

# EZmocks: extending the Zel’dovich approximation to generate mock galaxy catalogues with accurate clustering statistics

Chia-Hsun Chuang<sup>1\*</sup>, Francisco-Shu Kitaura<sup>2†</sup>, Francisco Prada<sup>1,3,4</sup>, Cheng Zhao<sup>5</sup>, Gustavo Yepes<sup>6</sup>,

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

<sup>2</sup> *Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany*

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

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

4 September 2014

## ABSTRACT

We present a new methodology to generate mock halo or galaxy catalogues, which have accurate clustering properties, nearly indistinguishable from full  $N$ -body solutions, in terms of the one-point, two-point, and three-point statistics. In particular, the agreement is remarkable, within 1% up to  $k = 0.65 \ h \text{ Mpc}^{-1}$  and down to  $r = 10 \ h^{-1} \text{ Mpc}$ , for the power spectrum and two-point correlation function respectively, while the bispectrum agrees in general within 20% for different scales and shapes. Our approach is based on the Zel’dovich approximation, however, effectively including with the simple prescriptions the missing physical ingredients, and stochastic scale-dependent, non-local and nonlinear biasing contributions. The computing time and memory required to produce one mock is similar to that using the log-normal model. With high accuracy and efficiency, the effective Zel’dovich approximation mocks (EZmocks) provide a reliable and practical method to produce massive mock galaxy catalogues for the analysis of large-scale structure measurements.

**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 decade. 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) is surveying 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).

Mock galaxy catalogues are essential for analysing the cluster-

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

† E-mail: kitaura@aip.de, Karl-Schwarzschild-Fellow

<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>

ing signal drawn from these surveys. The most reliable technique for creating mock catalogues is  $N$ -body cosmological simulations, 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; Chuang & Wang 2012, 2013b,a; Samushia et al. 2012). However, the total run-time and memory required to generate a large suit of simulations make this effort prohibited in most of the cases, hence their use for ongoing and future surveys is impractical. Recent techniques permit to speed up  $N$ -body codes (see, e.g., COLA<sup>15</sup>; Tassev et al. 2013), however, the memory requirements are still large with these methods.

Alternatively, log-normal mock catalogues (Coles & Jones 1991) have been widely used (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 model produces by construction mock catalogues with precise two-point statistics, transforming Gaussian density fields with a given variance. Yet, its co-moving description neglects the relative displacement of matter through cosmic evolution, and does not capture the spatial pattern of the cosmic web. Hence, it leads to large deviations in the higher-order statistics of the large-scale structure.

More realistic mock catalogues can be produced describing the gravity induced motion of matter with perturbation theory. Two codes have pioneered the field based on this technique, namely PTHalos (Scoccimarro & Sheth 2002) and PINOCCHIO<sup>16</sup> (Monaco et al. 2002, 2013). More recently, PTHalos has been applied for the analysis of the BOSS galaxy survey (Manera et al. 2012, 2014). The accuracy of these methods is limited by the approximate solutions of perturbation theory, unable to model the nonlinear regime towards small scales, and hence, to accurately resolve the formation of haloes hosting the galaxies. To overcome this problem, recent approaches, relying on a statistical description of the halo distribution have been introduced. These methods require only the large scale density field and therefore can rely on approximate gravity solvers, either based on improved perturbation theory algorithms (see PATCHY<sup>17</sup>; Kitaura et al. 2013, 2014), or on approximate particle mesh solutions (see QPM<sup>18</sup>; White et al. 2013). This density field is populated with haloes or galaxies in a posterior step. We want to further explore this approach, reducing the computational requirements to a minimum, albeit without loosing accuracy.

In this work, we develop a new methodology to generate mock halo or galaxy catalogues, which have accurate clustering properties, nearly indistinguishable from full  $N$ -body solutions, in terms of the one-point, two-point, and three-point statistics. Our approach is based on the Zel'dovich approximation (ZA, Zel'dovich 1970), however, effectively including stochastic scale-dependent, non-local and nonlinear biasing contributions.

The computation of the effective Zel'dovich approximation mock catalogues (EZmocks, hereafter) demands a minimum run-time, i.e., three fast Fourier transform (FFT) to compute the displacement field in three directions and populate haloes nearby each grid point with some random assignment process, and a minimum memory requirement, i.e., a few arrays of the same size as the grid

used by FFT to store the information of the displacement field and the number of generated haloes. The ongoing and upcoming large galaxy surveys will demand precise mock catalogues scanning huge parameter space. Simple and efficient, but accurate methods will be required to cover those needs. While this year celebrates the 100<sup>th</sup> anniversary of the birth of Yakov Zel'dovich, we find that the Zel'dovich approximation can still provide one of the most powerful tools for analysing the large-scale structure survey data.

This paper is organised as follows. In Section 2, we describe the BigMultiDark  $N$ -body simulations and the computational facilities used for this study. In Section 3, we provide the details of the methodology to construct the EZmocks. In Section 4, we show its performance by comparing with  $N$ -body simulations. We summarise and conclude in Section 5.

