.

The authors contributed in the following ways to this paper: FP coordinated the approximate method workshop programme from which this comparison project and paper originated. The analysis presented here was performed by CC and CZ. The paper was written by CC, EM, and FP. The mock catalogues and descriptions are run and provided by AI (COLA), SA (HALOGEN), CC (EZMOCK), FK (PATCHY), MM (PTHALOS), PM (PINOCCHIO), and SM (HALOGEN). AK, AAK, CS, GY SG, and VT prepared the reference BigMD simulation catalogues; EM developed and applied an HOD scheme for this study; VM and FM helped to develop the code for computing three-point correlation function; other authors contributed towards the content of the paper and helped to proof read it.

[18] http://www.mockingastrophysics.org

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Abell P. A. et al., 2009, preprint (arXiv:0912.0201)
Avila S. G., Murray S., Knebe A., Power C., Robotham A. S., Garcia-Bellido J., 2015, MNRAS, 450, 1856
Benitez N. et al., 2015, in Cenarro A. J., Figueras F., Hernández-Monteagudo C., Trujillo Bueno J., Valdivielso L., eds, Highlights of Spanish Astrophysics VIII, p. 148
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Beutler F. et al., 2011, MNRAS, 416, 3017
Bhattacharya S., Habib S., Heitmann K., Vikhlinin A., 2013, ApJ, 766, 32
Blake C. et al., 2011, MNRAS, 415, 2892
Bryan G., Norman M., 1998, ApJ, 495, 80
Chuang C.-H., Wang Y., Hemantha M. D. P., 2012, MNRAS, 423, 1474
Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621
Cole S. et al., 2005, MNRAS, 362, 505
Coles P., Jones B., 1991, MNRAS, 248, 1
Colless M. et al., 2001, MNRAS, 328, 1039
Colless M. et al., 2003, preprint (astro-ph/0306581)
Davis M., Efstathiou G., Frenk C. S., White S. D., 1985, ApJ, 292, 371
Dawson K. S. et al., 2013, AJ, 145, 10
de Jong R. S. et al., 2012, Proc. SPIE, 8446, 84460T
de la Torre S. et al., 2013, A&A, 557, A54
Drinkwater M. J. et al., 2010, MNRAS, 401, 1429
Eisenstein D. J. et al., 2011, AJ, 142, 72
Frieman J., Dark Energy Survey Collaboration, 2013, BAAS, 221, 335.01
Gardner J. P., Connolly A., McBride C., 2007, preprint (arXiv:0709.1967)
Green J. et al., 2012, preprint (arXiv:1208.4012)
Hill G. et al., 2008, ASP Conf. Ser. Vol. 399, Panoramic Views of Galaxy Formation and Evolution. Astron. Soc. Pac., San Francisco, p. 115
Hubble E., 1934, ApJ, 79, 8
Kazin E. A., Koda J., Blake C., Padmanabhan N., 2014, MNRAS, 441, 3524
Kitaura F., Angulo R., 2012, MNRAS, 425, 2443
Kitaura F.-S., Hess S., 2013, MNRAS, 435, 78
Kitaura F. S., Jasche J., Li C., Ensslin T. A., Metcalf R. B., Wandelt B. D., Lemson G., White S. D. M., 2009, MNRAS, 400, 183
Kitaura F.-S., Yepes G., Prada F., 2014, MNRAS, 439, L21
Kitaura F.-S., Gil-Marín H., Scoccola C., Chuang C.-H., Müller V., Yepes G., Prada F., 2015, MNRAS, 450, 1836
Klypin A., Holtzman J., 1997, preprint (astro-ph/9712217)
Klypin A., Yepes G., Gottlober S., Prada F., Hess S., 2014, preprint (arXiv:1411.4001)
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottloeber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Levi M. et al., 2013, preprint (arXiv:1308.0847)
McBride C. K., Connolly A. J., Gardner J. P., Scranton R., Newman J. A., Scoccimarro R., Zehavi I., Schneider D. P., 2011, ApJ, 726, 13
Manera M. et al., 2012, MNRAS, 428, 1036
Manera M. et al., 2015, MNRAS, 447, 437
Monaco P., Theuns T., Taffoni G., 2002, MNRAS, 331, 587
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Nuza S. et al., 2013, MNRAS, 432, 743
Parkinson D. et al., 2012, Phys. Rev. D, 86, 103518
Percival W. J. et al., 2010, MNRAS, 401, 2148
Pope A. C., Szapudi I., 2008, MNRAS, 389, 766
Reid B. A. et al., 2010, MNRAS, 404, 60
Riebe K. et al., 2011, preprint (arXiv:1109.0003)

