

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

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

<sup>1</sup> Instituto de Física Teórica, (UAM/CSIC), Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain

<sup>2</sup> Tsinghua Center for Astrophysics, Department of Physics, Tsinghua University, Haidian District, Beijing 100084, P. R. China

<sup>3</sup> Campus of International Excellence UAM+CSIC, Cantoblanco, E-28049 Madrid, Spain

<sup>4</sup> Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, E-18080 Granada, Spain

<sup>5</sup> Dipartimento di Fisica - Sezione di Astronomia, Università di Trieste, via Tiepolo 11, I- 34131 Trieste, Italy

<sup>6</sup> Departamento de Física Teórica, Universidad Autónoma de Madrid, Cantoblanco, 28049, Madrid, Spain

<sup>7</sup> Institut de Ciències de l'Espai, IEEC-CSIC, Campus UAB, Facultat de Ciències, Torre C5 par-2, Barcelona 08193, Spain

<sup>8</sup> Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany

<sup>9</sup> University College London, Gower Street, London WC1E 6BT, UK

<sup>10</sup> INAF, Osservatorio Astronomico di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy

<sup>11</sup> ICRAR, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia

<sup>12</sup> ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO)

<sup>13</sup> Facultad de Ciencias Astronómicas y Geofísicas - Universidad Nacional de La Plata. Paseo del Bosque S/N (1900). La Plata, Argentina

<sup>14</sup> CONICET, Rivadavia 1917, 1033, Buenos Aires, Argentina

<sup>15</sup> Technology, P.O. Box 218, Hawthorn, VIC 3122, Australia

<sup>16</sup> Centre for Astrophysics & Supercomputing, Swinburne University of Technology, P.O. Box 218, Hawthorn, VIC 3122, Australia

<sup>17</sup> Department of Physics, Center for Astrophysics and Space Sciences, University of California, 9500 Gilman Drive, San Diego, CA 92093

<sup>18</sup> Astronomy Department, New Mexico State University, Las Cruces, NM, USA

<sup>19</sup> CPPM, Université Aix-Marseille, CNRS/IN2P3, Case 907, 13288 Marseille Cedex 9, France

<sup>20</sup> Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, 125047 Moscow, Russia

13 May 2015

## 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 2-point correlation function in real- and redshift-space), and the three-point clustering statistics (bispectrum and 3-point correlation function). The reference catalogues are drawn from the BigMultiDark  $N$ -body simulation. Both friend-of-friends (including distinct halos only) and spherical overdensity (including distinct halos 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 quasilinear 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% level) at small scales, i.e.,  $r < 25 h^{-1}$  Mpc or  $k > 0.15 h$  Mpc<sup>-1</sup>. 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 percent accuracy in terms of clustering statistics at the scales interesting for

\* E-mail: chia-hsun.chuang@uam.es, MultiDark Fellow

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<sup>1</sup> (2dFGRS) obtained 221,414 galaxy redshifts at  $z < 0.3$  (Colless et al. 2001, 2003), and the Sloan Digital Sky Survey<sup>2</sup> (SDSS, York et al. 2000) collected 930,000 galaxy spectra in the Seventh Data Release (DR7) at  $z < 0.5$  (Abazajian et al. 2009). WiggleZ<sup>3</sup> collected spectra of 240,000 emission-line galaxies at  $0.5 < z < 1$  over 1,000 square degrees (Drinkwater et al. 2010; Parkinson et al. 2012), and the Baryon Oscillation Spectroscopic Survey<sup>4</sup> (BOSS, Dawson et al. 2013) of the SDSS-III project (Eisenstein et al. 2011) has surveyed 1.5 million luminous red galaxies (LRGs) at  $0.1 < z < 0.7$  over 10,000 square degrees. There are new upcoming ground-based and space experiments, such as 4MOST<sup>5</sup> (4-metre Multi-Object Spectroscopic Telescope, de Jong et al. 2012), DES<sup>6</sup> (Dark Energy Survey, Frieman J. 2013), DESI<sup>7</sup> (Dark Energy Spectroscopic Instrument, Schlegel et al. 2011; Levi et al. 2013), eBOSS<sup>8</sup> (Extended Baryon Oscillation Spectroscopic Survey), HETDEX<sup>9</sup> (Hobby-Eberly Telescope Dark Energy Experiment, Hill et al. 2008), J-PAS<sup>10</sup> (Javalambre Physics of accelerating universe Astrophysical Survey, Benitez et al. 2014), LSST<sup>11</sup> (Large Synoptic Survey Telescope, Abell et al. 2009), Euclid<sup>12</sup> (Laureijs et al. 2011), and WFIRST<sup>13</sup> (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<sup>14</sup> (Large Suite of Dark Matter Simulations), which has been used to analyse the SDSS-II galaxy sample (e.g., Chuang et al. 2012; Samushia et al. 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 Log-Normal 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 halos which can be provided by COLA, PINOCCHIO, and PTHalos. The other methods, i.e., EZmock, HALOgen, Log-Normal, and PATCHY, generate halos with some biasing models calibrated with the  $N$ -body simulations.

This paper – emerging out of the ‘nIFTy cosmology’ workshop<sup>15</sup> – 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<sup>16</sup> (Klypin et al. 2014), which was performed using GADGET-2 Springel (2005) with  $3840^3$  particles on a volume of  $(2500 h^{-1} \text{Mpc})^3$  assuming  $\Lambda$ CDM Planck cosmology with  $\{\Omega_M = 0.307115, \Omega_b = 0.048206, \sigma_8 = 0.8288, n_s = 0.96\}$ , and a Hubble constant ( $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$ ) given by  $h = 0.6777$ . Within the MultiDark project a series of 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).

<sup>1</sup> <http://www2.aao.gov.au/2dfgrs/>

<sup>2</sup> <http://www.sdss.org>

<sup>3</sup> <http://wigglez.swin.edu.au/site/>

<sup>4</sup> <https://www.sdss3.org/surveys/boos.php>

<sup>5</sup> <http://www.4most.eu/>

<sup>6</sup> <http://www.darkenergysurvey.org>

<sup>7</sup> <http://desi.lbl.gov/>

<sup>8</sup> <http://www.sdss.org/sdss-surveys/eboss/>

<sup>9</sup> <http://hetdex.org>

<sup>10</sup> <http://j-pas.org>

<sup>11</sup> <http://www.lsst.org/lsst/>

<sup>12</sup> <http://www.euclid-ec.org>

<sup>13</sup> <http://wfirst.gsfc.nasa.gov>

<sup>14</sup> <http://lss.phy.vanderbilt.edu/lasdamas/>

<sup>15</sup> <http://popia.ft.uam.es/nIFTyCosmology>

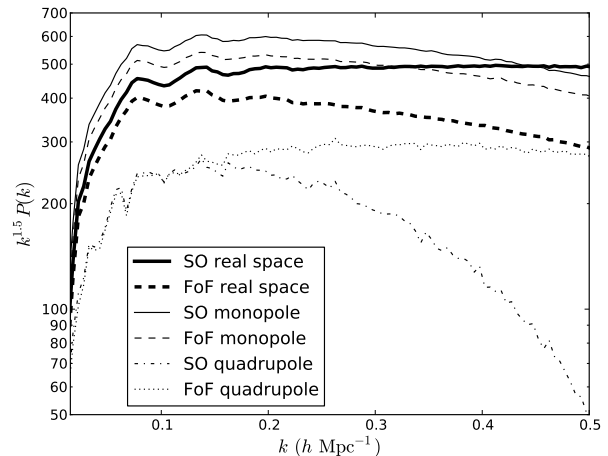
<sup>16</sup> <http://www.multidark.org/>

Methodology	reference
<b>Log-Normal</b>	Coles & Jones 1991
<b>PTHalos</b>	Manera et al. 2012, 2015
<b>PINOCCHIO</b> (PINpointing Orbit-Crossing Collapsed Hierarchical Objects)	Monaco et al. 2002, 2013
<b>COLA</b> (COmoving Lagrangian Acceleration simulation)	Tassev et al. 2013
<b>PATCHY</b> (PerturbAtion Theory Catalog generator of Halo and galaxY distributions)	Kitaura et al. 2014a,b
<b>QPM</b> (quick particle mesh)	White et al. 2013
<b>EZmock</b> (Effective Zel'dovich approximation mock catalogue)	Chuang et al. 2015
<b>HALOgen</b>	Avila et al. 2014

