Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **516,** 1062–1071 (2022) https://doi.org/10.1093/mnras/stac2219
Advance Access publication 2022 August 6

Solving small-scale clustering problems in approximate light-cone mocks

Alex Smith ,[1,2,3]★ Shaun Cole,[1] Cameron Grove ,[1] Peder Norberg[1,4] and Pauline Zarrouk[5]

[1] *Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*
[2] *IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France*
[3] *Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[4] *Centre for Extragalactic Astronomy, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*
[5] *Sorbonne Université, Université Paris Diderot, CNRS/IN2P3, Laboratoire de Physique Nucléaire et de Hautes Energies, LPNHE, 4 Place Jussieu, F-75252 Paris, France*

Accepted 2022 August 2. Received 2022 August 2; in original form 2022 June 17

ABSTRACT

Realistic light-cone mocks are important in the clustering analyses of large galaxy surveys. For simulations where only the snapshots are available, it is common to create approximate light-cones by joining together the snapshots in spherical shells. We assess the two-point clustering measurements of central galaxies in approximate light-cones built from the Millennium-XXL simulation, which are constructed using different numbers of snapshots. The monopole and quadrupole of the real-space correlation function is strongly boosted on small scales below 1 h^{-1} Mpc, due to some galaxies being duplicated at the boundaries between snapshots in the light-cone. When more snapshots are used, the total number of duplicated galaxies is approximately constant, but they are pushed to smaller separations. The effect of this in redshift space is small, as long as the snapshots are cut into shells in real space. Randomly removing duplicated galaxies is able to reduce the excess clustering signal. Including satellite galaxies will reduce the impact of the duplicates, since many small-scale pairs come from satellites in the same halo. Galaxies that are missing from the light-cone at the boundaries can be added to the light-cone by having a small overlap between each shell. This effect will impact analyses that use very small-scale clustering measurements, and when using mocks to test the impact of fibre collisions.

Key words: catalogues – galaxies: statistics – large-scale structure of Universe.

1 INTRODUCTION

The use of realistic mock galaxy catalogues is an essential requirement in the clustering analysis of galaxies in large galaxy surveys. In order to make redshift-space distortion (RSD; Kaiser 1987) and baryon acoustic oscillations (BAO; e.g. Cole et al. 2005; Eisenstein et al. 2005) measurements from the two-point clustering statistics, the theoretical models must be validated on mock catalogues, where the underlying cosmology is known. Additionally, the use of mocks allows the systematics affecting these measurements to be quantified, and for the development of methods to mitigate them (e.g. Smith et al. 2019; Sugiyama et al. 2020; DeRose et al. 2022). These allow us to place constraints on theories of dark energy, and theories of modified gravity (e.g. Guzzo et al. 2008).

Current and future galaxy surveys, such as the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016a,b; Abareshi et al. 2022), Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), *Roman Space Telescope* (Spergel et al. 2015), and *Euclid* (Laureijs et al. 2011), will map many millions of galaxies. In order to reach the required high precision BAO and RSD measurements, it is essential that the mock catalogues used are as accurate as possible.

Mock galaxy catalogues are constructed using the outputs of *N*-body simulations. Because of the large volumes required, dark matter-only simulations are typically used, which are then populated with galaxies. There are several methods that are commonly used to add galaxies to the dark matter haloes, such as the halo occupation distribution (HOD; e.g. Smith et al. 2017, 2020; Alam et al. 2020, 2021b; Rossi et al. 2021), where the probability a halo contains central and satellite galaxies depends on the halo mass. Subhalo abundance matching (SHAM; e.g. Rodríguez-Torres et al. 2016; Safonova, Norberg & Cole 2021) ranks the subhaloes in the simulation based on a property (e.g. circular velocity), placing the brightest galaxies in the most massive subhaloes (with scatter). Semi-analytic models (SAM; e.g. Cole et al. 2000; Benson & Bower 2010) describe the physics of galaxy formation and evolution, using analytic techniques.

To emulate an observed data set, it is necessary to create light-cone mocks, where the galaxy properties evolve with the distance to the observer. Distant haloes or galaxies in the light-cone are output at early times in the simulation, and nearby galaxies at late times. Ideally, a direct light-cone output of the simulation would be used, where particles are output on the fly at the time at which they cross the observer's light-cone. For some simulations, light-cone outputs are available, such as the *Hubble* volume simulation (Evrard et al. 2002), the *Euclid* flagship simulation (Potter, Stadel & Teyssier 2017), and the DESI AbacusSummit simulations (Maksimova et al. 2021;