Samushia L., Percival W. J., Raccanelli A., 2012, MNRAS, 420, 2102
Schlegel D. et al., 2011, preprint (arXiv:1106.1706)
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Skibba R. A., Sheth R. K., 2009, MNRAS, 392, 1080
Springel V., 2005, MNRAS, 364, 1105
Tassev S., Zaldarriaga M., 2012, J. Cosmol. Astropart. Phys., 1204, 013
Tassev S., Zaldarriaga M., Eisenstein D., 2013, J. Cosmol. Astropart. Phys., 1306, 036
Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, ApJ, 646, 881
Watson W. A., Iliev I. T., D'Aloisio A., Knebe A., Shapiro P. R., Yepes G., 2013, MNRAS, 433, 1230
White M., 2014, MNRAS, 439, 3630
White M., Tinker J. L., McBride C. K., 2014, MNRAS, 437, 2594
Wild V. et al., 2005, MNRAS, 356, 247
York D. G. et al., 2000, AJ, 120, 1579
Zehavi I. et al., 2011, ApJ, 736, 59
Zhao C., Kitaura F.-S., Chuang C.-H., Prada F., Yepes G., Tao C., 2015, MNRAS, preprint (arXiv:1501.05520)
Zheng Z. et al., 2005, ApJ, 633, 791

APPENDIX A: ASSIGNING SUBHALOES WITH AN HOD PRESCRIPTION

The approximative mock methods are all designed to give halo catalogues, but (due to aforementioned limitations) not all of them are capable of adding subhaloes to them. Therefore, we applied a post-processing step, i.e. the HOD, to them augmenting their submitted catalogues with subhaloes. The HOD approach is based on a statistical assignment of the number, positions, and velocities of substructures residing in a halo as a function of the halo mass, e.g. Berlind & Weinberg 2002, Kravtsov et al. 2004, Zheng et al. 2005, Skibba & Sheth 2009, Zehavi et al. 2011.

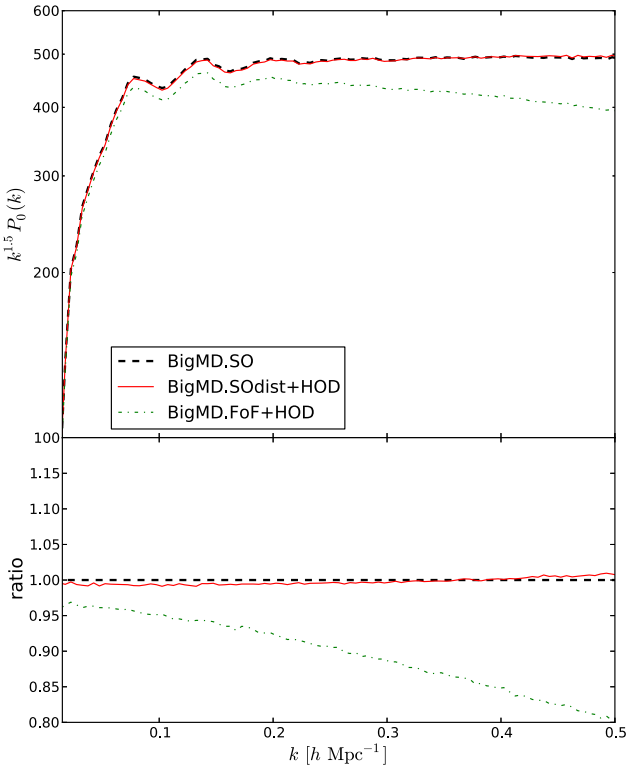


Figure A1. HOD Power spectrum comparison, in real space, among the BigMD SO catalogue, SO distinct haloes catalogue with HOD applied, and FOF catalogue with HOD applied.