**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.

Halos 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  $3.5 \times 10^{-4} h^3 \text{Mpc}^{-3}$ , which is similar to that of the BOSS galaxy sample at  $z \sim 0.5$ . This abundance is equivalent to a mass cut of  $\sim 1 \times 10^{13} M_{\odot}/h$  for the FoF catalogue and  $\sim 8.5 \times 10^{12} M_{\odot}/h$  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  $k \lesssim 1 h \text{Mpc}^{-1}$  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  $\lambda > 2\pi/k \sim 6 h^{-1} \text{Mpc}$  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) are 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

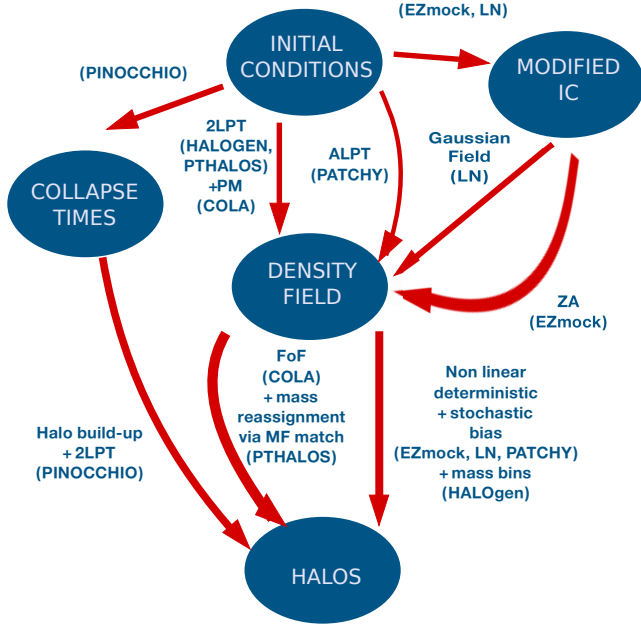


**Figure 1.** Clustering statistics in real and redshift-space for the reference BigMultiDark SO and FoF catalogues, both with the same number density. Monopole of the power spectrum in real space for BigMD SO catalogue (thick solid line) and FoF catalogue (thick dashed line); monopole of power spectrum 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.

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 dark matter 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 LogNormal perform a modification of the initial conditions (see Chuang et al. 2015, and Coles & Jones 1991 for more details).



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

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 2LPT evolution allows 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:

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

so that the equation of motion schematically reads

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

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. 2014 in preparation.

### 3.2 EZmock

EZmock (Effective Zel'dovich approximation mock catalogue, Chuang et al. 2015) is constructed from the Zel'dovich approximation density field. It absorbs the nonlinear effect and halo bias (i.e. linear, nonlinear, deterministic, and stochastic bias) into some effective modeling 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 dark matter density field on a grid using the Zel'dovich approximation (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

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

where  $\rho_s(\mathbf{r})$  and  $\rho_o(\mathbf{r})$  are the ZA density field after and before the scattering respectively.  $G(\lambda)$  is a random number drawn from the Gaussian distribution with width  $\lambda$ . 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

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

where  $\rho_{o'}(\mathbf{r})$  is the modified density,  $\rho_o(\mathbf{r})$  is the original ZA density,  $\rho_{th}^{low}$  and  $\rho_{th}^{high}$  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

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

where  $A$  is a free parameter;

(VI) fitting baryon acoustic oscillations (BAOs) by enhancing the amplitude of BAOs in the initial input power spectrum by

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

where  $P_{eBAO}(k)$  is the BAO enhanced power spectrum,  $P_{lin}(k)$  is the linear power spectrum,  $P_{nw}(k)$  is the smoothed no-wiggle power spectrum obtained by applying a cubic spline fit to  $P_{lin}(k)$ , 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 modeled by a random Gaussian distribution, i.e.,

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

where  $B$  is a constant corresponding to linear growth;  $\psi$  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  $\lambda'$ .

### 3.3 HALOGEN

The aim of HALOgen (Avila et al. 2014) is to provide a simple and efficient approximate method for generating mock halo catalogues with correct mass-dependent 2-point statistics. The basic algorithm is as follows:

	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
Initial conditions	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
# free params	0	7	1 (each mass bin)	–	7	5	1
# free params for z-space dist.	0	1	1	–	2	0	0
# free params for MF	0	–	adopt MF	–	–	5	adopt MF
# free params for bias	0	6	1	–	5	0	0

**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. in prep.). The post-processing can be used to assign mass and other mass related quantities, e.g. circular velocity. HALOgen constructs halo catalogues in mass bins; different initial condition codes are used to construct the dark matter density field for different methodologies; all the codes are using parallelisation 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.

(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  $l_{cell} \approx 2d_{part}$  (twice the mean-interparticle distance, for this comparison we used  $l_{cell} = 4 h^{-1}$  Mpc).

(IV) Distribute haloes into mass bins (for this comparison we use 8 bins), and for each bin  $M_j$  from highest to lowest mass, place each halo in the following way:

- Choose a cell with probability  $P(i|M_j) \propto \rho_i^{\alpha(M_j)}$ .
- Place the halo on a random 2LPT particle within the cell.
- Ensure the halo does not overlap previous halo centres (if so, repeat the cell choice).
- Decrease the mass of the cell by the mass of the halo (ensuring mass conservation on scales of  $l_{cell}$ ).

(V) Assign particle velocities to haloes with a factor  $\vec{v}_h = f_{vel} \cdot \vec{v}_p$ , computed as the ratio of the velocity dispersions of the selected particles and the reference halo catalog:  $f_{vel} = \frac{\sigma(\vec{v}_p)}{\sigma(\vec{v}_{ref})}$

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  $l_{cell}$ , and used for any number of random initial conditions. 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 Log-Normal

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 co-moving 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

$$\xi_G(r) = \ln[1 + \xi(r)], \quad (8)$$

and this can be Fourier transformed to the power spectrum,  $P_G(k)$ ,

(II) A Gaussian density field  $\delta_G(r)$  is generated on the grid with the power spectrum,  $P_G(k)$ ,

(III) A lognormal field is calculated by

$$\delta_{LN}(r) = \exp\left[\delta_G(r) - \frac{\sigma_G^2}{2}\right] - 1, \quad (9)$$

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

$$\sigma_G^2 = \sum_{i,j,l=1}^{N_{grid}} P_G\left[(k_{x_i}^2 + k_{y_j}^2 + k_{z_l}^2)^{\frac{1}{2}}\right], \quad (10)$$

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

(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. 2013), but it is not done in this study.

### 3.5 PATCHY

PATCHY (Kitaura et al. 2014a) relies on modeling 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 second order LPT (2LPT) in the high and low density regimes with a nonlinear local term derived from the spherical collapse model

(SC) matching  $N$ -body simulations. In this approximation the displacement field  $\Psi_{\text{ALPT}}(\mathbf{q}, z)$ , mapping a distribution of dark matter particles at initial Lagrangian positions  $\mathbf{q}$  to the final Eulerian positions  $\mathbf{x}(z)$  at redshift  $z$  ( $\mathbf{x}(z) = \mathbf{q} + \Psi(\mathbf{q}, z)$ ), is split into a long-range and a short-range component, given by 2LPT and SC, respectively:

$$\Psi_{\text{ALPT}}(\mathbf{q}, z) = \mathcal{K}(\mathbf{q}, r_S) \circ \Psi_{2\text{LPT}}(\mathbf{q}, z) + (1 - \mathcal{K}(\mathbf{q}, r_S)) \circ \Psi_{\text{SC}}(\mathbf{q}, z) \quad (11)$$

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

$$\rho_h = f_h \theta(\rho_M - \rho_{\text{th}}) \rho_M^\alpha \exp\left[-\left(\frac{\rho_M}{\rho_\epsilon}\right)^\epsilon\right], \quad (12)$$

with

$$f_h = \bar{N}_h / \langle \theta(\rho_M - \rho_{\text{th}}) \rho_M^\alpha \exp\left[-\left(\frac{\rho_M}{\rho_\epsilon}\right)^\epsilon\right] \rangle, \quad (13)$$

and  $\{\rho_{\text{th}}, \alpha, \epsilon, \rho_\epsilon\}$  the parameters of the model;

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

$$P(N_i | \rho_{hi}, \beta) = \frac{\lambda_i^{N_i}}{N_i!} \frac{\Gamma(\beta + N_i)}{\Gamma(\beta)(\beta + \rho_h)^{N_i}} \frac{1}{(1 + \rho_h/\beta)^\beta} \quad (14)$$

with  $\beta$  being the stochastic bias parameter.

