HALOGEN: a tool for fast generation of mock halo catalogues

Santiago Avila,1,2∗ Steven G. Murray,3,4 Alexander Knebe,1 Chris Power,3,4 Aaron S. G. Robotham3 and Juan Garcia-Bellido1,2
1Departamento de Física Teórica, Módulo C-15, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Cantoblanco, Madrid, Spain
2Instituto de Física Teórica, UAM-CSIC, Universidad Autonoma de Madrid, E-28049 Cantoblanco, Madrid, Spain
3ICRAR, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia
4ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO)

Accepted 2015 March 27. Received 2015 March 24; in original form 2014 December 19

ABSTRACT

We present a simple method of generating approximate synthetic halo catalogues: HALOGEN. This method uses a combination of second-order Lagrangian Perturbation Theory (2LPT) in order to generate the large-scale matter distribution, analytical mass functions to generate halo masses, and a single-parameter stochastic model for halo bias to position haloes. HALOGEN represents a simplification of similar recently published methods. Our method is constrained to recover the two-point function at intermediate (10 h−1 Mpc < r < 50 h−1 Mpc) scales, which we show is successful to within 2 per cent. Larger scales (∼100 h−1 Mpc) are reproduced to within 15 per cent. We compare several other statistics (e.g. power spectrum, point distribution function, redshift space distortions) with results from N-body simulations to determine the validity of our method for different purposes. One of the benefits of HALOGEN is its flexibility, and we demonstrate this by showing how it can be adapted to varying cosmologies and simulation specifications. A driving motivation for the development of such approximate schemes is the need to compute covariance matrices and study the systematic errors for large galaxy surveys, which requires thousands of simulated realizations. We discuss the applicability of our method in this context, and conclude that it is well suited to mass production of appropriate halo catalogues. The code is publicly available at https://github.com/savila/halogen.

Key words: galaxies: distances and redshifts – galaxies: haloes – large-scale structure of Universe.

1 INTRODUCTION

We have entered an observational era where it is customary for redshift surveys to map millions of galaxies in the sky with the volumes of these surveys exceeding Gpc3 scales. Recent and upcoming galaxy survey projects include PAU (Castander et al. 2012), BOSS (Dawson et al. 2013), DES (Frieman & Dark Energy Survey Collaboration 2013), DESi (Levi et al. 2013), Euclid (Laureijs et al. 2011), etc. The interpretation of such surveys demands a new generation of theory tools in order to better understand and interpret the large amounts of data. One important component is the need for accurate simulations of the expected results, to which the observations should be compared. However, models of large-scale structure and the clustering of (dark matter) haloes forming in it are inherently non-linear, and require the production of simulations based on N-body calculations. Such simulations are extremely costly, and consequently very few realizations can be run for a given application. However, investigating the effects of systematic errors, cosmic variance, and their interplay require many hundreds of realizations of a single simulation (e.g. BOSS survey used 600 Manera, Scoccimarro & Percival 2013). These are necessary to compute covariance matrices which characterize the resultant uncertainty on the final parameters.

To mitigate this situation, many have now turned to approximate schemes in order to calculate the required realizations of the simulations. Early such work used the so-called log-normal realizations (Coles & Jones 1991), which placed particles randomly according to a log-normal distribution, given the true power spectrum. While this is indeed efficient, and reproduces two-point statistics faithfully, its lack of physical motivation for the particle placement results in poor higher order statistics, such as the three-point function or counts-in-cells moments. Improved methods developed in the past decade include PTHALOS (Scoccimarro & Sheth 2002; Manera et al. 2013), PINOCCHIO (Monaco, Theuns & Taffoni 2002; Monaco et al. 2013), PATCHY (Kitaura, Yepes & Prada 2014), COLA (Tassev, Zaldarriaga & Eisenstein 2013), QPM (White, Tinker & McBride 2014), EZMOCKS (Chuang et al. 2015), etc. For a comparison of these

∗E-mail: santiagoavilaperez@gmail.com
Table 1. Properties of the two reference N-body halo catalogues. From left to right: side length of the simulated cubic volume (in $h^{-1}$ Mpc), number of particles (for N-body and HALOGEN), redshift of the snapshot, cosmological parameters (density of baryons, total matter and dark energy, Hubble parameter, power spectrum normalization and spectral index), halo-finding technique, halo number density (in ($Mpc h^{-1}$)$^3$), method used to generate the ICs and redshift at which they were generated.

| Name             | $L_{box}$ | $N_{part}$ | $z$ | $\Omega_b$ | $\Omega_M$ | $\Omega_A$ | $h$ | $\sigma_8$ | $n_S$ | Finder | $n$ | IC   | $z_{IC}$ |
|------------------|----------|------------|-----|-------------|-------------|-------------|-----|------------|------|--------|-----|------|---------|
| GOLIAT           | 1000     | 512$^3$    | 0   | 0.044       | 0.27        | 0.69        | 0.7 | 0.8        | 0.96 | AHF    | $2.0 \times 10^{-4}$ | 2LPT | 32     |
| BigMultiDark     | 2500     | 3840$^3$, 1280$^3$ | 0.56 | 0.048       | 0.31        | 0.73        | 0.68 | 0.82       | 0.96 | FOF    | $3.5 \times 10^{-4}$ | ZA   | 100    |

methods (including HALOGEN presented here), we refer the reader to Chuang et al. (2014).

One may segregate these methods into two classes – predictive-type methods which are required to ‘find’ haloes in a given density field (e.g. PINOCCHIO, COLA and PTHALOS), and statistical-type methods which merely stochastically sample a density field to locate haloes (e.g. PATCHY, QPM and EZMocks). The former have the advantage of being predictive, and often not requiring an N-body reference simulation for calibration, while the latter have the advantage of computational speed and resources, as the number of particles used is reduced.

We present a new (statistical-type) approximate scheme, called HALOGEN whose prime objective is to generate halo catalogues with the correct two-point clustering and mass-dependent bias using a simple and rapid approach.

We note that statistical-type methods tend to follow a standard pattern of four steps.

(i) Produce a density field.
(ii) Sample halo masses.
(iii) Sample particles as haloes with some bias.
(iv) Assign halo velocities.

In this paper, we seek to abstract this pattern, providing a framework in which each step is highly modular. Whilst modular, HALOGEN implements default behaviour with very simple (and rapid) components using the popular second-order Lagrangian Perturbation Theory (2LPT) as the gravity solver, theoretical mass functions, a single-parameter bias prescription (as opposed to two or more parameters for other statistical-type methods) and a direct linear transformation of the velocities. As such, HALOGEN can be rapidly calibrated, and easily extended. In addition, we introduce physically motivated constraints for halo exclusion and mass conservation, which tie the individual steps together.

In this paper, we will compare the results from HALOGEN to a pair of reference N-body simulations to be presented in Section 2. We introduce the general ideas of the method in Section 3, leaving a more detailed explanation of the spatial placement of haloes – which we consider the essence of HALOGEN – for Section 4. Section 5 demonstrates the effects of each parameter of HALOGEN and how to optimize them. We conclude with some applications and results in Section 6.