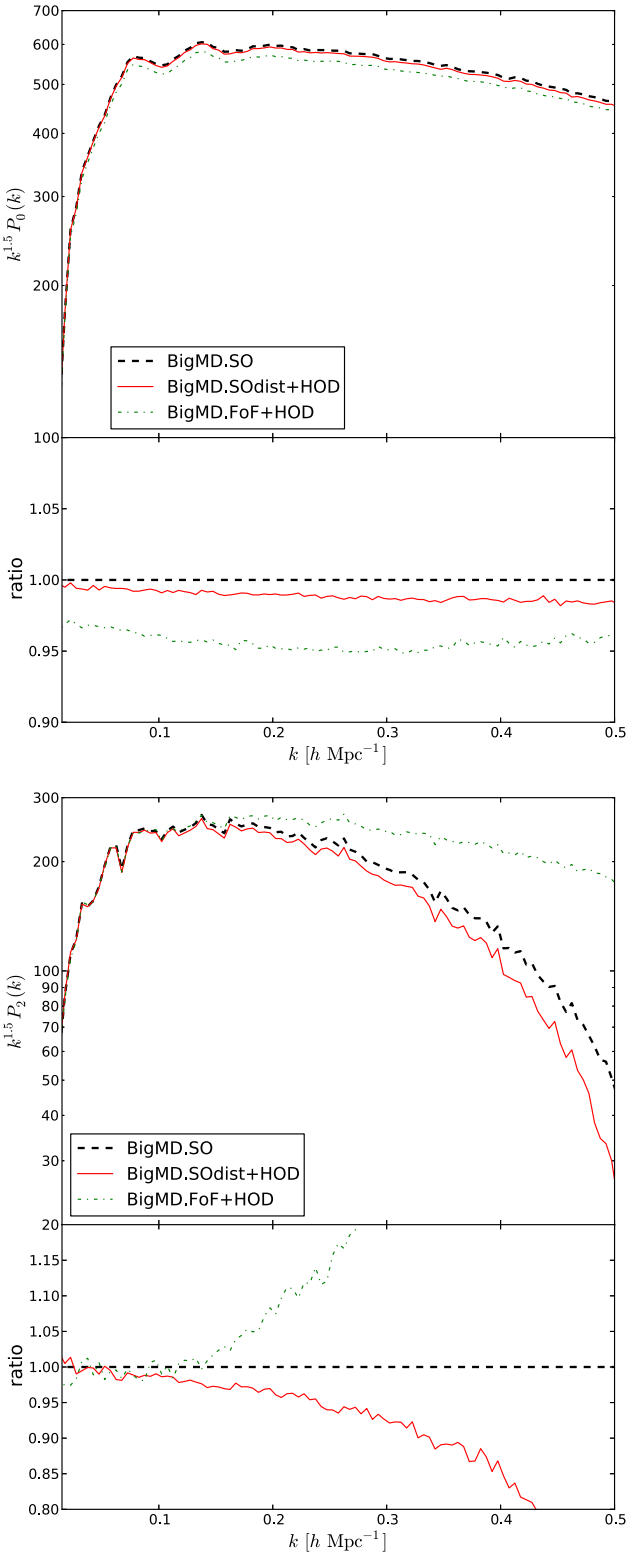


Figure A2. Top panel: HOD performance results for the monopole of the power spectrum in redshift space. Bottom panel: comparison of the quadrupole of the power spectrum in redshift space.