(IV) The parameters are constrained to efficiently match the halo (or galaxy) probability distribution function (PDF) and the power spectrum for a given number density. In this way we can match the 3-point statistics,

(V) Peculiar velocities are split into a coherent and a quasi-virialised 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<sup>17</sup> Monaco et al. (2002, 2013) is based on the ellipsoidal collapse, solved with the aid of 3rd-order LPT, to compute the time at which mass elements collapse (in the orbit-crossing sense), and Extended Press & Schechter (EPS) to deal with multiple smoothing radii. It starts from the generation of a linear density field on a regular grid in Lagrangian space, in the same way as initial conditions 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 center 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 2-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 summarised as follows (Manera et al. 2012, 2015):

(I) Create a dark matter particle field based 2LPT.

(II) Identify halos using a Friends-of-Friends (FoF, Davis et al. 1985) halo finder with an appropriately chosen linking length. Alternatively, one can identify halos with Spherical Overdensity with an equivalent density threshold.

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

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

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

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

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

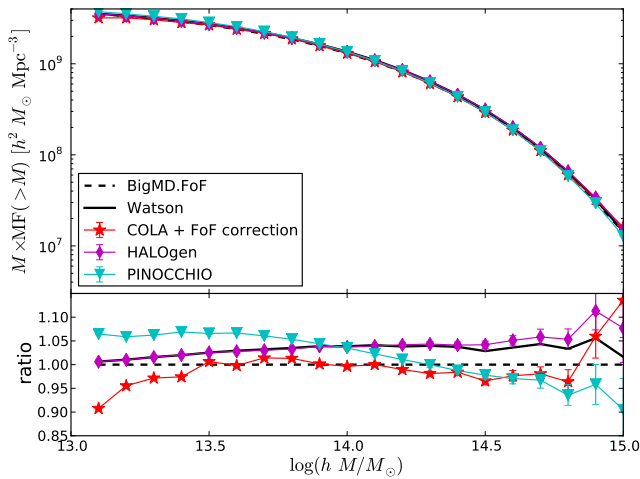
where

$$\Omega_m(z) = \Omega_m(1+z)^3 / H^2(z). \quad (17)$$

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

Scoccimarro & Sheth (2002) originally constructed a merger tree to assign halos masses in cells. This method adopts a mass function and imparts it to the rank-ordered halos found by the halo

<sup>17</sup> <http://adlibitum.oats.inaf.it/monaco/Homepage/Pinocchio/index.html>



**Figure 3.** Cumulative mass functions comparing with the BigMultiDark FoF reference catalog. The error-bars were estimated using Jack-knife resampling using 64 different sub-volumes. All the methods reproduce the numerical mass function to 5% accuracy.

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 initial condition directly. We compute the reduced white noise by averaging and rescaling the noise on the neighbor 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.**

### 4.1 Mocks for FoF catalogues

Here, we compare the different mocks with the BigMultiDark FoF reference catalog (see Sec. 2). The mesh size used for computing the statistics is  $960^3$  if applicable.

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

$$M = N m_p (1 - N^{-0.6}), \quad (18)$$

where  $N$  is the number of particles in the halo and  $m_p$  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% 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 2-point clustering statistics of FoF catalogues

2-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% at the scales between 10 to 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 has at large scales are mainly due to sample variance (COLA did not use the BigMultiDark white noise). For the quadrupole, COLA agrees with the BigMultiDark within 5% down to the minimum scale we measured (10  $h^{-1}$  Mpc); PINOCCHIO agrees within 10%; EZmock and PATCHY are within 15%.

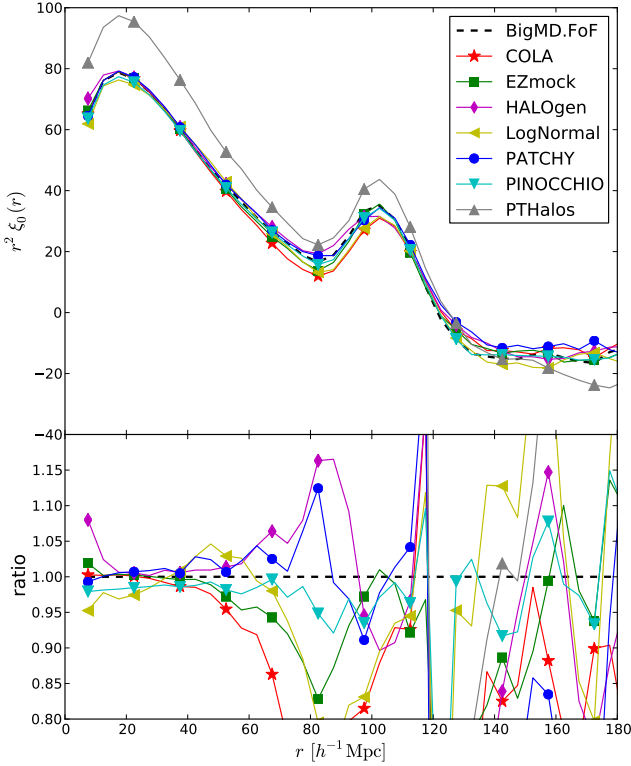
Although, theoretically, the power spectrum is simply a Fourier transform of the 2-point correlation function, the performance can be very different. The uncertainties at small scales in the configuration space will propagate to the relative large scales in Fourier space. Fig. 6 shows the monopole of the power spectrum in real-space. COLA, EZmock, and PATCHY agree with the simulation within 5% for all the scales. HALOgen, Log-Normal, and PINOCCHIO agree with the simulation within 10% up to  $k = 0.2 - 0.25 h \text{ Mpc}^{-1}$ . PTHalos has  $\sim 20\%$  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  $\sim 10$  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% for all the scales shown in the plot; for the quadrupole, COLA agrees with the simulation within 5% for all the scales; PINOCCHIO agrees within 10%; EZmock and PATCHY agree with the simulation within 15 – 20%. 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  $k > 0.15 h \text{ Mpc}^{-1}$ , 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 3-point clustering statistics of FoF catalogues

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

	BigMD	COLA	EZmock	HALOGEN	Log-normal	PATCHY	PINOCCHIO	PTHalos
Particle mesh size	3840 <sup>3</sup>	1280 <sup>3</sup> (3840 <sup>3</sup> for force)	960 <sup>3</sup>	1280 <sup>3</sup>	1280 <sup>3</sup>	960 <sup>3</sup>	1920 <sup>3</sup>	1280 <sup>3</sup>
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

**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 BigMultiDark 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 two or even more.



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

fixed sides,  $r_1 = 10 h^{-1} \text{Mpc}$  and  $r_2 = 20 h^{-1} \text{Mpc}$ , 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,  $k_1 = 0.1 h \text{Mpc}^{-1}$  and  $k_2 = 0.2 h \text{Mpc}^{-1}$ , and a varying angle  $\theta_{12}$ . 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 3-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 spherical overdensity catalogue (obtained using BDM code) from BigMultiDark 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 2-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 for the monopole and quadrupole of the correlation function in redshift-space. For the monopole, COLA+HOD shows some deviation at scales  $> 40 h^{-1} \text{Mpc}$ , which may be due to not using the BigMultiDark white noise. For the quadrupole, EZmock, PATCHY and PINOCCHIO+HOD agree with the simulation catalogue within 10% for all the scales considered. COLA+HOD agrees within 10% down to  $r = 15 h^{-1} \text{Mpc}$ .