2 THE REFERENCE SIMULATIONS

To tune HALOGEN to a specific cosmology, we require an N-body simulation. In order to show the adaptability of HALOGEN to varying setups, we have not limited ourselves to a single simulation but used two with differing box size, mass resolution, and cosmology. Further, the reference halo catalogues have been obtained by applying different halo-finding techniques, and have different number density. We summarize the characteristics of both reference catalogues in Table 1 and describe them below.

**Goliat Simulation.** This simulation was run with the GADGET2 code (Springel 2005) from initial conditions (ICs) generated by 2LPTIC at $z = 32$. It uses $N = 512^3$ dark matter particles in a box with side length $L_{box} = 1000 h^{-1}$ Mpc. The cosmological parameters used in this simulation are $\Omega_M = 0.27$, $\Omega_A = 0.73$, $\Omega_b = 0.044$, $h = 0.7$, $\sigma_8 = 0.8$, $n_s = 0.96$ yielding a mass resolution of $m_p = 5.58 \times 10^{11} h^{-1} M_{\odot}$. The halo catalogue was obtained from a $z = 0$ snapshot and has been generated with the halo finder AHF (Knollmann & Knebe 2009), a spherical-overdensity (SO) algorithm. Though AHF identifies subhaloes, they have been discarded for the present analysis as these scales are too small for 2LPT to resolve. There is a possibility of phenomenologically adding substructure in a later step using a halo occupation distribution (HOD) prescription (Skibba & Sheth 2009), but we leave that to a future study. In this catalogue, we use a halo reference density of $n = 2.0 \times 10^{-4}$ ($Mpc h^{-1}$)$^3$.

HALOGEN requires an input density field obtained from 2LPT (see Section 3.1). For this purpose, we run a 2LPTIC snapshot at $z = 0$ with the same IC phases as those used in GOLIAT.

**Big MultiDark Simulation.** BigMultiDark described in Klypin et al. (2014), employs the cosmology from the Planck Collaboration XVI (2014), which for some parameters represents a significant change with respect to the GOLIAT simulation: $\Omega_M = 0.31$, $\Omega_A = 0.69$, $\Omega_b = 0.048$, $h = 0.68$, $\sigma_8 = 0.82$, $n_s = 0.96$. The halo catalogue is extracted with a Friends-of-Friends (FOF; Davis et al. 1985) algorithm (which intrinsically neglects substructure) at $z = 0.56$, and we choose a reference halo number density of $n = 3.5 \times 10^{-4}$ ($Mpc h^{-1}$)$^3$.

Compared to GOLIAT, is both larger ($L_{box} = 2500 h^{-1}$ Mpc) and more resolved ($N = 3840^3$ particles of mass $m_p = 2.3 \times 10^{10} h^{-1} M_{\odot}$). It was run with L-GADGET2 from ICs based on the Zel’dovich Approximation (ZA) at $z = 100$. Given the large scales that it explores while resolving large numbers of haloes, it is well suited to probing the Baryon Acoustic Oscillation (BAO) peak.

For the input of HALOGEN, we run 2LPTIC to $z = 0.56$ with the same IC phases as BigMultiDark. The cosmology and $L_{box}$ used are the same, but with a lower resolution of $N = 1280^3$.

3 METHOD OUTLINE

In this section, we briefly outline our method, leaving a more detailed presentation of the actual modus operandi of HALOGEN for Section 4. The general algorithm consists of four (major) steps:

(i) generate a dark matter density field,
(ii) draw halo masses by sampling a halo mass function (HMF),
(iii) populate the volume with haloes in the box, and
(iv) assign velocities to the haloes.

1 http://cosmo.nyu.edu/roman/2LPT
2 http://www.cosmosim.org

1857
We aim to decouple each of these steps from the others as far as possible so that different algorithms may be used at each point. The first two steps are relatively trivial, as they use pre-developed prescriptions from the literature, and we discuss these, and basic outlines of the last two steps, in this section.

3.1 Density field

The basic scaffolding of HALOGEN is an appropriate dark matter density field realized at the desired redshift, sampled by \( N \) particles. For simplicity, we choose to use second-order perturbation theory (2LPT) (Moutarde et al. 1991; Bouchet et al. 1995) to produce this field, which can be obtained with the public code 2LPTIC.

We show in Fig. 1, the density distribution of an \( N \)-body simulation (top panel) and a 2LPT representation (bottom panel) at \( z = 0.5 \). Notably, the 2LPT distribution appears to be blurred in comparison to the \( N \)-body simulation. This is due to the fact that 2LPTIC – as the name suggests – was originally designed only to generate ICs (Scoccimarro 1998), since even second-order perturbation theory breaks down at low redshift when overdensities become highly non-linear. The small-scale difference in Fig. 1 can be explained by shell crossing, an effect in which particles following their 2LPT trajectories cross paths and continue rather than gravitationally attracting each other in a fully non-linear manner (Sahni & Shandarin 1996; Neyrinck 2013). In order to compensate for shell-crossing, Manera et al. (2013) advocates the use of a smoothing kernel over the input power spectrum. We tested the effect of this smoothing in HALOGEN but did not find any improvement in the final catalogue.

Nevertheless, 2LPT provides a suitable approximation of the large-scale distribution of matter, where perturbations have not yet entered into the highly non-linear regime and this is sufficient for HALOGEN. Note that HALOGEN is in principle agnostic about the method in which this density field snapshot is produced. Other methods, for instance the ‘Quick-PM’ (cf. the QPM method described by White et al. 2014), COLA (Tassev et al. 2013) or 3LPT could equally be employed by the user. A different choice of density field will yield somewhat different results, especially at smaller scales. As long as the chosen method reconstructs large scales correctly, the remaining steps of HALOGEN should be unmodified.

Despite this, we have by default incorporated 2LPTIC as part of the HALOGEN code (which bypasses the costly I/O of writing the snapshot to disc), but also allow the user to provide an arbitrary snapshot with a distribution of \( N \) particles in a cosmological volume. Our choice for 2LPT was mainly driven by its low computational cost and success in the distribution of matter at large scales. We use this approach for all results in this paper.

3.2 The mass function

The HMF \( n(>M) \) measures the number density of haloes above a given mass scale. It is required to generate mass-conditioned clustering, which in turn is a pre-requisite for extension to HOD-based galaxy mock generation.

We produce a sampled mass function by the standard inverse-Cumulative Distribution Function (CDF) method, utilizing an arbitrary input HMF.

The most accurate HMF for a given cosmology, over a range of suitable scales, may be obtained from an \( N \)-body simulation via a halo-finding algorithm – although there are notable variations depending on the technique (Knebe et al. 2011). Since we require a full \( N \)-body simulation for the tuning of HALOGEN, it would be perfectly acceptable to use this simulation to generate the HMF. However, in the hope of future improvements, we wish to avoid using the full simulation as far as possible. Fortunately, there is a wealth of literature concerning accurate predictions of the HMF for widely varying cosmologies and redshifts using extended Press–Schechter theory (Press & Schechter 1974; Bond et al. 1991).

The mass function may be calculated by any means, so long as a discretized function of \( n(>M) \) is provided. For simplicity, we decided to use the online HMF calculator HMFcalc\(^3\) (Murray, Power & Robotham 2013) for obtaining the halo mass distribution in this paper.

In the remainder of the paper, we use the fit of Watson et al. (2013) for BigMultiDark and that of Tinker et al. (2008) for GOLIAT which both constitute reliable fits.