★ E-mail: alex.smith@ed.ac.uk

© The Author(s) 2022.
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

Hadzhiyska et al. 2022), which are designed to meet the requirements of the new generation of large-scale structure surveys. However, for many simulations, the positions of particles and haloes are only output in the cubic box, at certain discrete time snapshots. In this case, the simulation snapshots must be used to build approximate light-cones. Creating approximate light-cones from the simulation snapshots also provides more flexibility, allowing the observer position to be chosen after the simulation has been run, e.g. in an analogue of the Local Group. Multiple light-cones can also be created with observers at different locations in the box. Creating light-cones on-the-fly is a very specialized use of a simulation, increasing the input/output (I/O) and storage requirements, when simulation boxes are used for most applications. Additionally, light-cones currently available do not cover the full sky (e.g. the AbacusSummit light-cones cover one octant), so the snapshots can be used to create mocks that cover a larger footprint.

The simplest way to create a mock with galaxies positioned on the sky, cut to the survey geometry, is to use a single simulation snapshot, at the median redshift of the galaxy sample being considered. In the final analysis of the extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016; Alam et al. 2021a), this is what was done to create mocks for the different galaxy tracers (Smith et al. 2020; Alam et al. 2021b; Rossi et al. 2021). These mocks were used in the eBOSS mock challenges, to validate the theoretical models used in the two-point clustering analyses. While this was good enough for the precision of eBOSS, these mocks lacked evolution over the redshift ranges of the tracers. To include this evolution, multiple snapshots can be used to build a light-cone.

There are two ways in which multiple snapshots can be combined to make a light-cone, each with their own advantages and disadvantages. First, haloes can be interpolated between snapshots in order to build a light-cone (e.g. Merson et al. 2013; Smith et al. 2017; Izquierdo-Villalba et al. 2019). Interpolation has also been used to make LSST mocks in Korytov et al. (2019). This requires the use of halo merger trees to easily identify the descendants and progenitors of a halo. However, halo interpolation is not perfect, and there are additional complications, such as halo mergers that take place between snapshots. Different methods for interpolating haloes are compared in Smith et al. (2022). Halo merger trees require a lot of computational effort to produce, and are not always available. Particle IDs are required to be able to track the same halo at different times. For some large simulations, halo IDs are not tracked in order to maximize the number of dark matter particles (e.g. Potter, Stadel & Teyssier 2017), so merger trees cannot be produced.

The second method that can be used to build light-cones is to join together the snapshots in spherical shells. This is straightforward to implement for any simulation, requiring less computational effort than interpolation, but has the disadvantage that there are discontinuities at the boundaries between shells. Galaxies or haloes that cross the light-cone close to a boundary can be duplicated, appearing at both sides, or conversely they can never appear at all (e.g. Kitzbichler & White 2007).

The method of joining snapshots in spherical shells has been applied to many simulations to create light-cones and mocks for a wide range of applications. It is commonly used to create light-cones for studies of weak lensing, e.g. in the 'onion shell' Marenostrum Institut de Ciències de l'Espai (MICE) simulations (Fosalba et al. 2008, 2015), and in the lensing light-cones of Giocoli et al. (2016, 2017). A comparison of codes used to create lensing simulations can be found in Hilbert et al. (2020). Light-cones have also been created in spherical shells in lensing studies of the cosmic microwave background (e.g. Carbone et al. 2008; Sgier et al. 2021), and in studies of anisotropies in the gamma-ray background (e.g. Zavala, Springel & Boylan-Kolchin 2010; Fornasa et al. 2013). Other applications include mocks of active galactic nuclei (e.g. Comparat et al. 2019) and light-cones of galaxy clusters (e.g. Zandanel et al. 2018). Light-cones have been used in the Baryon Oscillation Spectroscopic Survey (BOSS) to predict the HOD and clustering of the constant-mass (CMASS) galaxies (Rodríguez-Torres et al. 2016), and have been created for the Dark Energy Survey (DES; Avila et al. 2018). Mocks for the upcoming *Roman Space Telescope* have been made in Wang et al. (2022). This method is also being used to create galaxy light-cones for the Sloan Digital Sky Survey (SDSS) and the DESI Bright Galaxy Survey (Dong-Páez et al. 2022).

While the issue of repeated or missing galaxies in the light-cones has been known about for some time, it is not accounted for in many of the light-cones that have been created. One way to correct for this is to linearly interpolate the galaxies that are close to the interface (Kitzbichler & White 2007). In this paper, we quantify the effect of duplicated galaxies on the two-point clustering statistics of light-cone mocks that are constructed in spherical shells. We compare light-cones constructed using different numbers of snapshots, and assess the impact of duplicated galaxies on the small-scale clustering. We propose a method to correct the issue of duplicated galaxies, which is simple to implement for any simulation, and does not require halo merger trees and interpolation.