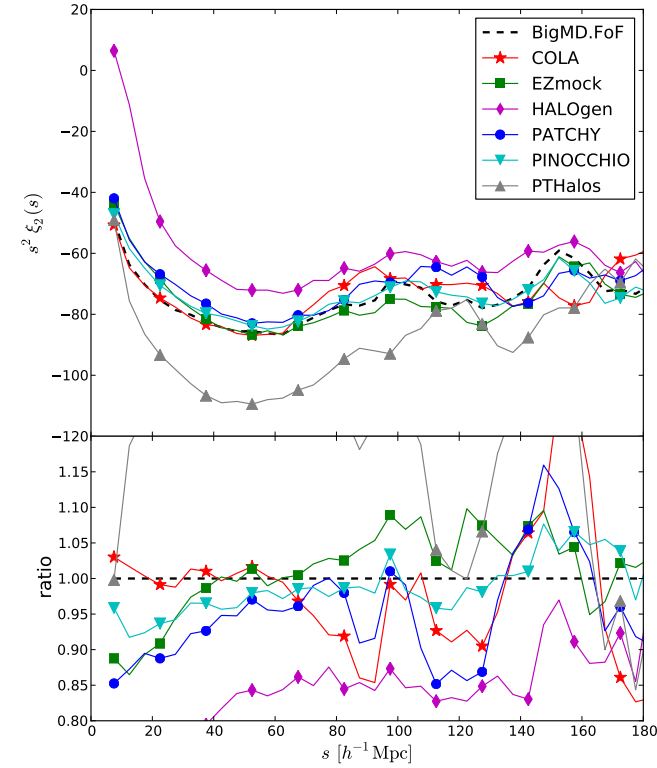
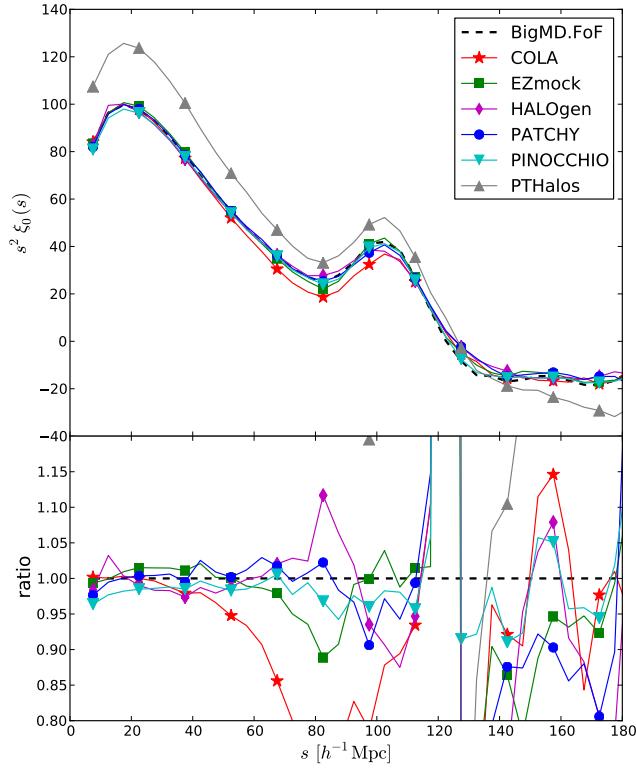
Fig. 11 shows the monopole of the power spectrum in real-space. EZmock and PATCHY agree with BigMultiDark within 5% for all the scales. COLA+HOD and HALOgen+HOD are within 10% up to  $k \sim 0.35 h \text{Mpc}^{-1}$ , and PINOCCHIO+HOD and Log-Normal are within 10% up to  $k \sim 0.1 h \text{Mpc}^{-1}$ . 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 BigMultiDark monopole within 10% for all the scales; and up to  $k \sim 0.1 h \text{Mpc}^{-1}$  for HALOgen+HOD and PINOCCHIO+HOD. For the quadrupole, EZmock and PATCHY agree with the simulation within 10% for all the scales; COLA+HOD and PINOCCHIO+HOD agree up to  $k = 0.25 h \text{Mpc}^{-1}$ .

As discussed in the Appendix, we test our HOD scheme by applying it to the SO distinct halos 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.

### 4.2.2 3-point clustering statistics of SO catalogues

Fig. 13 shows the bispectrum and 3-point correlation function in real space. The configurations are the same as for FoF catalogues.





**Figure 6.** FoF power spectrum comparison, in real space, between the different approximate methods and BigMultiDark.

For the 3-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%. For the bispectrum, COLA+HOD, EZmock, and PATCHY agree within 10-20% with the reference simulation catalogue. We conclude that an appropriate bias model is the key to reach high accuracy for the power spectrum and 3-point clustering statistics.

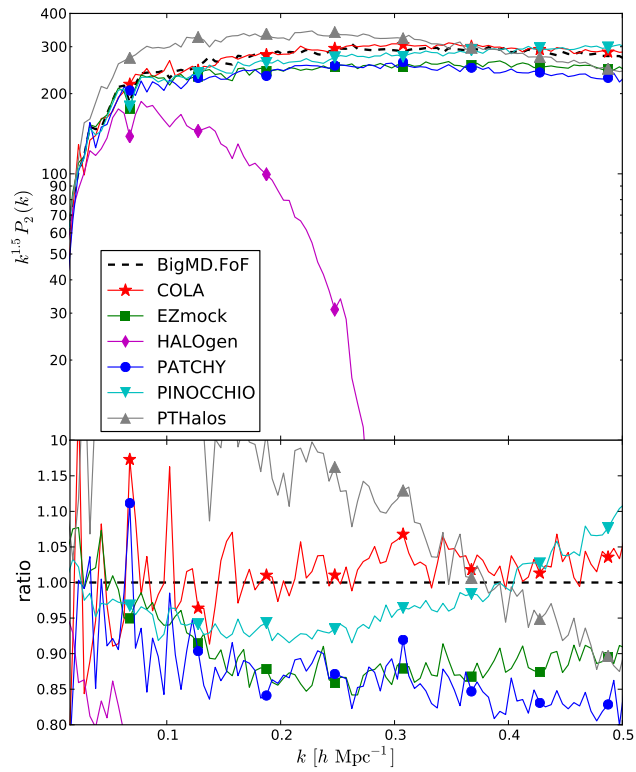
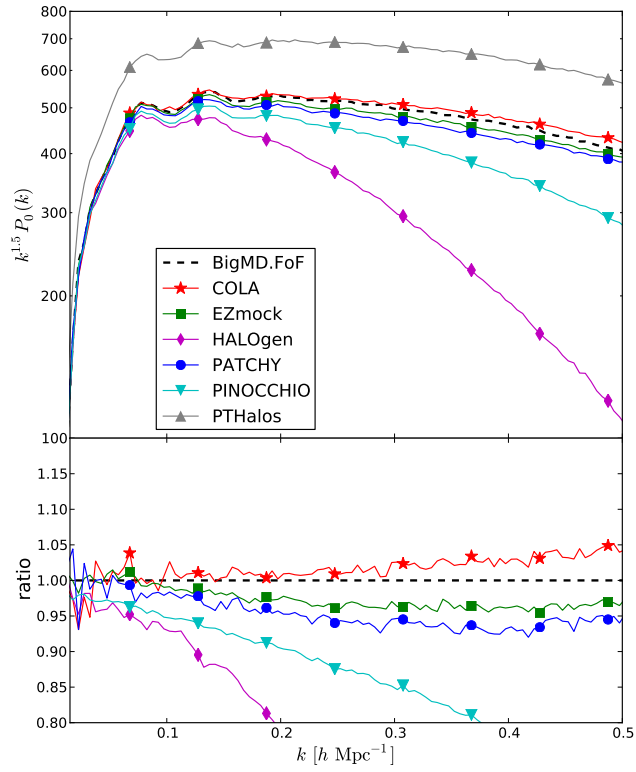
## 5 SUMMARY

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 BigMultiDark 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.