## 2 REFERENCE SIMULATIONS AND COMPUTATIONAL REQUIREMENTS

To test our method, we use a reference halo catalogue at redshift  $z = 0.5618$  extracted from one of the BigMultiDark (BigMD) simulations<sup>19</sup> (Klypin et al. in prep.), which was performed using GADGET-2 Springel (2005) with 3840<sup>3</sup> 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  $h = 0.6777$ . Haloes were defined based on two different algorithms. One identifies density peaks including substructures using the Bound Density Maximum (BDM) halo finder (Klypin & Holtzman 1997; Riebe et al. 2011) and the other one group up the particles using a friend-of-friend (FOF) halo finder. In this work, we use the FOF catalogue as our reference. 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  $z \sim 0.5$ . The MultiDark simulations have been used to interpret the clustering of the BOSS galaxy sample (e.g., see Nuza et al. 2013).

We are using the thin-nodes of the Curie supercomputer<sup>20</sup> to generate EZmocks. It has 16 cores (2.7GHz) with 64GB of memory per node (we only need less than 32GB in this study) and 5040 nodes in total. We use shared memory multiprocessing (i.e., OpenMP) to speed up the computation. It takes less than 5 minutes to construct one EZmock with the grid size of 960<sup>3</sup> using one node. In other words, it takes 10 minutes to generate 10,000 EZmocks using full power of the machine. This make EZmock really competitive as compared to other approximation methods mentioned above, which certainly demand more computational resources.

## 3 METHODOLOGY

In this section, we introduce the methodology to construct EZmocks. The ZA yields a crude approximation of the dark matter density field on small scales, which cannot be directly used to compute the halo or galaxy catalogue. Therefore, the different components, which are not modelled by the ZA, have to be added in subsequent steps. The main assumption we test in this study, is that the ZA is sufficiently accurate on large scales to account for the three dimensional shape of the cosmic web. We aim at including

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

<sup>15</sup> COLA (COmoving Lagrangian Acceleration simulation)

<sup>16</sup> PINOCCHIO (PINpointing Orbit-Crossing Collapsed Hierarchical Objects)

<sup>17</sup> PATCHY (PerturbAtion Theory Catalog generator of Halo and galaxy distributions)

<sup>18</sup> QPM (quick particle mesh)

<sup>19</sup> <http://www.multidark.org/MultiDark/>

<sup>20</sup> <http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm>

with simple prescriptions the missing physical ingredients, such as tidal fields (included in second order LPT), and nonlinear, non-local stochastic biasing. 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;
- (IV) fitting the amplitude of the power spectrum and bispectrum with a density threshold and saturation;
- (V) fitting the shape of the final power spectrum by modifying the tilt in the initial input power spectrum with a scale-dependent function;
- (VI) fitting baryon acoustic oscillations (BAOs) by enhancing the amplitude of BAOs in the initial input power spectrum;
- (VII) computing the velocity field within the ZA for each object.

The steps listed above have a physical analogy. While the scatter in step III emulates stochastic biasing, the combination of PDF mapping (step II), thresholding (step IV) and tilting of the power spectrum (step V) mimics the effects of nonlinear, deterministic biasing. Non-local effects, such as tidal fields not included in linear LPT or other biasing contributions, are effectively included in both, the scatter relation and the tilting of the initial power spectrum. The missing power towards large  $k$ s (small scales) of perturbative approaches, is included in the modulation of the initial power spectrum when fitting the resulting halo population. In this way, multiple effects are accounted for in a single procedure.

We note that the approach presented in this work focuses on fitting the clustering statistics with direct and explicit prescriptions, circumventing the problem of finding the parameters to a particular biasing model, like it is done for instance in the PATCHY code. This makes the parameter finding in EZmocks much more efficient at the expense of giving up more explicit physical relations.

### 3.1 Generation of the Zel'dovich density field

A particle located at Lagrangian position  $\mathbf{q}$  will be mapped to its Eulerian position  $\mathbf{x}$  at cosmic time  $t$  by the displacement field  $\Psi(\mathbf{q}, t)$ , i.e.,

$$\mathbf{x}(\mathbf{q}, t) = \mathbf{q} + \Psi(\mathbf{q}, t). \quad (1)$$

The first order Lagrangian perturbation theory solution to the equations of motions is given by the Zel'dovich approximation (for a review, see, e.g., Bernardeau et al. 2002). The displacement field in the ZA is given by

$$\Psi(\mathbf{q}) = \int \frac{d^3k}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{q}} \frac{i\mathbf{k}}{k^2} \hat{\delta}(\mathbf{k}), \quad (2)$$

where  $\hat{\delta}(\mathbf{k})$  is the fractional density perturbation in Fourier-space. We construct the displacement field using the ZA to the same redshift as the reference catalog, i.e.,  $z = 0.5618$ . The adopted grid size,  $960^3$ , is a factor of 64 smaller, as compared to the one used by the BigMD simulation. We have tested that smaller grid sizes than the one presented in this study significantly damp the BAO peak. We have also verified that we get the same results with larger grid sizes, however, at a higher computational cost. Therefore, we consider that the resolution chosen in this study is very close to be the optimal one. We use the reduced white noise from the BigMD simulation to reduce the cosmic variance throughout this work. The

density field is obtained using the cloud-in-cells particle assignment scheme (CIC, e.g., Hockney & Eastwood 1981).

### 3.2 PDF mapping scheme

In this step, we map the probability distribution function measured from the halo catalog extracted from the  $N$ -body simulation to that from the ZA density field based on a rank ordering procedure. We note that this step is done on the halo field rather than on the dark matter field, and is not sufficient to solve the damping of power in perturbative approaches alone (e.g., see Leclercq et al. 2013).