3.3 Spatial placement of haloes

The crucial step in the generation of approximate halo catalogues is the commissioning of halo positions. In keeping with the philosophy of modularity, the halo-placement step is decoupled from the rest.

---

\(^3\) http://hmf.icrar.org
Any routine which takes a vector of halo masses and an array of dark matter particle positions and returns a subset of those positions as the halo locations is acceptable. However, we consider this step to be at the heart of the HALOGEN method, as it is responsible of generating the correct mass-dependent clustering.

To achieve an efficient placement that reconstitutes the target two-point statistics, we recognize the validity of the clustering on large scales from the broad-brush 2LPT field. We place haloes on 2LPT field particles, essentially using the estimated density field as scaffolding on which to build an approximate halo field. We will follow a series of steps in the construction of the method of spatial placement to be presented in Section 4 below.

3.4 Assignment of velocities

The most obvious way to assign velocities to each halo would be to use the velocity of the particle on which it is centred. However, haloes are viralized systems whose velocities tend to be lower than that of their constituent particles. This is potentially mitigated by using the average velocity of all particles within a defined radius of the artificially placed halo. However, this is not robust as there are often very few particles inside the halo radius. Additionally, the 2LPT particle velocities will differ from their N-body counterparts due to shell-crossing, especially on the small scales associated with haloes.

Thus, we prefer to take a phenomenological approach, and assume that a simple mapping via a factor $f_{\text{cel}}$ can be applied to the collection of halo velocities to recover the results of the N-body distribution

\[ v_{\text{halo}} = f_{\text{cel}} \cdot v_{\text{part}}. \]

This factor could a priori depend on the velocity (i.e. a non-linear mapping) and the mass of the halo $f_{\text{cel}}(v_{\text{part}}, M_{\text{halo}})$. However, we will show in Section 5.2 that a linear mapping is sufficient and present a way to compute $f_{\text{cel}}(M_{\text{halo}})$.

4 HALOGEN

Though HALOGEN is a four-stage process, the most crucial aspect is the assignment of halo positions, which this section describes in some detail. The general concept is to specify a sample of particles from an underlying density field as haloes.

The motivating philosophy of HALOGEN is to start from the simplest idea and improve if necessary. In this vein, we present here successive stages of evolution of the HALOGEN method, which we hope will show satisfactorily that the method as it stands is optimal. Fig. 2 will serve as the showcase for the various stages of HALOGEN. In it, we present the two-point correlation function (2PCF) for each stage of development to verify that the method approaches the GOLIAT reference catalogue as new characteristics are added.

Note that the 2PCF is computed with the publicly available parallel code CUTE\(^4\) (Alonso 2012). In the fitting routine that is included in the HALOGEN package and described in Section 5.1 we also use the same code.

4.1 Random particles

We start with the simplest approach, using random particles from the 2LPT snapshot as the sites for haloes. We expect to recover the large-scale shape of the 2PCF in this way, as this is encoded in the 2LPT density field which we trace.

However, it is clear from Fig. 2 that this method (‘random no-exc’) consistently underestimates the 2PCF over all scales except $r < 1 h^{-1}$ Mpc, where it should sharply drop to −1, but rather remains positive.

The consistent underestimate is a realization of an inaccurate linear bias, $b$, defined as the scaling factor between the two-point function of the haloes and the underlying matter density field,

\[ \xi_{\text{halo}}(r) = b^2 \xi_{\text{dm}}(r). \]

We begin to address this in Section 4.3.

The small-scale clustering can be explained by the fact that particles can be arbitrarily close, whereas distinct haloes – recall that subhaloes have been removed – have a well-defined minimum separation (otherwise they merge). The turnover in the simulation based 2PCF occurs around the mean halo radius scale.

4.2 Random particles (with exclusion)

The simplest improvement to the random case is to eliminate the artificial small-scale correlations. Though the primary application of HALOGEN will be for large scales, a simple improvement at small scales is useful.

As we have noted, the artificial clustering at small scales arises from the fact that particles can be arbitrarily close, whereas simulated haloes have a minimum separation. The radius of a halo is a rather subjective quantity, and its definition is modified in various applications and halo finders. However, we may parametrize this by

\[ R_\Delta = \left( \frac{3M_{\text{halo}}}{4\pi\Delta_\text{v} \rho_{\text{crit}}} \right)^{1/3}, \]

where $\Delta_\text{h}$ is the overdensity of the halo with respect to the critical density of the Universe. For the work presented here we used $\Delta_\text{h} = 200$.

Using this scale, we introduce exclusion, a modifiable option which controls the degree to which haloes can overlap, which we set to mimic the halo finder’s specification. For example, in this work we use both AHF and FOF (see Knebe et al. 2013, for a comparison

\(^{4}\) http://members.ift.uam-csic.es/dmonge/CUTE.html

Figure 2. 2PCF of the GOLIAT haloes in comparison to HALOGEN for the various evolutionary stages presented in Sections 4.1 through 4.5. The dashed vertical line indicates the cell size of $l_{\text{cel}} = 5 h^{-1}$ Mpc applied for the approaches 4.3 through 4.5.
and an introduction to halo-finding techniques). For the latter, we
do not allow any overlap whereas for the former HALOGEN’s halo
centres are not allowed to lie inside another halo’s radius.

The effect of exclusion is presented in Fig. 2 (‘random exc’). As
expected, scales of $r < 1 \, h^{-1} \, \text{Mpc}$ show a turnover while larger
scales are unaffected. We note that the turnover is at smaller scales
for HALOGEN than for AHF. This is to be expected, as it is unlikely to
find two AHF haloes separated by a distance slightly exceeding $R_\alpha$,
due to reasons akin to the FOF overlapping problem. In such cases,
there is an increased likelihood of the two haloes being subsumed
into one, or one becoming a subhalo of the other. It is conceivable
that one could empirically model these effects by tuning the value of
$\Delta_\alpha$ by some factor which captures this suppressed probability.
However, as we are more interested in large scales and these con-
siderations touch upon the subtleties of halo definition, we consider
these exclusion criteria sufficient for present purposes. We will use
this form of exclusion (in an appropriate form) for all following
work.

4.3 Ranked approach

We return now to the problem of underestimation of the correlations,
which we noted due to an incorrect realization of the linear halo
bias. In effect, a random choice of particle position corresponds to
sampling the matter power spectrum uniformly, and therefore
$\beta = 1$. However, halo bias is generally greater than unity (especially
for higher mass halo samples) (Tinker et al. 2005).

Increasing the bias corresponds to sampling higher density re-
gions. The simplest way to achieve this is to rank-order the density
of regions in the particle distribution, and assign haloes to these
regions based on their mass.

To calculate densities from the particle distribution, we simply
create a uniform grid with cell size $l_{\text{cell}}$, and obtain the density in
each cell using a Nearest-Grid-Point (NGP) assignment scheme
(Hockney & Eastwood 1988). We consider specification of the op-
timal $l_{\text{cell}}$ in Section 5.3. The cells are ordered by density, and the
haloes by mass, and each halo is assigned to its corresponding cell
(a random particle is chosen within the cell).