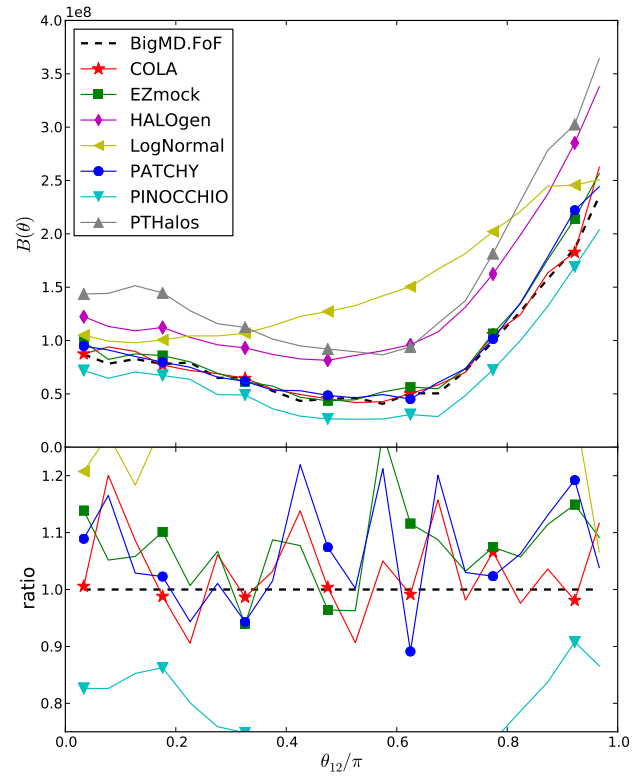
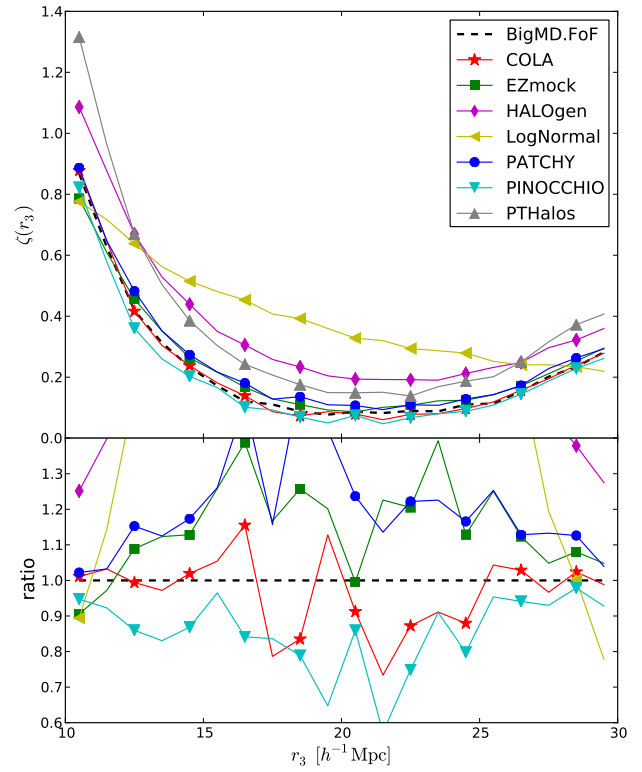
We are listing some items we have learned from this comparison study and have more discussion following the list.

- (I) Most of the methods are able to reproduce the 2-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 3-point statistics, including bispectrum and 3-point correlation function,
- (III) in redshift space, so far, only the semi- $N$ -body simulation, i.e. COLA, could reach high accuracy (1% level) at small scales,

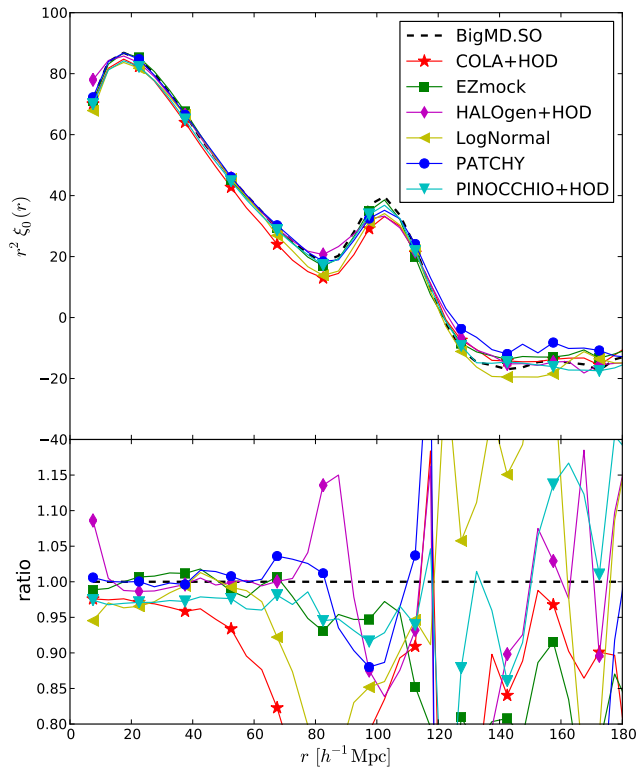
**Figure 5.** Top panel: comparison of the monopole of the correlation function in redshift space. Bottom panel: performance results for the quadrupole of the correlation function in redshift-space. Dashed lines correspond to the BigMultiDark FoF reference catalogue.



**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 BigMultiDark FoF reference catalogue.



**Figure 8.** Top panel: performance results for the 3-point correlation function in real-space. Bottom panel: bispectrum in real-space. Dashed lines correspond to the BigMultiDark FoF reference catalogue.

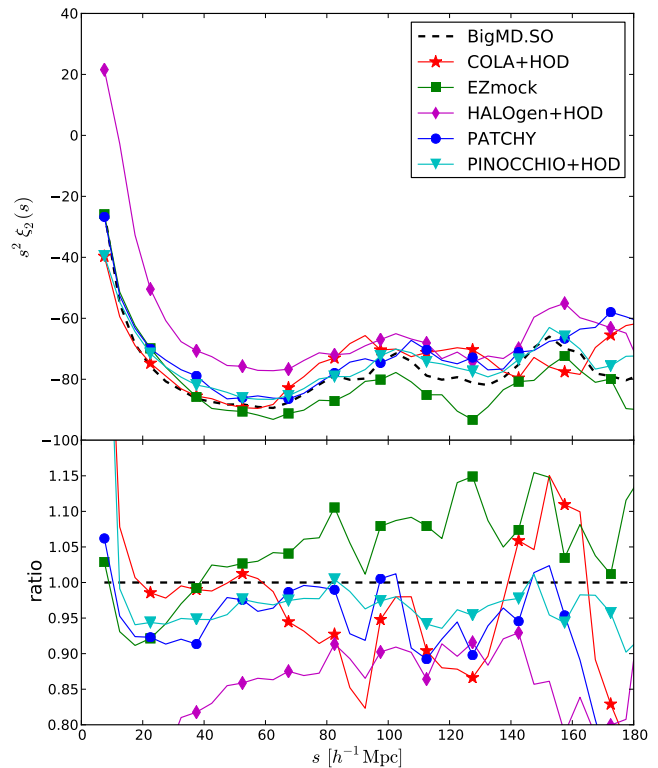
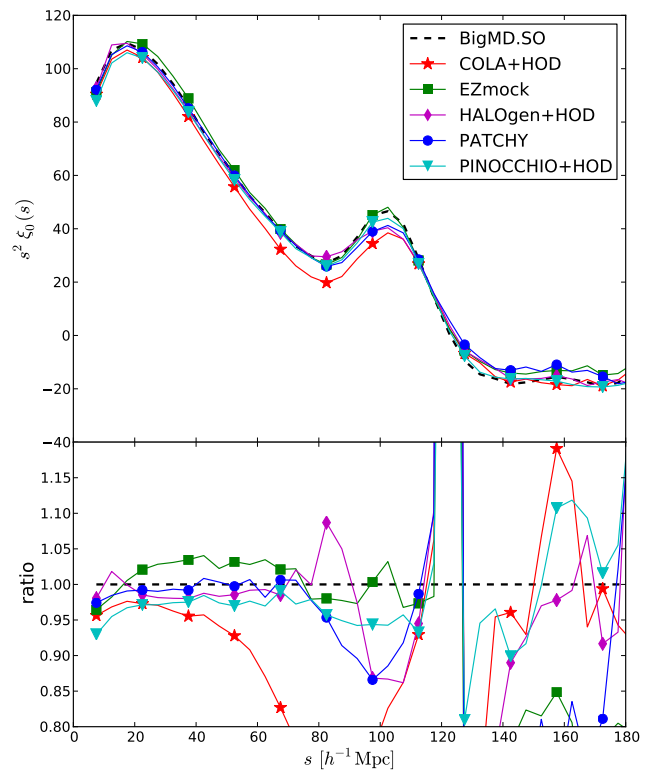