Our rank ordering procedure is as follows:

- (a) computing the halo density field from BigMD with CIC;
  - (b) generating an integer on each grid point using a Poisson random number generator based on the density computed from step (a);
  - (c) assigning these integers to the grid of our ZA density field by rank ordering;
  - (d) populating the mock catalogue following a CIC distribution.
- The number of haloes populated on each grid point is the one assigned from step (c). For each halo, the CIC distribution can be implemented by

$$\Delta x = \begin{cases} (1 - \sqrt{R}) \times \text{cell-size} & \text{if } R \geq 0; \\ (-1 + \sqrt{-R}) \times \text{cell-size} & \text{if } R < 0, \end{cases} \quad (3)$$

where  $\Delta x$  is the distance from the halo to the grid point in  $x$  direction,  $R$  is a random number drawn between  $-1$  and  $1$ , and cell-size is the separation of the neighboring grid points, i.e.  $2500/960 h^{-1}$  Mpc in this work.  $\Delta y$  and  $\Delta z$  are determined with different draws of  $R$ .

### 3.3 Adding scatter to the PDF mapping

This step models the uncertainty in the PDF relation and hereby also the stochasticity of the tracers. While mapping the PDF with strict rank ordering, one finds that the amplitude of the power spectrum of the mock catalog is too high. Therefore, we introduce a scatter to the ZA density field by

$$\rho_s(\mathbf{r}) = \begin{cases} \rho_o(\mathbf{r})(1 + G(\lambda)) & \text{if } G(\lambda) \geq 0; \\ \rho_o(\mathbf{r}) \exp(G(\lambda)) & \text{if } G(\lambda) < 0, \end{cases} \quad (4)$$

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. To simplify the problem, we fix  $\lambda = 10$ . The choice of  $\lambda$  will affect the values of the other parameters in our models but does not affect the performance of the mock catalogue.

### 3.4 Density threshold and saturation

We apply a density threshold and a density saturation before the scattering scheme described in Sec. 3.3 by

$$\rho_{o'}(\mathbf{r}) = \begin{cases} 0, & \text{if } \rho_o(\mathbf{r}) < \rho_{\text{th}}^{\text{low}}; \\ \rho_{\text{th}}^{\text{high}}, & \text{if } \rho_o(\mathbf{r}) > \rho_{\text{th}}^{\text{high}}, \end{cases} \quad (5)$$

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

One can adjust three-point statistics (e.g., bispectrum) by tuning the density threshold. An effective field theory description of the large-scale structure has been studied to constrain the dark matter three-point statistics (see Angulo et al. 2014; Baldauf et al. 2014). Thereafter, Kitaura et al. (2014) found that a density threshold is essential to accurately model the halo three-point statistics. The density saturation can be used to adjust the amplitude of the power spectrum. It can be considered as modifying the weights of the scatter for different density regions, i.e., the density regions above the density saturation will have the same weight during the scatter process. While the scatter parameter for the PDF mapping can also be used to adjust the amplitude of the power spectrum, we find that the amplitude of the power spectrum is more sensitive to the density saturation. Therefore, we fix the scatter parameter to a constant and vary the density saturation only.

### 3.5 Nonlinear effects correction

Nonlinear evolution of the dark matter density field and halo bias have partially been accounted for by the steps described above. We correct for residual nonlinear effects by modifying the input power spectrum. In addition, our approximative method might introduce uncertainties of a few Mpc in the position of haloes, which will reduce the power at small scales and the signal of baryon acoustic oscillations (BAO). The correction includes enhancing the BAO feature, and overall power spectrum at small scales.

The BAO is enhanced by

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

where  $P_{\text{eBAO}}(k)$  is the BAO enhanced power spectrum,  $P_{\text{lin}}(k)$  is the linear power spectrum,  $P_{\text{nw}}(k)$  is the smoothed no-wiggle power spectrum obtained by applying a cubic spline fit to  $P_{\text{lin}}(k)$ , and  $k_*$  is usually known as the damping factor (however, for the damping model, one should use  $\exp(-k^2/k_*^2)$  instead).

We also enhance the power spectrum at small scales by

$$P_{\text{ePK}}(k) = P_{\text{eBAO}}(k) \cdot (1 + Ak), \quad (7)$$

where  $A$  is a free parameter.

### 3.6 Peculiar velocity

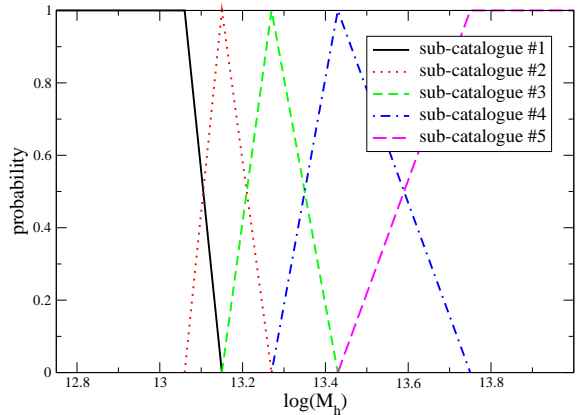
We model peculiar motions  $v$  by adding to the linear coherent motion, which is proportional to the ZA displacement field, a dispersion term modeled by a random Gaussian distribution, i.e.,

$$v_i(\mathbf{r}) = B\psi_i(\mathbf{r}) + G(\lambda'), \quad (8)$$

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  $\lambda'$ .