Using $l_{\text{cell}} = 5 \, h^{-1} \, \text{Mpc}$ in this case, we obtain the results shown
in Fig. 2 labelled ‘ranked exc’. The resulting 2PCF is now overesti-
mated. This is not surprising, since even if we expect haloes to form
in dense environments, the bias is not completely deterministic: in
reality the nth most massive halo does not need to reside in the nth
densest place.

The effect of introducing a scalelength, $l_{\text{cell}}$, is also clearly seen
in this result. There is a turnover in the 2PCF below $l_{\text{cell}}$, which
corresponds to a significant reduction of bias on these scales since
a random particle is chosen within the cell.

4.4 $\alpha(M)$ approach

We find that selecting completely random particles yields too low
a bias, whereas the ranked approach is highly biased. We require
an intermediate solution, which has higher probability of selecting
dense areas than the random approach, and lower probability than
the ranked approach.

The probability that a cell is chosen is a function of its density,

$$P_{\text{cell}} \propto G(\rho_{\text{cell}}).$$ (4)

In the completely random case, we have $G(\rho_{\text{cell}}) = \rho_{\text{cell}}$. In principle,
we can tailor $G(\rho_{\text{cell}})$ so that the probability of selecting a cell

reproduces the appropriate bias. We choose to constrain $G(\rho_{\text{cell}})$ to
have a power-law form, i.e.

$$G(\rho_{\text{cell}}) = \rho_{\text{cell}}^\alpha.$$ (5)

When $\alpha = 1$, we recover the random approach, and as $\alpha \to \infty$ we
obtain the ranked approach.

In Fig. 2, we show results for $\alpha = 1.5, 2$, demonstrating the
effectiveness of our model for tuning the normalization (i.e. bias)
of the 2PCF. The $\alpha = 1.5$ curve closely matches the 2PCF of the
AHF catalogue, at least at scales larger than the applied cell size
$l_{\text{cell}} = 5 \, h^{-1} \, \text{Mpc}$.

The exact value of $\alpha$ for a particular application may be deter-
bined by a least-squares fit, which we describe in more detail in
Section 5.1 (note that here the choice of $\alpha$ was not formally fit).

In corollary with this prescription, we also introduce a means to
roughly ensure mass conservation in cells; once a halo is placed,
if the total halo mass in the cell exceeds the original mass, the cell
is eliminated from future selections. However, we do not update
the value of the probability after every halo placement because it
is computationally very expensive ($O(N_h^3)$) and we have checked
that doing so has a negligible effect on output statistics.

We note that a similar method was employed in QPM (White
et al. 2014). In fact, the physically meaningful quantity is $f_{\text{halo}}(\rho)$ –
the distribution of halo density (i.e. the fraction of haloes in cells
with density $\rho$). This can be written as

$$f_{\text{halo}}(\rho) = P(\text{cell}|\rho) f_{\text{cell}}(\rho),$$ (6)

where $P(\text{cell}|\rho)$ specifies the relative probability of choosing a cell
given its density (in our case, $\rho^\alpha$), and $f_{\text{cell}}(\rho)$ is the intrinsic distri-
bution of cell densities given the cell size and cosmology (heavily
related to the cosmological parameter $\sigma_8$). QPM specifies the target
distribution $f_{\text{halo}}(\rho)$ directly, as a Gaussian. In HALOGEN, we instead
specify $P(\text{cell}|\rho)$, which is more closely tied to our algorithm.
In principle, one can convert from QPM-like methods to HALOGEN with
equation (6).

4.5 $\alpha(M)$ approach

The approach as it stands reproduces the 2PCF accurately down to
the scale of $l_{\text{cell}}$. If the 2PCF of a sample of given number density is
all that is required for a specific application, then this will do well.
However, if we were to select a subsample of the most massive
haloes of our catalogues and recompute the 2PCF, the bias would be
incorrect, since more massive haloes are more biased (Tinker et al.
2005). For a truly representative catalogue, in which the haloes
are conditionally placed based on their mass, the bias model is
required to be mass-dependent. Failing this, there is no physical
meaning attached to the assignment of masses in the second step
(Section 3.2).

Mass-dependent halo bias is also crucial for implementing HOD
models on the catalogue, for use in galaxy survey statistics, as the
game of galaxies associated with a halo depends on its mass.

We incorporate this mass-dependence into the $\alpha$ parameter, so
that we finally have

$$G(\rho_{\text{cell}}, M) = \rho_{\text{cell}}^{\alpha(M)},$$ (7)

with $\alpha(M)$ an increasing function.

In practice, we use discrete mass bins, and for each bin $i$, with
masses $M_{\text{in}}^{i-1} > M > M_{\text{in}}^i$, we use a different $\alpha_i$. We describe how
we obtain the best fit to this mass-dependent $\alpha$ using the fiducial
halo catalogue from the simulation in Section 5.1.
Table 2. Properties of the selected mass bins for the GOLIAT simulation: mass threshold \( M_{\text{th}} \), equivalent number density \( n(M > M_{\text{th}}) \), and best-fitting \( \alpha_i \) in \( M_{\text{th}}^{-1} < M < 2 M_{\text{th}} \) for the HALOGEN \( \alpha(M) \) approach.

| Bin | \( M_{\text{th}} \) (\( h^{-1} M_\odot \)) | \( n_i \) (\( h^{-1} \text{Mpc}^{-3} \)) | \( \alpha_i \) |
|-----|---------------------------------|----------------------------|----------|
| 0   | 1.64 \times 10^{14}           | 0.05 \times 10^{-4}       | 3.54     |
| 1   | 4.80 \times 10^{13}           | 0.40 \times 10^{-4}       | 2.26     |
| 2   | 2.65 \times 10^{13}           | 0.90 \times 10^{-4}       | 1.77     |
| 3   | 1.86 \times 10^{13}           | 1.40 \times 10^{-4}       | 1.48     |
| 4   | 1.38 \times 10^{13}           | 2.00 \times 10^{-4}       | 1.41     |

Using just five mass bins, we illustrate this approach in Fig. 2, labelled ‘\( \alpha(M) \) exc’ (magenta line) using the best-fitting values for \( \alpha(M) \). We list in Table 2 the mass thresholds, applied \( \alpha \)-values, and corresponding number densities of all haloes with \( M_{\text{halo}} > M_{\text{th}} \). Note that though the probability is not recomputed after placing a halo, it is recomputed with updated \( \rho \) and \( \alpha \) when changing mass bins.

Though the \( \alpha(M) \) approach does not improve the 2PCF with respect to the \( \alpha \) approach in Fig. 2, it has the clear advantage of reproducing a mass-dependent clustering, which as we noted is essential for further HOD analyses, and useful for being able to use any mass range in the same realization.

4.6 Summary

In conclusion, HALOGEN constitutes a method for generating a halo catalogue which exhibits correct two-point clustering statistics, while not only positioning the haloes correctly, but also imbuing them with physically meaningful masses. The method can be summarized as follows.

The particles generated by 2LPT (Section 3.1) are covered by a grid of cell size \( L_{\text{cell}} \), the halo masses \( M_i \) generated from the HMF (Section 3.2) are ordered by mass, and starting from the most massive halo they are placed by

(i) selecting a cell with probability \( P_{\text{cell}} \propto \rho^{\alpha_i} \),

(ii) randomly selecting a particle within the cell and using its coordinates as the halo position,

(iii) ensuring that the halo does not overlap (following an exclusion criterion) with any previously placed halo in any cell, and

(iv) re-choosing a different random particle in that case.