**Figure 9.** Comparison of the monopole of the correlation function in real-space. Dashed line corresponds to the BigMultiDark SO reference catalogue.

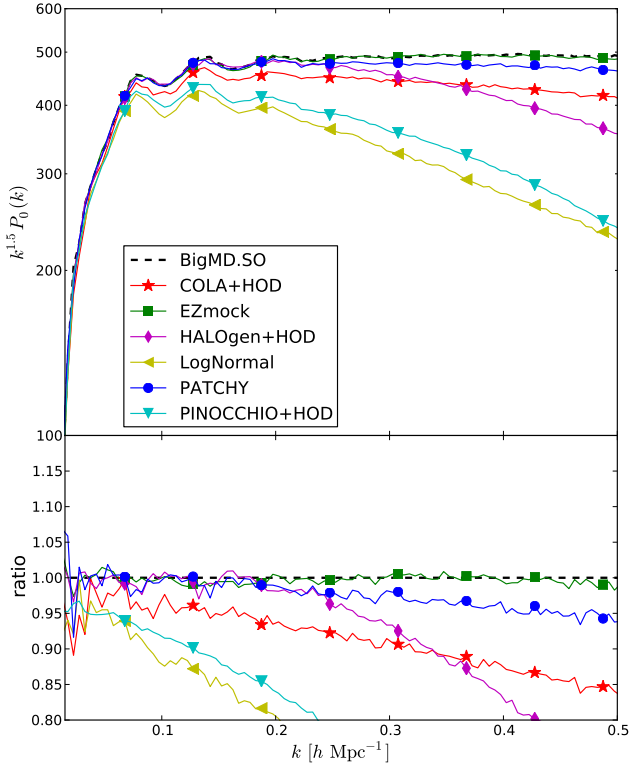
i.e.,  $r < 25 h^{-1} \text{ Mpc}$  or  $k > 0.15 h \text{ Mpc}^{-1}$ , on the quadrupole of the correlation function or the power spectrum.

(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 dark matter particles after cosmic evolution according to perturbative approaches show a typical uncertainty of roughly a few Mpc, depending on the chosen approximation (e.g., see Monaco et al. 2013; Kitaura & Hess 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  $ks$ . 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 dark matter 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



**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 BigMultiDark SO reference catalogue.

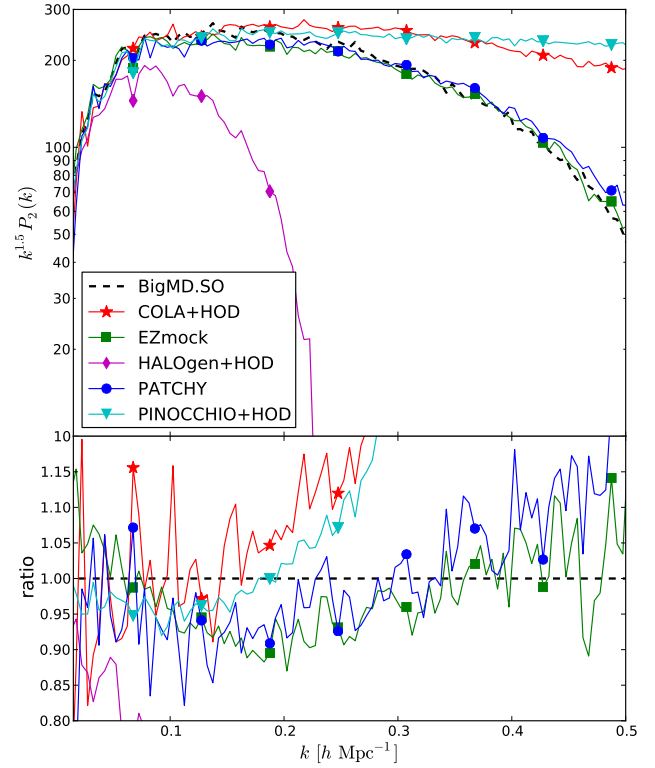
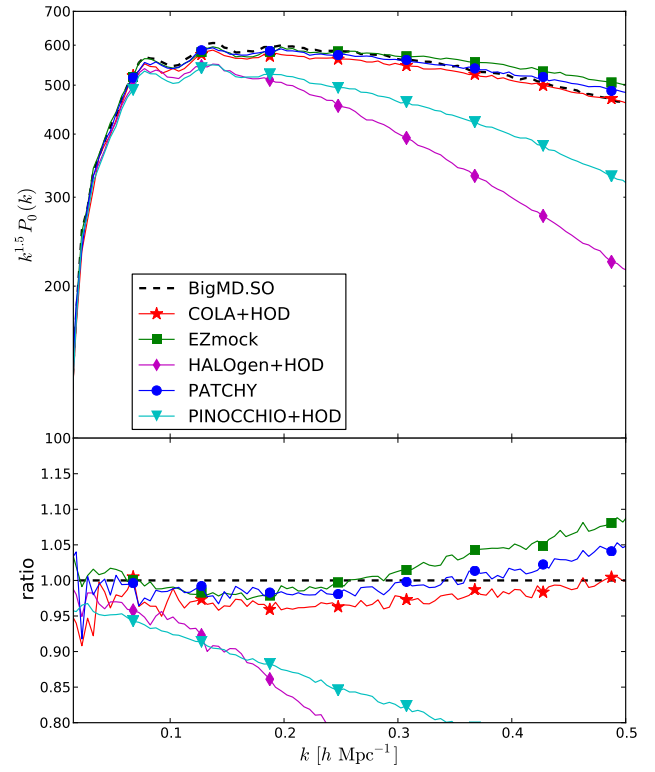