### 3.7 Mass assignment

We describe here how to assign masses to the distribution of tracers obtained in the previously described procedure with a simple additional step. First, we divide the BigMD halo catalogue into multiple catalogues according to the halo masses. Instead of taking the traditional mass bins (sharp mass cuts), we divide the sample in a statistic way. A halo will be placed into one of the sub-catalogues with certain probability as shown in Fig. 1. For example, a halo with  $10^{13.6}$  solar mass has about half a chance to be collected by



**Figure 1.** The probability of each halo in the BigMD catalogue to be collected by one of the five sub-catalogues, represented by different colours.  $M_h$  is the halo mass in units of solar mass. These sub-catalogues are prepared to construct the EZmock including masses. The divisions are chosen, so that the sub-catalogues have a comparable number of haloes.

sub-catalogue #4. If it is not collected by sub-catalogue #4, it will be collected by sub-catalogue #5. For each sub-catalogue, we construct an EZmock catalogue that will reproduce the clustering of each sub-catalogue. For each EZmock, we assign the halo mass with the value randomly selected from the sub-catalogue to which it corresponds. We demonstrate in the next section that the combined EZmock catalogue comprising all the sub-catalogues for different mass bins smoothly depends on arbitrary mass cuts.

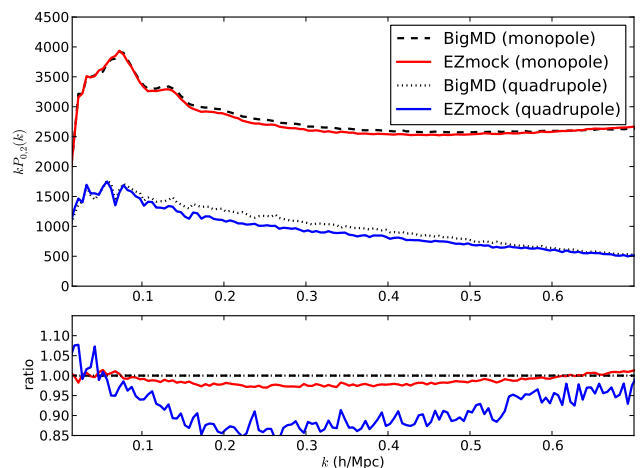
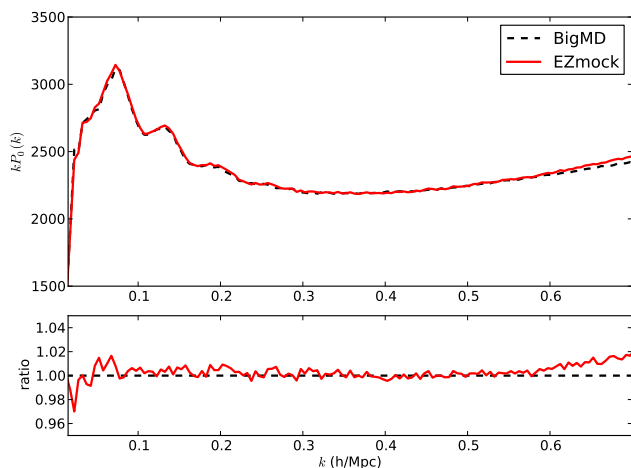
## 4 VALIDATION OF THE METHOD

In this section we validate our method comparing different statistics between EZmock and BigMD. The statistics include probability distribution function, power spectrum and two-point correlation function (monopole and quadrupole), bispectrum, and mass function.

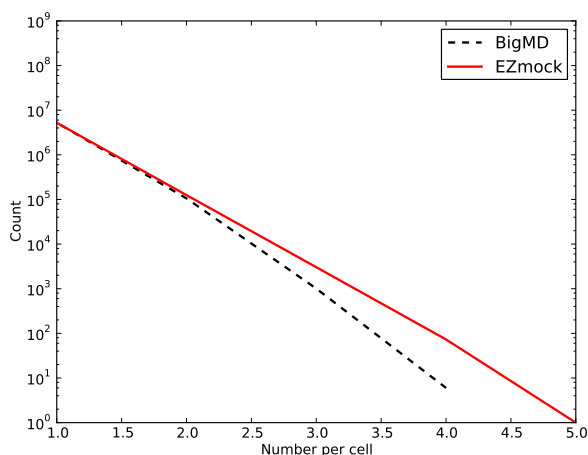
### 4.1 Probability distribution function

The PDF mapping ensures that the integral of the PDF, i.e., the number density, is fitted by construction. It also encodes information on the higher-order correlation functions (see, e.g., Kitaura et al. 2014). Therefore, an accurate mock should also reproduce the shape of the PDF. We note, however, that it is difficult to perfectly match the tail of the PDF towards large numbers of haloes due to the small number statistics in such bins. Fig. 2 shows the PDFs of EZmock and BigMD. The PDFs are computed using nearest-grid-point (NGP) with a grid of  $960^3$ .