5 PARAMETER STUDY

We have mentioned several parameters of the HALOGEN method, and these are of particular importance in producing accurate realizations. In this section, we will discuss each parameter, its effects and how to optimize for it if possible.

There are three parameters in HALOGEN (with other options and parameters being expressly determined by the required output, such as the size of the simulation box \( L \)): the two physical parameters of the model, \( \alpha \) – controlling the linear bias – and \( f_{\text{cell}} \) – controlling the velocity bias – and the one parameter of the algorithm, \( L_{\text{cell}} \).

These are summarized in Table 3, and detailed in the following subsections.

In the previous section, we used GOLIAT as a reference. We now turn to BigMultiDark and its FOF catalogue; this simulation has a larger volume, allowing us to probe BAO scales. The increased volume also reduces cosmic variance on intermediate scales. HALOGEN primarily aims at reproducing clustering statistics for even larger volumes, hence it is beneficial to assess the performance of HALOGEN and its parameters in this regime. Furthermore, this demonstrates independence from the underlying simulation and halo-finding technique.

5.1 Fitting \( \alpha(M) \)

The value of \( \alpha(M) \) is crucial to the performance of HALOGEN, as it constitutes the only physical parameter controlling the bias. The HALOGEN package contains a stand-alone routine which determines a best fit for \( \alpha(M) \), which can then be passed to BigMultiDark to generate any number of realizations. We describe this routine here, and illustrate it with application to BigMultiDark. The fitting of \( \alpha(M) \) is based on the standard \( \chi^2 \)-minimization technique. However, a few details are worth mentioning.

Mass-dependence. We perform the fit in sharp-edged mass bins to determine a mass-dependent \( \alpha(M) \), i.e. for each bin \( i \) we fit a \( \alpha_i \) for the mass range \( M_i^{-1} < M < 2 M_i \). There are two conceivable ways of doing this – differentially or cumulatively. We have experimented with both and find that the cumulative procedure has better performance. That is, we fit the first mass bin, and then the first and second together (keeping the best value of \( \alpha_i \) for the first bin), and so on. This has the advantage of being able to properly correct for deviations in previous bins, which is particularly important since the first bins to be fit are the high mass bins, for which fewer haloes exist. Misestimation of \( \alpha \) here is more likely, but is compensated for when fitting to lower mass bins by including the high-mass estimates in the fit.

HALOGEN variance. The halo placement in HALOGEN is probabilistic, even given a constant underlying density field. Using different random seeds can slightly affect the final placement, and thus the clustering statistics (the extent of this is dependent on the volume, \( n \) and \( \alpha \)). We term this HALOGEN variance”, and note that it is not to be confused with cosmic variance. Cosmic variance is introduced by modifying the random seed of 2LPTC, which in effect results in a different realization of the universe. During the fit, each mass bin is realized several times (ten in the case of BigMultiDark) with HALOGEN to average out the effects of HALOGEN variance, and also...
provide an error $\sigma_H$ (computed as the standard deviation) to use in the definition of $\chi^2$.

**$\chi^2$ minimization.** The fit is performed by minimizing $\chi^2$,

$$\chi^2(\alpha) = \sum_j \left( \frac{\xi_H(r_j|\alpha) - \xi_{NB}(r_j)}{\sigma_H(r_j)} \right)^2,$$

where $\xi_H$ and $\xi_{NB}$ are the 2PCFs of HALOGEN and the reference catalogue, respectively. We note that minimizing this statistic is susceptible to systematic errors in HALOGEN in bins where the stochastic error ($\sigma_H$) is much smaller than the systematic error ($\Delta \xi$). This is especially likely when the region of the fit approaches $k_{\text{cell}}$. To test whether the region is stable, we may choose a distance estimator to be minimized that treat all scales with the same weight, e.g. $\Delta = (\xi_H - \xi_{NB})^2/\xi_{\text{NB}}^2$. We have tried with both quantities in our fitted range, and the results are left unchanged, indicating that the range of the fit is stable.

We use a grid of $\alpha$ to cover the expected result for each mass bin. We use a cubic spline interpolation over $\chi^2(\alpha)$ to locate a precise minimum for the best-fitting $\alpha$.

**Fitting Range.** We restrict the range of the fit to scales in which the shape of $\xi_H(r)/\xi_{NB}(r)$ is flat. This corresponds to mid-range scales of $15 \ Mpc < r < 47 \ Mpc$, which avoids small-scale effects of HALOGEN, and large-scale cosmic variance.

**Number of mass bins.** The number of bins to use in this procedure will depend on the needs of the user, and the size and resolution of the reference simulation. It determines the reliability of the mass-dependent clustering. For B\textsc{BigMultiDark}, we distribute the haloes into eight roughly equinumbered bins with the mass thresholds $M_{\text{th}}$ as shown in Table 4. In that table, we also show the best-fitting $\alpha_i$, and the equivalent number density $n_i$ for each mass threshold.

The 2PCFs for our eight values of $n_i$ are shown in Fig. 3, where we compare the results from HALOGEN against the B\textsc{BigMultiDark} reference catalogue. The range used during the fitting procedure and for the $\chi^2$-minimization is indicated by the vertical lines.

We note that the choice of $\alpha$ finely controls the bias. This is demonstrated in Fig. 4, in which we show the resultant $\xi(r)$ for the entire grid of $\alpha_i$ for this fit (left-hand side). There is a $\sim 10$ per cent deviation in $\xi_H(r)$ over the grid range (1 per cent between consecutive lines). On the right-hand side of the figure, we show the $\chi^2$ of each of those curves and the cubic spline fit interpolation.
used to find the minimum, which corresponds to the $\alpha_I$ best-fitting value shown in Table 4.

5.2 Velocity factor $f_{vel}$

In Section 3.4, we outlined a method of converting the velocity of 2LPTIC particles (designated as halo sites), $v_{x,p}$, to the velocity of a HALOGEN halo, $v_x$. We stated that the transformation was linear in $v_{x,p}$, and thus we can write

$$v_x = f_{vel}(M) \cdot v_{x,p},$$

where we have retained a mass-dependence in the conversion factor. This section will explore the means to calculate this factor.

We begin by justifying our choice of a linear function. Fig. 5 shows the one-component velocity distribution of BIGMULTIDARK and the particles selected by HALOGEN. Both curves are well-described by a Gaussian with $v_0 = 0$, where the standard deviation of the $N$-body haloes is reduced compared to that of $v_{x,p}$, i.e. $\sigma_x > \sigma_{NB}$. This confirms our claim in Section 3.4 that the particle velocities are larger than the halo velocities, and also shows that a simple linear transformation suffices to map the distribution of $v_{x,p} \rightarrow v_x$.

This simple characterization leads to a transformation of $f_{vel} = \sigma_{NB}/\sigma_p$, which is verified by the blue dotted line where this remapping has been applied.