We have applied an HOD scheme to PINOCCHIO, COLA, and HALOGEN halo catalogues. For the first two methods, we have first converted the values of mass into the values corresponding to bound masses, in order to be compatible with the definition adopted in the BigMD simulation. For PINOCCHIO and COLA, we have looked for a transformation that maps the halo masses into new mass values imposing that the mass function matches the one of the BigMD SO reference catalogue.

The following step consists in looking for a relation that associates the halo mass of the BigMD with the average number of substructures in the haloes of that mass.

We have considered logarithmically equispaced mass bins. In each bin, the distribution of haloes with a given number of substructures (main haloes included) is verified to be Poisson distributed, and the best-fitting Poisson parameter $\lambda(M)$ is assigned to that bin as representative of the mean number of substructures.

It is now possible to populate the haloes obtained with PINOCCHIO, COLA, and HALOGEN, with a population of substructures statistically identical to that of the BigMD reference catalogue. The actual number of substructures in a halo is assigned as a random number taken from a Poisson distribution having the mean value $\lambda(M)$.

Substructures are spatially distributed in order to have an NFW number density profile, with concentration equal to the main halo's one. The latter is computed following Bhattacharya et al. (2013). Peculiar velocities in each of the three directions are randomly extracted from a Gaussian distribution having null mean and dispersion equal to $\sqrt{GM(r)/r}$.

We test and validate our HOD scheme by applying it on BigMD SO distinct halo catalogue and BigMD FOF catalogue. Fig. A1 shows the power spectrum in real space. One can see that BigMD SO distinct haloes with HOD scheme applied agrees with the full BigMD SO catalogue very well. BigMD FOF catalogue with HOD scheme applied has 5 per cent deviation which will propagate to the mocks to which we apply the HOD scheme in this study. Fig. A2 shows the monopole and quadrupole of power spectrum in redshift space. For the monopole, BigMD SO distinct haloes with HOD scheme applied agrees with the full BigMD SO catalogue very well; for quadrupole, it agree within 20 per cent up to $k = 0.4\,h\,\mathrm{Mpc}^{-1}$.

[1]*Instituto de Física Teórica, (UAM/CSIC), Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*
[2]*Tsinghua Center for Astrophysics, Department of Physics, Tsinghua University, Haidian District, Beijing 100084, People Republic of China*
[3]*Campus of International Excellence UAM+CSIC, Cantoblanco, E-28049 Madrid, Spain*
[4]*Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, E-18080 Granada, Spain*
[5]*Dipartimento di Fisica – Sezione di Astronomia, Università di Trieste, via Tiepolo 11, I-34131 Trieste, Italy*
[6]*INAF, Osservatorio Astronomico di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy*
[7]*Departamento de Física Teórica, Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*
[8]*Institut de Ciències de l'Espai, IEEC-CSIC, Campus UAB, Facultat de Ciències, Torre C5 par-2, E-08193 Barcelona, Spain*
[9]*Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany*
[10]*University College London, Gower Street, London WC1E 6BT, UK*
[11]*ICRAR, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[12]*ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO), 44 Rosehill Street, Redfern, NSW 2016, Australia*
[13]*Facultad de Ciencias Astronómicas y Geofísicas – Universidad Nacional de La Plata. Paseo del Bosque S/N, 1900 La Plata, Argentina*
[14]*CONICET, Rivadavia 1917, 1033Buenos Aires, Argentina*
[15]*Technology, PO Box 218, Hawthorn, VIC 3122, Australia*
[16]*Centre for Astrophysics & Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia*
[17]*Department of Physics, Center for Astrophysics and Space Sciences, University of California, 9500 Gilman Drive, San Diego, CA 92093, USA*
[18]*Department of Astronomy, New Mexico State University, PO Box 30001, MSC 4500, Las Cruces, NM 88003, USA*
[19]*CPPM, Université Aix-Marseille, CNRS/IN2P3, Case 907, F-13288 Marseille Cedex 9, France*
[20]*Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, 125047 Moscow, Russia*

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