Since the haloes are populated using CIC (see Sec. 3.2) to generate a continuous halo number density field, it is not guaranteed that the reference PDF is restored in the PDF mapping. Nevertheless, the PDF is sufficiently matched to ensure that the higher-order statistics is accurately reproduced, as we demonstrate below.



**Figure 3.** Left panel: monopole of the power spectrum in real space. Right panel: monopole and quadrupole of the power spectrum in redshift-space.



**Figure 2.** Halo probability distribution functions for EZmock (solid red line) and BigMD (dashed black line). The PDFs are computed using nearest-grid-point (NGP) with a grid of  $960^3$ .

## 4.2 Power spectrum and two-point correlation function

One of the main goals of mock galaxy catalogues is to accurately reproduce the theoretically expected two-point statistics (power spectrum in Fourier-space), including cosmic variance, i.e., the uncertainties due to the particular realisation of seed perturbations. When using Gaussian or log-normal realisations, one is able to impose an arbitrary power spectrum.

However, modelling gravitational evolution with perturbation theory introduces a bias, due to the approximate modelling of the nonlinear regime with respect to full gravity solutions. This is seen in the power spectrum as a damping of the small scale structures towards high  $k$ s. Therefore, some works suggested to use a transfer function connecting the full with the perturbative solution (see Tashev & Zaldarriaga 2012a,b). Yet, this transfer function introduces ringing effects in the fields coming from Fourier-space deconvolution effects and hence, it cannot be applied to produce mock catalogues. It was therefore suggested in Kitaura et al. (2013) to include

the perturbation theory induced damping effect in the bias. We follow this approach here obtaining remarkable results..

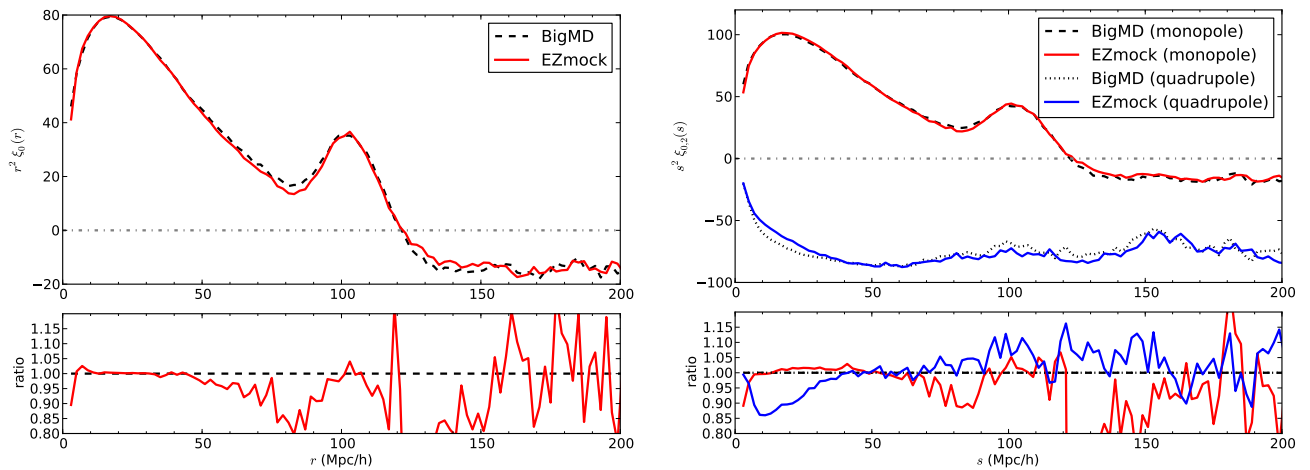
Fig. 3 (left panel) shows the comparison of the monopole of the power spectrum between the EZmock and BigMD in real-space. They agree within 1% accuracy in the scale range up to  $k = 0.65 \, h \, \text{Mpc}^{-1}$ . Since we generate the EZmock halo catalogues with a finite grid ( $960^3$ ), it is not meaningful to compare the power spectrum at the scales which are close to the Nyquist frequency, i.e., at  $k > 0.7 \, h \, \text{Mpc}^{-1}$ . It would be interesting to explore the validity of our methodology at smaller scales, however, such a study is out of the scope of this paper.

Fig. 3 (right panel) shows the monopole and quadrupole of the power spectrum in redshift-space. Our simple model (linear + Gaussian, see Sec. 3.6) to assign velocities to haloes has a limited performance, yielding a worse fit than in real-space. However, while there is still some room for improvement, it should be already good enough for most of practical applications, as the Kaiser factor is well recovered.