**Figure 11.** SO power spectrum comparison, in real space, between the different approximate methods and BigMultiDark.

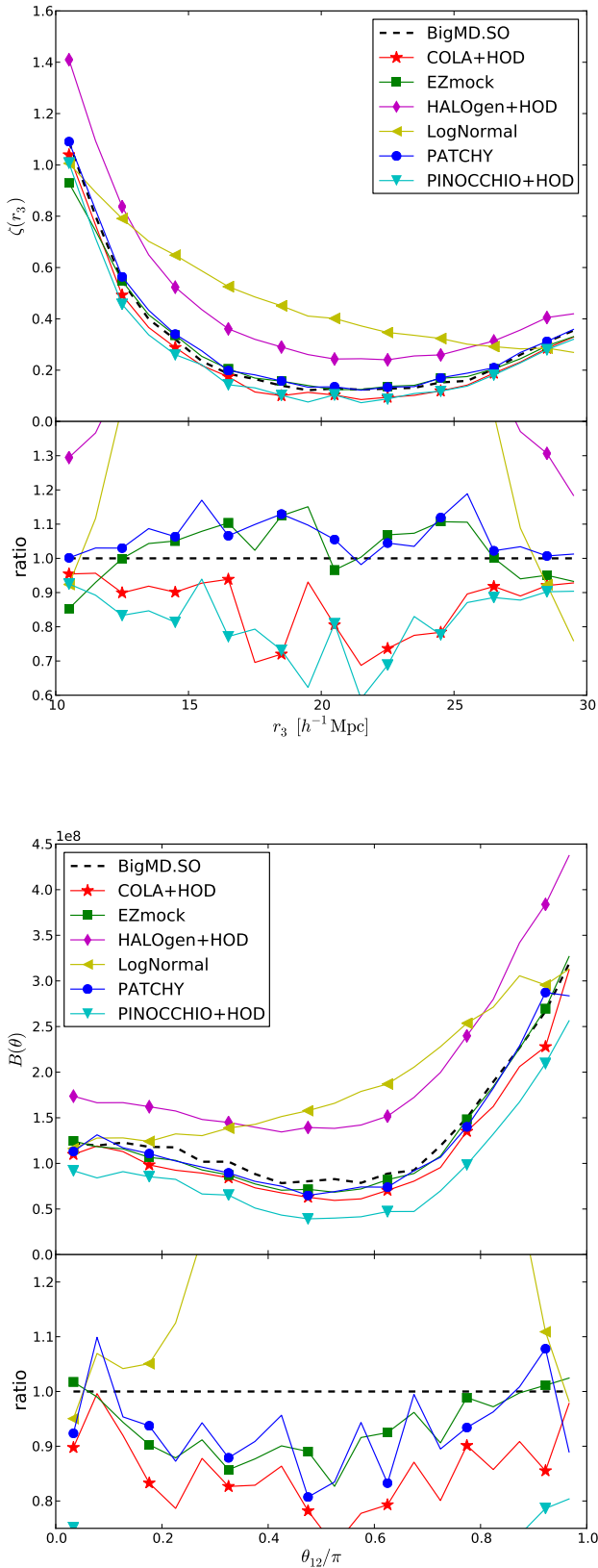
with additional prescriptions to fit the simulations, and introduce more complex relations.

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.

A HOD model is typically used to analyze 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 catalog, and there is no guarantee that the resulting mock catalog reproduces the expected clustering signal. In addition, if a model is calibrated only to the clustering length or bias (i.e., the 2-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 it's 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.



**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 BigMultiDark SO reference catalogue.



**Figure 13.** Top panel: performance results for the 3-point correlation function in real-space. Bottom panel: bispectrum in real-space. Dashed lines correspond to the BigMultiDark SO reference catalogue.

## 6 ACKNOWLEDGEMENT

The authors would like to express special thanks to the Instituto de Física Teórica (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<sup>18</sup> program 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 MICINN’s Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064 and AYA2010-21231-C02-01 grant, and Spanish MINECO’s Centro de Excelencia Severo Ochoa Programme under grant SEV-2012-0249. CZ and CT acknowledges support from Tsinghua University, and 973 program 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](http://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 have 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 were run in COSMA, the HPC facilities at the ICC in Durham, for a project granted by of the STFC DiRAC

<sup>18</sup> <http://www.mockingastrophysics.org>

HPC Facility ([www.dirac.ac.uk](http://www.dirac.ac.uk)). Some other computation 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 (EZ-mock), FK (PATCHY), MM (PTHalos), PM (PINOCCHIO), and SM (HALOgen). AK, AAK, CS, GY SG, and VT prepared the reference BigMultiDark simulation catalogues; EM developed and applied a HOD scheme for this study; VM and FM helped to develop the code for computing 3-point correlation function; other authors contributed towards the content of the paper and helped to proof-read it.

## APPENDIX A: ASSIGNING SUBHALOS WITH A 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 halo occupation distribution, to them augmenting their submitted catalogues with subhaloes.** The Halo Occupation Distribution (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).

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 BigMultiDark 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.

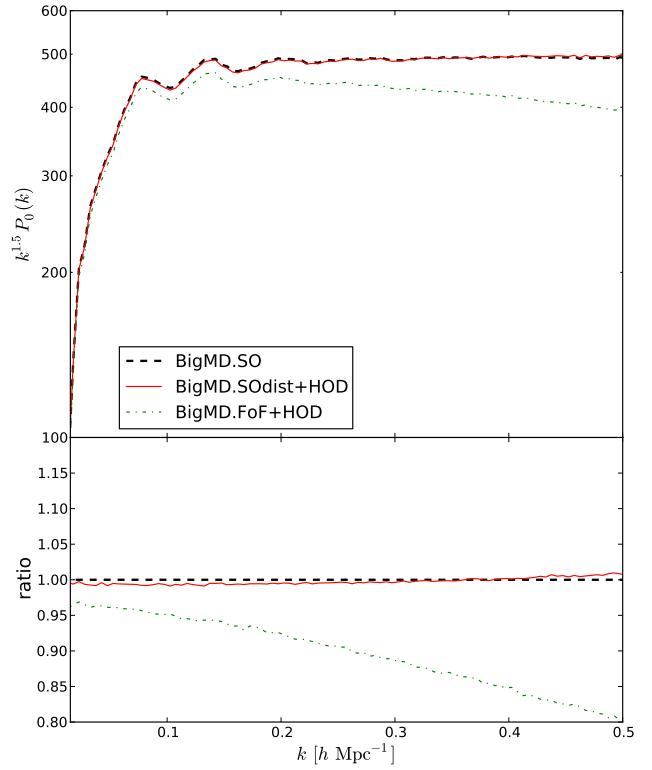
The following step consists in looking for a relation that associates the halo mass of the BigMultiDark with the average number of substructures in the halos of that mass.

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

It is now possible to populate the halos obtained with PINOCCHIO, COLA and HALOgen, with a population of substructures statistically identical to that of the BigMultiDark 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 halos with HOD scheme applied agrees with

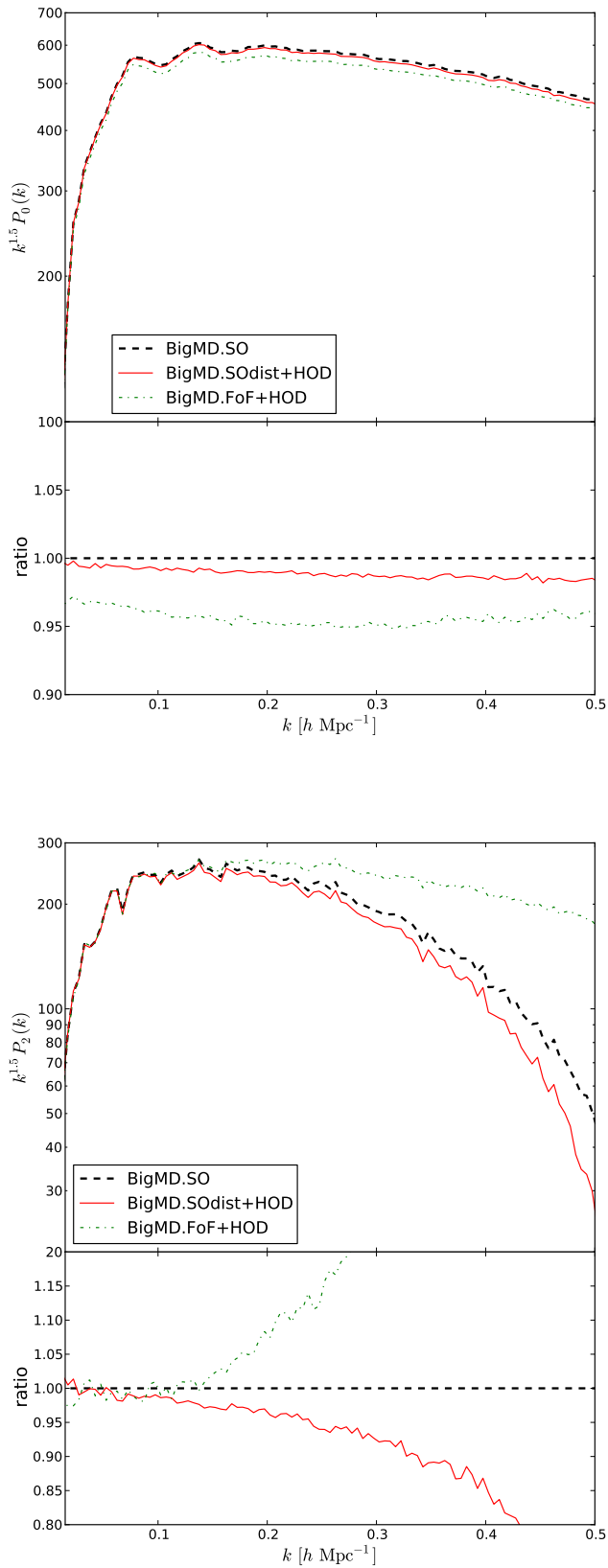


**Figure A1.** HOD Power spectrum comparison, in real space, among the BigMultiDark SO catalogue, SO distinct halos catalogue with HOD applied, and FoF catalogue with HOD applied.