We expect that the velocity bias (Colín, Klypin & Kravtsov 2000) will be dependent on mass-scale in general. We can easily incorporate this into our fitting routine by calculating

$$f_{vel} = \frac{\sigma_{NB}}{\sigma_p},$$

for each interval of mass $M = (M_{i-1}^h : M_{i}^h)$ while performing the fit for $\alpha$. These results are also listed in Table 4. There is a noticeable decrease in $f_{vel}$ towards higher mass haloes. We will see in Section 6.4 how this affects the modelling of redshift space distortions.

We finally note that there may be other more complex models of velocity bias accounting for the physics of low scales and adjusting other statistics beyond the overall velocity distribution. However, the model presented here is very simple and capable of reproducing the halo velocity distribution with a great accuracy.

5.3 Cell size: $l_{cell}$

We have previously mentioned the cell size $l_{cell}$ which is introduced to HALOGEN to provide a simple local density via the NGP scheme (Hockney & Eastwood 1988). We have also noted that it defines a lower limit of reliability of the resultant 2PCF. In this section, we explore this parameter further, describing its effects and how to optimize for it.

In Fig. 6, we show the 2PCF of the BIGMULTIDARK catalogue against HALOGEN results for several values of $l_{cell}$. We note two effects, $l_{cell}$

(i) determines the minimum scale at which the 2PCF is reliable and
(ii) controls the broadening of the BAO peak.

The first effect is clearly noticeable in the left-hand panel where the HALOGEN 2PCF detaches from the BIGMULTIDARK curve at $r \approx l_{cell}$. This is expected, since particles are chosen at random inside the cell, reducing the bias at these scales.

The second effect is more noticeable in the right-hand panel. As $l_{cell}$ is decreased, the broadening and dampening (best seen in the lower panel as the difference between the artificial peak at $r = 80\, h^{-1}$ Mpc and trough at $r = 100\, h^{-1}$ Mpc) is decreased. The reason for this is that we introduce an uncertainty (on a scale $l_{cell}$) in the position of the haloes which propagates to an uncertainty in the determination of $r_{BAO}$. In effect, the density field has been filtered by a quasi-top-hat function (Hockney & Eastwood 1988), which has the known effect of peak broadening.

Clearly, $l_{cell}$ should be set as small as possible to mitigate these effects. However, a limit is imposed by the mean-interparticle-separation, $d_p$, of the input density field. We cannot hope to reliably probe scales smaller than $d_p$, and even just above this scale we run into the problem of having poor statistics within cells. We recommend using a value of $l_{cell} \gtrsim 1.5d_p$ (ensuring $>3$ particles per cell on average), and in this work we take $l_{cell} = 4\, h^{-1}$ Mpc $\approx 2d_p$ as the reference.

We comment here that the choice of $l_{cell}$ affects the optimal $\alpha(M)$ relation. This is unfortunate, because it would be useful to be able to perform the fit for $\alpha$ using a lower resolution (since this is the bottleneck). The mechanism by which this effect occurs is known, and we hope to be able to correct for it in the future.

Let us illustrate the mechanism with an example; suppose we take a cell with cell size $l_{cell}^I$ and density $\rho_{cell}^I$ from a volume $(N_{cell}^I)^3$. For the same distribution, we could also use $l_{cell}^II = l_{cell}^I/2$, which forms eight subcells $i$ with densities $\rho_{cell,i}^II$. For the same $\alpha$, the probability of choosing the cell in case I is

$$P_{cell}^I = \frac{\left(\rho_{cell}^I\right)^\alpha}{\sum_i \left(\rho_i\right)^\alpha},$$

whereas in case II we have

$$P_{cell}^{II} = \frac{\sum_i \left(\rho_{cell,i}^II\right)^\alpha}{\sum_i \left(\rho_i\right)^\alpha},$$

and clearly these are not in general equivalent if $\alpha \neq 1$. We expect the difference in the distributions to be dependent on $\alpha$, the two cell sizes and their ratio and the cosmology, via the mass variance $\sigma(r)$. In future studies, we hope to be able to quantify this relationship to enable faster fitting.

Fig. 7 shows the effect of changing $l_{cell}$ on the best-fitting $\alpha(M)$ and we notice two characteristics. First, $\alpha(M)$ is an increasing function for all $l_{cell}$, as expected since $b(M)$ is increasing. Secondly, low
masses are less sensitive to $l_{\text{cell}}$, which we expect mathematically from equations (11) and (12) with an increasing $\alpha(M)$ (the greater $\alpha$ is, the greater the differences expected).

In Fig. 6, we have re-fit the $\alpha(M)$ relation for each value of $l_{\text{cell}}$, ensuring proper comparison between curves. Furthermore, we run five realizations of each and display the average, to reduce the effects of HALOGEN variance.

6 RESULTS AND APPLICATIONS

While previous sections were dedicated to the design and optimization of HALOGEN, we have now defined the final method and fixed the optimal parameters. In this section, we discuss the performance of HALOGEN in more detail, both in the clustering statistics so far analyzed, and in other statistics that HALOGEN is not constrained to match. We begin by demonstrating the power of HALOGEN for mass production of halo catalogues for use in deriving covariance matrices to measure cosmic variance, which we envision as the primary application of the HALOGEN machinery.

6.1 Mass production of halo catalogues

The driving motivation of developing fast methods for synthetic halo catalogues is to accurately produce robust covariance matrices for large galaxy survey statistics. Though HALOGEN requires a full $N$-body simulation to calibrate its two parameter sets, once these parameters have been established, we are free to run as many realizations (with different phases for the ICs of the halo catalogue (using the same cosmological parameters, volume, mass resolution etc.) as we like. This process is expected to purely simulate the effects of cosmic variance, and thus is extremely valuable for deriving the covariance matrices.

In order to verify that the variance seen in the resulting data traces the expected cosmic variance, we complemented the generation of the HALOGEN catalogues with several corresponding $N$-body simulations. Due to the computational time constraints, we were only able to run five simulations, which were based on GOLIAT, and in which only the seed for the random IC phases was changed. The ICs for these runs were generated with 2PTIC at redshift $z = 32$ (for the $N$-body) and $z = 0$ (for HALOGEN), using the same seed for each pair. The $N$-body particle distributions were evolved to $z = 0$ using GADGET2 (and subsequently analysed with AHF).

In Fig. 8, we present the 2PCF of those five pairs of catalogues (with HALOGEN as lines, and $\text{AHF as points}$). The HALOGEN lines are the average of five realizations of HALOGEN placement (maintaining the same phases) and the error bars show the HALOGEN variance. Given that the GOLIAT box size is rather small ($1h^{-1}$Gpc) and the error bars show the HALOGEN variance. Given that the GOLIAT box size is rather small ($1h^{-1}$Gpc), scales $r > \sim 60h^{-1}$Mpc are dominated by cosmic variance effects. This makes it easy to identify the signature of each set of ICs. Though the realizations are significantly different, we note that the HALOGEN catalogue follows the $N$-body result, and maintains the correct normalization at intermediate scales ($20h^{-1}$Mpc $< r < 50h^{-1}$Mpc).

We stress that the fitting procedure has only been performed once; all five cases used fixed parameters. The similarity of the goodness of fit in each case (as compared to that directly fitted to) demonstrates that the fitted $\alpha(M)$ is universal with respect to input seed. We note also that the HALOGEN variance is significantly subordinate to the cosmic variance.