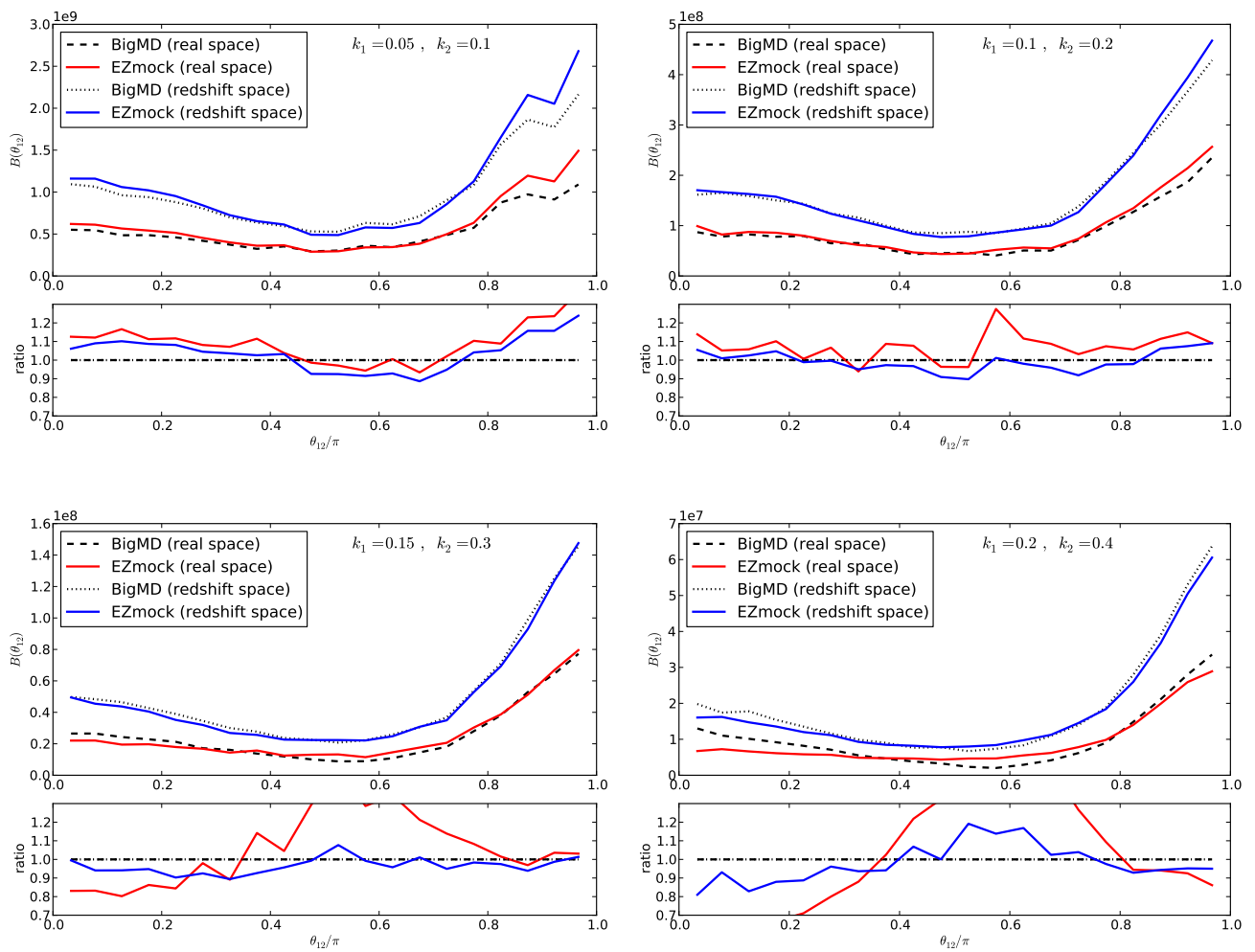
This can be seen in Fig. 4, where the monopole and quadrupole of the two-point correlation function in real and redshift-space are shown. The correlation function is restored down to  $10 \, h^{-1} \, \text{Mpc}$  within 1%.

## 4.3 Bispectrum

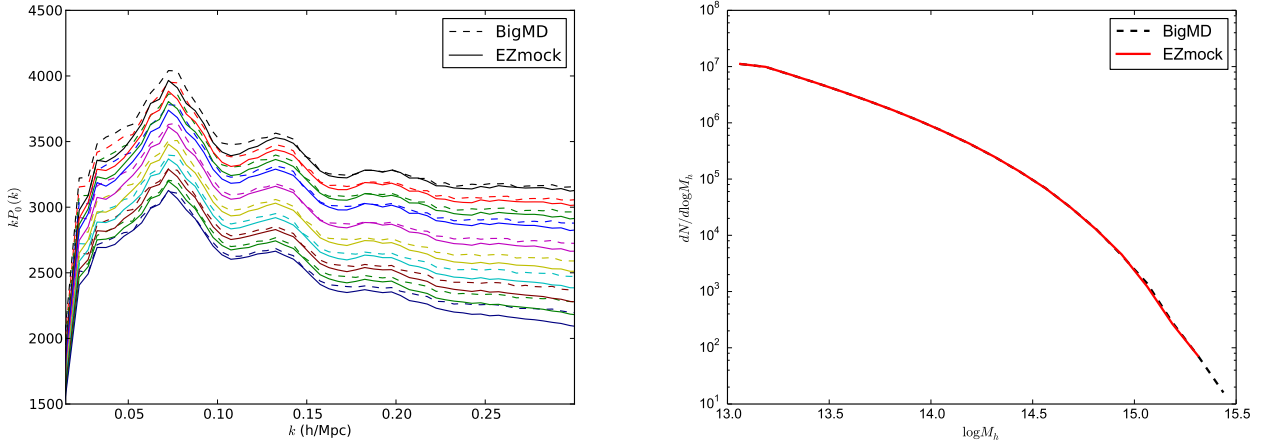
We find that, in general terms, the three-point statistics is well modelled by the effective Zel'dovich approximation on large scales. Fig. 5 shows the bispectrum of different configurations,  $k_2 = 2k_1 = 0.1$ ,  $k_2 = 2k_1 = 0.2$ ,  $k_2 = 2k_1 = 0.3$ , and  $k_2 = 2k_1 = 0.4 \, h \, \text{Mpc}^{-1}$ , in real and redshift-space. The bispectrum is fit mostly within 20% accuracy for the configurations,  $k_2 = 2k_1 = 0.1$  and  $k_2 = 2k_1 = 0.2 \, h \, \text{Mpc}^{-1}$ , in real-space, and for all the configurations tested, i.e.,  $k_2 = 2k_1 = 0.1$ ,  $k_2 = 2k_1 = 0.2$ ,  $k_2 = 2k_1 = 0.3$ , and  $k_2 = 2k_1 = 0.4 \, h \, \text{Mpc}^{-1}$ , in redshift-space. The accuracy in the three point statistics could be improved using a more accurate structure formation model, such as Augmented Lagrangian Perturbation theory (ALPT, Kitaura & Hess 2012). This would have, of course, an impact on the computational efficiency of our method. We will leave such an improvement for a later work.