the full BigMD SO catalogue very well. BigMD FoF catalog with HOD scheme applied has 5% 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 halos with HOD scheme applied agrees with the full BigMD SO catalogue very well; for quadrupole, it agree within 20% up to  $k = 0.4 h \text{ Mpc}^{-1}$ .

## REFERENCES

- Abazajian K. N., et al., 2009, *ApJS*, 182, 543, arXiv:0812.0649  
 Abell P. A., et al., 2009, 0912.0201  
 Avila S., G.Murray S., Knebe A., Power C., Robotham A. S., Garcia-Bellido J., 2014, 1412.5228  
 Benitez N., et al., 2014, 1403.5237  
 Berlind A. A., Weinberg D. H., 2002, *ApJ*, 575, 587, astro-ph/0109001  
 Beutler F., Blake C., Colless M., Jones D. H., Staveley-Smith L., et al., 2011, *MNRAS*, 416, 3017, 1106.3366  
 Blake C., Davis T., Poole G., Parkinson D., Brough S., et al., 2011, *MNRAS*, 415, 2892, 1105.2862  
 Bryan G., Norman M., 1998, *ApJ*, 495, 80, astro-ph/9710107  
 Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, *MNRAS*, 446, 2621, 1409.1124  
 Chuang C.-H., Wang Y., Hemantha M. D. P., 2012, *MNRAS*, 423, 1474, 1008.4822  
 Cole S., et al., 2005, *MNRAS*, 362, 505, astro-ph/0501174



**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.

- Coles P., Jones B., 1991, MNRAS, 248, 1  
 Colless M., et al., 2001, MNRAS, 328, 1039, astro-ph/0106498  
 Colless M., Peterson B. A., Jackson C., Peacock J. A., Cole S., et al., 2003, 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, 1208.0022  
 de Jong R. S., Bellido-Tirado O., Chiappini C., Depagne E., Haynes R., et al., 2012, 1206.6885  
 de la Torre S., Guzzo L., Peacock J., Branchini E., Iovino A., et al., 2013, A&A, 557, A54, 1303.2622  
 Drinkwater M. J., Jurek R. J., Blake C., Woods D., Pimblett K. A., et al., 2010, MNRAS, 401, 1429, 0911.4246  
 Eisenstein D. J., et al., 2011, AJ, 142, 72, 1101.1529  
 Frieman J. D. E. S. C., 2013, The Dark Energy Survey: Overview. Vol. 221, American Astronomical Society, AAS Meeting  
 Gardner J. P., Connolly A., McBride C., 2007, 0709.1967  
 Green J., Schechter P., Baltay C., Bean R., Bennett D., et al., 2012, 1208.4012  
 Hill G., Gebhardt K., Komatsu E., Drory N., MacQueen P., et al., 2008, ASP Conf.Ser., 399, 115, 0806.0183  
 Hubble E., 1934, ApJ, 79, 8, ADS  
 Kazin E. A., Koda J., Blake C., Padmanabhan N., 2014, MNRAS, 441, 3524, 1401.0358  
 Kitaura F., Angulo R., 2012, MNRAS, 425, 2443, 1111.6617  
 Kitaura F.-S., Gil-Marín H., Scoccola C., Chuang C.-H., Müller V., et al., 2014, 1407.1236  
 Kitaura F.-S., Hess S., 2013, MNRAS, 435, 78, 1212.3514  
 Kitaura F. S., Jasche J., Li C., Ensslin T. A., Metcalf R. B., et al., 2009, MNRAS, 400, 183, 0906.3978  
 Kitaura F.-S., Yepes G., Prada F., 2014, MNRAS, 439, L21, 1307.3285  
 Klypin A., Holtzman J., 1997, astro-ph/9712217  
 Klypin A., Yepes G., Gottlober S., Prada F., Hess S., 2014, 1411.4001  
 Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlober S., et al., 2004, ApJ, 609, 35, astro-ph/0308519  
 Laureijs R., et al., 2011, 1110.3193  
 Levi M., et al., 2013, 1308.0847  
 McBride C. K., Connolly A. J., Gardner J. P., Scranton R., Newman J. A., et al., 2011, ApJ, 726, 13, 1007.2414  
 Manera M., Samushia L., Tojeiro R., Howlett C., Ross A. J., et al., 2015, MNRAS, 447, 437, 1401.4171  
 Manera M., Scoccamarro R., Percival W. J., Samushia L., McBride C. K., et al., 2012, MNRAS, 428, 1036, 1203.6609  
 Monaco P., Sefusatti E., Borgani S., Croce M., Fosalba P., et al., 2013, MNRAS, 433, 2389, 1305.1505  
 Monaco P., Theuns T., Taffoni G., 2002, MNRAS, 331, 587, astro-ph/0109323  
 Nuza S., Sanchez A., Prada F., Klypin A., Schlegel D., et al., 2013, MNRAS, 432, 743, 1202.6057  
 Parkinson D., Riemer-Sorensen S., Blake C., Poole G. B., Davis T. M., et al., 2012, Phys.Rev., D86, 103518, 1210.2130  
 Percival W. J., et al., 2010, MNRAS, 401, 2148, 0907.1660  
 Pope A. C., Szapudi I., 2008, MNRAS, 389, 766, 0711.2509  
 Reid B. A., Percival W. J., Eisenstein D. J., Verde L., Spergel D. N., et al., 2010, MNRAS, 404, 60, 0907.1659  
 Riebe K., Partl A. M., Enke H., Forero-Romero J., Gottlober S., et al., 2011, 1109.0003  
 Samushia L., Percival W. J., Raccanelli A., 2012, MNRAS, 420, 2102, 1102.1014  
 Schlegel D., et al., 2011, 1106.1706

- Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629, astro-ph/0106120
- Skibba R. A., Sheth R. K., 2009, MNRAS, 392, 1080, 0805.0310
- Springel V., 2005, MNRAS, 364, 1105, astro-ph/0505010
- Tassev S., Zaldarriaga M., 2012, JCAP, 1204, 013, 1109.4939
- Tassev S., Zaldarriaga M., Eisenstein D., 2013, JCAP, 1306, 036, 1301.0322
- Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, ApJ, 646, 881, astro-ph/0506395
- Watson W. A., Iliev I. T., D'Aloisio A., Knebe A., Shapiro P. R., et al., 2013, MNRAS, 433, 1230, 1212.0095
- White M., 2014, MNRAS, 439, 3630, 1401.5466
- White M., Tinker J. L., McBride C. K., 2013, 1309.5532
- Wild V., et al., 2005, MNRAS, 356, 247, astro-ph/0404275
- York D. G., et al., 2000, AJ, 120, 1579, astro-ph/0006396
- Zehavi I., et al., 2011, ApJ, 736, 59, 1005.2413
- Zhao C., Kitaura F.-S., Chuang C.-H., Prada F., Yepes G., et al., 2015, 1501.05520
- Zheng Z., Berlind A. A., Weinberg D. H., Benson A. J., Baugh C. M., et al., 2005, ApJ, 633, 791, astro-ph/0408564