To better appreciate the dominance of the cosmic variance in a more applicable scenario, we return to the BigMultiDark simulation. This has a reduced cosmic variance due to the larger volume, but has the disadvantage that we cannot run several $N$-body simulations of this magnitude. The blue line of Fig. 9 shows how the 2PCF of a single-run HALOGEN (neither HALOGEN nor cosmic variance has been averaged out) compares to the reference BigMultiDark catalogue when they have the same IC phases. We further show
used to obtain the random seeds. The first case corresponds to the original GOLAX used to obtain the $\alpha(M)$ relation whereas the following share the same setup besides the seed. All the HALOGEN lines have been averaged over five realizations and the error bars show the HALOGEN variance. Similarity in goodness of fit between the first case and the others indicates that the fitted $\alpha(M)$ is universal with respect to input seed.

The HALOGEN variance ($\sigma_H$) and cosmic variance ($\sigma_{\text{cosm}}$). The former has been computed as usual: running five realizations of HALOGEN on the same 2LPTC snapshot. For the latter, we run five 2LPTC snapshots with different IC seeds. In order to avoid mixing $\sigma_{\text{cosm}}$ and $\sigma_H$ for each of them we first averaged out HALOGEN variance by running five realizations of HALOGEN and $\sigma_{\text{cosm}}$ is computed as the dispersion of the five resulting ($\sigma_H$-free) lines. We find for all scales that the HALOGEN variance is dominated by the cosmic variance, $\sigma_H < \sigma_{\text{cosm}}$

### 6.2 Probability distribution function

A simple but powerful statistic for point particles is the probability distribution function (PDF), which is the distribution of particles per cell on a given scale. Though simple, it contains interesting information as it contains contributions from the entire hierarchy of $n$-point functions (Peebles 1980; Fry 1985; Saslaw 2000).

Covering the BigMultiDark simulation with meshes of various (regular) sizes, we show in Fig. 10 a histogram of the number of haloes per cell for both the HALOGEN and BigMultiDark catalogues; the cell size ranges from 2.5 to 10 $h^{-1}$ Mpc. We find good agreement, especially at lower numbers of $N_{\text{halo}}$/cell, where the contribution of non-linear scales is reduced. We note that the mesh used to calculate the PDF is not to be confused with the grid used by HALOGEN for the NGP density assignment.

### 6.3 Power spectrum

HALOGEN has been designed to recover the 2PCF $\xi(r)$ of a provided halo catalogue. As the power spectrum $P(k)$ is its Fourier Transform, it theoretically contains the same information. However, this information is distributed differently in the two functions and there is mode coupling when transforming from one to another; an error at a given scale in one of the magnitudes can propagate to an error at all scales in the other. So we expect to witness different strengths and weaknesses in $P(k)$.

In Fig. 11, we compare the power spectrum of the BigMultiDark FOF catalogue to the corresponding HALOGEN realization. We find agreement to 5 per cent across the scales $0.01 h \text{ Mpc}^{-1} < k < 0.3 h \text{ Mpc}^{-1}$, but note that smaller scales $k > 0.3 h \text{ Mpc}^{-1}$ ($r < 20 h^{-1} \text{ Mpc}$) are underestimated. This underestimation arises from the smallest scales of the 2PCF, $r < l_{\text{cell}}$, which integrate through higher scales in $P(k)$.

### 6.4 Correlation function in RS

Observed galaxies are not directly located in 3D space, but in 2D-angular ($\theta, \phi$) coordinates with redshift $z$ converted to a polar distance. However, such distances are modified by galaxies’ peculiar velocities–velocity components that are not due to the Hubble expansion. These modifications are encoded as redshift space distortions (RSD), and we can begin to account for them by assigning correct velocities to haloes.

Using the halo velocities, we can mimic this effect when calculating the 2PCF. We show the results of such an analysis in Fig. 12, in which the monopole of the 2PCF in RS is compared for the
We have presented a new method called HALOGEN for the construction of approximate halo catalogues. It consists of four major steps:

(i) create a distribution of particles in a cosmological volume using 2LPT and distribute them in a grid of cell size $l_{\text{cell}}$.

(ii) sample a theoretical HMF $n(> M)$ with a list of $N_h$ halo masses $M$ and order them in descending mass.

(iii) place the haloes at the position of particles with a probability dependent on the cell density and halo mass $P_{\text{cell}} \propto \rho_{\text{cell}}^{\alpha(M)}$. We select random particles within cells, respecting the exclusion criterion and conserving mass in cells (cf. Section 4).

(iv) assign the velocity of the selected particle to the halo through a factor $v_{\text{halo}} = f_{\text{vel}}(M) \times v_{\text{part}}$.

We noted the modularity of these steps and acknowledged alternatives for each of them. The 2LPT in step (i) provides us with the correct large-scale clustering at a low computational cost, while step (ii) reconstructs the HMF. The heart of HALOGEN is step (iii) where the mass dependent bias is modelled through the parameter $\alpha(M)$ that stochastically places more massive haloes in overdensities, recovering the correct 2PCF as a function of mass. We also preclude haloes from overlapping to match the small-scale behaviour of the two-point clustering. In the last step (iv), we remap particle velocities in order to obtain the correct halo velocity distribution.

We studied how the parameters of the method – $\alpha(M)$, $f_{\text{vel}}(M)$ and $l_{\text{cell}}$ – can be optimized and summarized the results in Table 3. Though HALOGEN needs a reference halo catalogue from an N-Body simulation to obtain $\alpha(M)$ and $f_{\text{vel}}(M)$, once they have been optimized for a given setup, HALOGEN can be used to generate a multitude of halo catalogues, allowing the quantification of cosmic variance.

The HMF is recovered by construction – with some negligible sampling noise – to the theoretical value. The two-point function at intermediate scales ($10 h^{-1} \text{Mpc} < r < 50 h^{-1} \text{Mpc}$, where the bias is controlled by $\alpha(M)$) can be obtained in a BigMultiDark-like simulation at the $\sim 2$ per cent level and to the 15 per cent level at BAO scales ($80 h^{-1} \text{Mpc} < r < 110 h^{-1} \text{Mpc}$, Fig. 9). In RS, the error at intermediate scales rises to $\sim 4$ per cent and remains at $\sim 15$ per cent at large scales (Fig. 12). The clustering has a mass-dependence, for which the accuracy is controlled by the number of bins in the $\alpha(M)$ fit (Fig. 3). The power spectrum can be recovered at the 5 per cent level in the range of scales 0.01 $h^{-1} \text{Mpc} < k < 0.3 h^{-1} \text{Mpc}$ (Fig. 11). The halo PDF is accurately reproduced at low $N_{\text{halo}}$/cell, but overpredicts the high-$N_{\text{halo}}$/cell tail where the contributions of non-linearities are higher (Fig. 10).

We remark upon the adaptability of HALOGEN to different setups: GOLiat and BigMultiDark have different characteristics (see Table 1) and HALOGEN can be used for both with little recalibration effort. This indicates that HALOGEN is not only capable of running on one specific boxsize, redshift or cosmology, which makes it a powerful tool for exploring the statistics of varying cosmologies etc.