**Figure 4.** Left panel: monopole of the two-point correlation function in real space. Right panel: monopole and quadrupole of the two-point correlation function in redshift-space.



**Figure 5.** Bispectrum in real and redshift-space. The configurations including  $k_2 = 2k_1 = 0.1$ ,  $k_2 = 2k_1 = 0.2$ ,  $k_2 = 2k_1 = 0.3$ , and  $k_2 = 2k_1 = 0.4$   $h \text{ Mpc}^{-1}$  are denoted in the panels.



**Figure 6.** Left panel: power spectra in real space with different number densities,  $\{2, 2.1, 2.2, 2.3, 2.5, 2.7, 2.9, 3.1, 3.3, 3.5\} \times 10^{-4} h^3 \text{ Mpc}^{-3}$ . Each subsample of a given number density is selected using a mass threshold, i.e., all the haloes with masses larger than a certain mass threshold with the right number density are selected. Right panel: mass functions for the EZmock and BigMD catalogues. They agree by construction.

#### 4.4 Mass vs. bias

Fig. 6 (left panel) shows the power spectrum in real-space for 10 different number densities, demonstrating that EZmock smoothly mimics the dependency between mass and bias using our methodology with 5 mass bins (see 4.4). One can still achieve a greater accuracy by increasing the number of mass bins. Fig. 6 (right panel) demonstrates that the mass function has been restored by construction.

### 5 CONCLUSION AND DISCUSSION

In this work, we have developed a new methodology to construct the effective Zel’dovich approximation mock catalogues (EZmocks), which provide an efficient way to generate massive mock catalogues with accurate one-, two-, and three-point clustering statistics. We have shown the good performance of EZmock by comparing the measurements in real and redshift-space with that from a large  $N$ -body simulation, BigMD. Constructing one mock, using a grid of  $960^3$  takes only a few minutes while using a machine with multicore processors (e.g., it takes 5 minutes while using 16 cores machine in our case). We have also presented an example to demonstrate the flexibility of assigning masses to EZmock, which gives a reasonable mass vs. bias relation.

We propose in this work to push the Zel’dovich approximation to its limits by extending it with effective models, which can compensate and alleviate the missing physical contributions. We demonstrate that such an approach is very promising to achieve high precisions, only reached with more complex approaches at higher computational costs. Our effective model including 1) PDF mapping with scattering, 2) density threshold and saturation, and 3) enhancing BAO and power spectrum at small scales, effectively absorbs the effects/corrections due to the nonlinear growth of the density field and halo bias (including deterministic and stochastic biasing), so that EZmock can precisely reproduce the clustering statistics. While we have used the simplest model with the minimum number of parameters (e.g., step function) to capture the major characteristics of the effects/corrections, we expect that the per-

formance of EZmock can be improved by using more precise models (e.g., using some smooth function instead of a step function).

Our method should be regarded as complementary to the efforts in producing catalogues based on full gravity solutions. High quality  $N$ -body simulations will remain indispensable to produce accurate reference catalogues, which can be massively reproduced with methods like the one presented in this paper. In addition, EZmock, at this stage, aims at providing a precise and practical tool for estimating the covariance matrices. It can be also conveniently used for future survey forecasts. EZmock provides a significant improvement with respect to log-normal mocks at a comparable computational cost.

In summary, provided the efficient production and accurate clustering statistics, EZmock can be one of the most practical methods to generate a large number of mock catalogues for analysing large-scale structure galaxy surveys.

### 6 ACKNOWLEDGEMENT

CC and FP were supported by the Spanish MICINN Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064 and AYA2010-21231-C02-01 grant, the Comunidad de Madrid under grant HEPHACOS S2009/ESP-1473, and Spanish MINECOs Centro de Excelencia Severo Ochoa Programme under grant SEV-2012-0249. CZ acknowledges support from Charling Tao and her grant from Tsinghua University, and 973 program No. 2013CB834906. CZ also thanks the support from MultiDark summer student program to visit the Instituto de Física Teórica, (UAM/CSIC), Spain. GY acknowledge 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).

The MultiDark Database 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.

We acknowledge PRACE for awarding us access to resource Curie supercomputer based in France.

REFERENCES

Abazajian K. N., et al., 2009, *Astrophys.J.Suppl.*, 182, 543, arXiv:0812.0649

Abell P. A., et al., 2009, 0912.0201

Angulo R. E., Foreman S., Schmittfull M., Senatore L., 2014, 1406.4143

Baldauf T., Mercolli L., Mirbabayi M., Pajer E., 2014, 1406.4135

Benitez N., et al., 2014, 1403.5237

Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, *Phys.Rept.*, 367, 1, astro-ph/0112551

Beutler F., Blake C., Colless M., Jones D. H., Staveley-Smith L., et al., 2011, *Mon.Not.Roy.Astron.Soc.*, 416, 3017, 1106.3366

Blake C., Davis T., Poole G., Parkinson D., Brough S., et al., 2011, *Mon.Not.Roy.Astron.Soc.*, 415, 2892, 1105.2862

Chuang C.-H., Wang Y., 2012, *Mon.Not.Roy.Astron.Soc.*, 426, 226, 1102.2251

Chuang C.-H., Wang Y., 2013a, *Mon.Not.Roy.Astron.Soc.*, 435, 255, 1209.0210

Chuang C.-H., Wang Y., 2013b, *Mon.Not.Roy.Astron.Soc.*, 431, 2634, 1205.5573

Chuang C.-H., Wang Y., Hemantha M. D. P., 2012, *Mon.Not.Roy.Astron.Soc.*, 423, 1474, 1008.4822

Cole S., et al., 2005, *Mon.Not.Roy.Astron.Soc.*, 362, 505, astro-ph/0501174

Coles P., Jones B., 1991, *Mon.Not.Roy.Astron.Soc.*, 248, 1

Colless M., et al., 2001, *Mon.Not.Roy.Astron.Soc.*, 328, 1039, astro-ph/0106498

Colless M., Peterson B. A., Jackson C., Peacock J. A., Cole S., et al., 2003, astro-ph/0306581

Dawson K. S., et al., 2013, *Astron.J.*, 145, 10, 1208.0022

de Jong R. S., Bellido-Tirado O., Chiappini C., Depagne E., Haynes R., et al., 2012, 1206.6885

Drinkwater M. J., Jurek R. J., Blake C., Woods D., Pimbblet K. A., et al., 2010, *Mon.Not.Roy.Astron.Soc.*, 401, 1429, 0911.4246

Eisenstein D. J., et al., 2011, *Astron.J.*, 142, 72, 1101.1529

Frieman J. D. E. S. C., 2013, *The Dark Energy Survey: Overview*. Vol. 221, American Astronomical Society, AAS Meeting

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

Hockney R. W., Eastwood J. W., 1981, *Computer Simulation Using Particles*. McGraw-Hill, New York

Kitaura F.-S., Gil-Marín H., Scoccola C., Chuang C.-H., Müller V., et al., 2014, 1407.1236

Kitaura F.-S., Hess S., 2012, 1212.3514

Kitaura F.-S., Yepes G., Prada F., 2013, 1307.3285

Klypin A., Holtzman J., 1997, astro-ph/9712217

Laureijs R., et al., 2011, 1110.3193

Leclercq F., Jasche J., Gil-Marín H., Wandelt B., 2013, *JCAP*, 1311, 048, 1305.4642

Levi M., et al., 2013, 1308.0847

Manera M., Samushia L., Tojeiro R., Howlett C., Ross A. J., et al., 2014, 1401.4171

Manera M., Scoccimarro R., Percival W. J., Samushia L., McBride C. K., et al., 2012, *Mon.Not.Roy.Astron.Soc.*, 428, 1036, 1203.6609

Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., et al., 2013, 1305.1505

Monaco P., Theuns T., Taffoni G., 2002, *Mon.Not.Roy.Astron.Soc.*, 331, 587, astro-ph/0109323

Nuza S., Sanchez A., Prada F., Klypin A., Schlegel D., et al., 2013, *Mon.Not.Roy.Astron.Soc.*, 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, *Mon.Not.Roy.Astron.Soc.*, 401, 2148, 0907.1660

Reid B. A., Percival W. J., Eisenstein D. J., Verde L., Spergel D. N., et al., 2010, *Mon.Not.Roy.Astron.Soc.*, 404, 60, 0907.1659

Riebe K., Partl A. M., Enke H., Forero-Romero J., Gottloeber S., et al., 2011, 1109.0003

Samushia L., Percival W. J., Raccanelli A., 2012, *Mon.Not.Roy.Astron.Soc.*, 420, 2102, 1102.1014

Schlegel D., et al., 2011, 1106.1706

Scoccimarro R., Sheth R. K., 2002, *Mon.Not.Roy.Astron.Soc.*, 329, 629, astro-ph/0106120

Springel V., 2005, *Mon.Not.Roy.Astron.Soc.*, 364, 1105, astro-ph/0505010

Tassev S., Zaldarriaga M., 2012a, *JCAP*, 1212, 011, 1203.5785

Tassev S., Zaldarriaga M., 2012b, *JCAP*, 1204, 013, 1109.4939

Tassev S., Zaldarriaga M., Eisenstein D., 2013, *JCAP*, 1306, 036, 1301.0322

White M., Tinker J. L., McBride C. K., 2013, 1309.5532

York D. G., et al., 2000, *Astron.J.*, 120, 1579, astro-ph/0006396

Zel’dovich Y., 1970, *Astron.Astrophys.*, 5, 84