We have also verified that changing the initial phases in 2LPTC for HALOGEN leads to changes in the correlation function (due to cosmic variance) that follow the N-body simulation. This implies that doing so will yield robust estimates of cosmic variance, over potentially hundreds to thousands of realizations.

We have demonstrated that HALOGEN is a powerful tool for modelling statistics of halo catalogues, and the effects cosmic variance on them. The most immediate application of HALOGEN is the generation of the many catalogues required to study the control of systematics and for computing covariance matrices for large galaxy surveys (e.g. DES, DESi, Euclid). However, it can conceivably be used for other applications involving the study of cosmic variance.

Future work will involve improvements to the method, for instance by exploring subcell adjustments (i.e. alternatives to the random choice inside cells) or by changing one of the four stages of HALOGEN (e.g. what happens if we use 3LPT, what is the best function for $G(\rho_{\text{cell}})$ in equation 4?). Furthermore, in this study we have neglected substructure and referred to a possible extension using HOD models. We anticipate a fully integrated HOD layer.

Figure 11. Power spectrum $P(k)$ of HALOGEN (blue line) and FOF (red line) for BigMultiDark. The bottom panel shows their ratio. The power spectrum has been computed using a $N = 1024^3$ mesh and corrected for shot noise as explained in Jing (2005).

Figure 12. 2PCF in redshift space (RS) for FOF (red points), and HALOGEN (blue line) of the BigMultiDark simulation. We also include in magenta the results of our catalogue without applying the velocity bias (i.e. $f_{\text{vel}} = 1$, ‘selected particles’) and find that a correct velocity bias is needed.

HALOGEN and BigMultiDark catalogues. To show the effect of our velocity transformation, we also include the 2PCF of the ‘selected particles’ in which the velocities were not transformed. The normalization and shape are significantly improved by the simple linear transformation (equation 9), and we find agreement to below 5 per cent per cent at intermediate scales.

7 CONCLUSIONS

We have presented a new method called HALOGEN for the construction of approximate halo catalogues. It consists of four major steps:

(i) create a distribution of particles in a cosmological volume using 2LPT and distribute them in a grid of cell size $l_{\text{cell}}$.

(ii) sample a theoretical HMF $n(> M)$ with a list of $N_h$ halo masses $M$ and order them in descending mass.

(iii) place the haloes at the position of particles with a probability dependent on the cell density and halo mass $P_{\text{cell}} \propto \rho_{\text{cell}}^{\alpha(M)}$. We select random particles within cells, respecting the exclusion criterion and conserving mass in cells (cf. Section 4).

(iv) assign the velocity of the selected particle to the halo through a factor $v_{\text{halo}} = f_{\text{vel}}(M) \times v_{\text{part}}$.
to the method in future releases, which will enable a more direct comparison to observed data.

ACKNOWLEDGEMENTS

SA and JGB acknowledge financial support from the Spanish MINECO under grant FPA2012-39684-C03-02 and Consolider-Ingenio ‘Physics of the Accelerating Universe (PAU)’ (CSD2007-00060). They also acknowledge the support from the Spanish MINECO’s ‘Centro de Excelencia Severo Ochoa’ Programme under Grant no. SEV-2012-0249.

SA is also supported by a PhD FPI-fellowship from the Universidad Autónoma de Madrid. He also thanks the ‘Estancias Breves’ programme from the UAM and the UWA Research Collaboration Award 2014 that supported his stay in ICRAR, where this project was born. He further thanks David Alonso for his advices at different stages of the project.

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. He further thanks Dinosaur Jr for the bug.

Part of this research was undertaken as part of the Survey Simulation Pipeline (SSimPL; ssimpl-universe.tk). The Centre for All-Sky Astrophysics (CAASTRO) is an Australian Research Council Centre of Excellence, funded by grant CE110100090.

The work was supported by iVEC through the use of advanced computing resources located at iVEC@Murdoch.

The MultiDark Database 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. The simulation and its FOF halo catalogue has been kindly made available to us courtesy Stefan Gottlöber, Anatoly Klypin, Francisco Prada, and Gustavo Yepes before its public release. We also acknowledge PRACE for awarding us access to resource Curie supercomputer based in France (project PA2259). Some computation were performed on HYDRA, the HPC-cluster of the IFT-UAM/CSIC.

This research has made use of NASA's Astrophysics Data System (ADS) and the arXiv preprint server.

REFERENCES

Alonso D., 2012, preprint (arXiv:1210.1833)
Bouchet F. R., Colombi S., Hivon E., Juszkiewicz R., 1995, A&A, 296, 575
Castander F. J. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 8464, The PAU camera and the PAU survey at the William Herschel Telescope. SPIE, Bellingham, p. 6
Chuang C.-H. et al., 2014, preprint (arXiv:1412.7729)
Coles P., Jones B., 1991, MNRAS, 248, 1
Colín P., Klypin A. A., Kravtsov A. V., 2000, ApJ, 539, 561
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Dawson K. S. et al., 2013, AJ, 145, 10
Friedman J., Dark Energy Survey Collaboration 2013, in AAS Meeting Abstracts, Vol. 221, The Dark Energy Survey: Overview. Am. Astron. Soc., Washington, DC, p. 335.01
Fry J. N., 1985, ApJ, 289, 10
Hockney R. W., Eastwood J. W., 1988, Computer Simulation Using Particles. CRC Press, Boca Raton, FL
Jing Y. P., 2005, ApJ, 620, 559
Klypin A., Yepes G., Gottloeber S., Prada F., Hess S., 2014, preprint (arXiv:1411.4001)
Knebe A. et al., 2011, MNRAS, 415, 2293
Knebe A. et al., 2013, MNRAS, 435, 1618
Knollmann S. R., Knebe A., 2009, ApJS, 182, 608
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Levi M. et al., 2013, preprint (arXiv:1308.0847)
Manera M. et al., 2013, MNRAS, 428, 1036
Monaco F., Theuns T., Taffoni G., 2002, MNRAS, 331, 587
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Moutarde F., Alimi J.-M., Bouchet F. R., Pellat R., Ramani A., 1991, ApJ, 382, 377
Murray S. G., Power C., Robotham A. S. G., 2013, Astron. Comput., 3, 23
Neyrinck M. C., 2013, MNRAS, 428, 141
Peebles P. J. E., 1980, Ann. New York Acad. Sci., 336, 161
Planck Collaboration XVI, 2014, A&A, 571, A16
Press W. H., Schechter P., 1974, ApJ, 187, 425
Sahni V., Shandarin S., 1996, MNRAS, 282, 641
Saslaw W. C., 2000, The Distribution of the Galaxies. Cambridge Univ. Press, Cambridge
Scoccimarro R., 1998, MNRAS, 299, 1097
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Skibba R. A., Sheth R. K., 2009, MNRAS, 392, 1080
Springel V., 2005, MNRAS, 364, 1105
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, J. Cosmol. Astropart. Phys., 6, 36
Tinker J. L., Weinberg D. H., Zheng Z., Zehavi I., 2005, ApJ, 631, 41
Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709
Watson W. A., Iliev I. T., D’Aloisio A., Knebe A., Shapiro P. R., Yepes G., 2013, MNRAS, 433, 1230
White M., Tinker J. L., McBride C. K., 2014, MNRAS, 437, 2594

This paper has been typeset from a TeX/LaTeX file prepared by the author.