This paper is organized as follows. The simulations and light-cone construction is described in Section 2. In Section 3, we compare light-cone mocks made with different numbers of snapshots, and assess the impact of duplicated galaxies on the small-scale clustering measurements. Our conclusions are summarized in Section 4.

2 LIGHT-CONES

In this paper, we create all-sky light-cones from the Millennium-XXL (MXXL) simulation. Galaxies are added to the simulation using the HOD methodology of Smith et al. (2017, 2022).

2.1 MXXL simulation

The MXXL simulation (Angulo et al. 2012) is a dark matter-only N-body simulation in a cubic box of side length 3 h^{-1} Gpc and particle mass $6.17 \times 10^9\ h^{-1}$ Mpc. The simulation was run in a 1-year *Wilkinson Microwave Anisotropy Probe* (*WMAP*1) cosmology (Spergel et al. 2003), with $\Omega_{\rm m} = 0.25$, $\Omega_\Lambda = 0.75$, $\sigma_8 = 0.9$, $h = 0.73$, and $n_{\rm s} = 1$.

Haloes in the simulation were first found using a friends-of-friends (FOF) algorithm (Davis et al. 1985), with linking length $b = 0.2$. Bound substructures within each FOF group were then identified using the SUBFIND algorithm (Springel et al. 2001).

Halo catalogues are output at a total of 64 simulation snapshots. There are 23 snapshots at $z < 1$, which are approximately evenly spaced in expansion factor, a.

2.2 Populating snapshots with galaxies

The MXXL simulation snapshots are populated with galaxies using the HOD methodology of Smith et al. (2017, 2022), where each galaxy is assigned an r-band magnitude. Here we briefly summarize this method. A set of nested HODs for different absolute magnitude thresholds is used, which are measured from the SDSS (Zehavi et al. 2011). For a given magnitude threshold, the HOD is modelled as a smoothed step function for central galaxies, and a power law for satellites. The smoothing of the step function that represents the

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

Table 1. MXXL simulation snapshots used in the construction of the light-cone mocks, built with a total of one, three, five, and nine snapshots. The redshift of the one-snapshot mock is closest to the median redshift of the galaxy sample we use.

Redshift	1 snapshot	3 snapshots	5 snapshots	9 snapshots
0			✓	✓
0.0199		✓		✓
0.0414			✓	✓
0.0644				✓
0.0892		✓	✓	✓
0.1159				✓
0.1444	✓		✓	✓
0.1749		✓		✓
0.2075			✓	✓

central galaxy HOD is performed using a pseudo-Gaussian spline kernel function rather than a simple Gaussian in order to prevent unphysical crossing of the HODs.

Central galaxies are placed at the centre of a halo, while satellites are randomly positioned following a Navarro–Frenk–White (NFW) profile (Navarro, Frenk & White 1997). The centrals are also assigned the same velocity as the halo, with a random virial velocity for the satellites, relative to the central. This is drawn from a Gaussian distribution in each dimension, with the velocity dispersion of the halo.

The HODs we use are evolved with redshift in order to match an evolving target luminosity function, from measurements from the SDSS and Galaxy And Mass Assembly (GAMA) surveys (Blanton et al. 2003; Loveday et al. 2012). When applied to a single snapshot, the mock that is produced will match exactly the target luminosity function at the redshift of the snapshot.

2.3 Creating light-cones

After populating the snapshots with galaxies, we create four full-sky light-cones by joining together the snapshots in spherical shells. This is done using one, three, five, and nine snapshots, where the snapshots used in each light-cone are summarized in Table 1. The joins between snapshots occur at the redshift exactly halfway between the snapshot redshifts. For the light-cone created using a single snapshot, we use the snapshot that is the closest to the median redshift of the volume-limited galaxy sample we examine in this work (see Section 2.5). Since the MXXL simulation is large, no periodic replications of the box are required.

When cutting each snapshot into spherical shells, there is a choice of whether this is done based on the real-space position of each galaxy, or based on the positions in redshift space. We therefore create two versions of each light-cone, to assess the impact of this choice on the galaxy clustering statistics.

When the light-cone is constructed from a single snapshot, the galaxy luminosity function is constant with redshift. For the other light-cones, there are sudden jumps in the luminosity function at the boundaries between snapshots. To make sure that all mocks have the same luminosity function, which smoothly evolves with redshift, we apply an abundance matching rescaling to the absolute magnitudes, as in Dong-Páez et al. (2022). The magnitude of a galaxy in a shell, M_r, can be converted to a corresponding cumulative number density, n, using the target luminosity function at the snapshot redshift, $z_{\rm snap}$. The target luminosity function at the redshift of the galaxy in the light-cone, z, is then used to convert the cumulative number density back a magnitude at redshift z.

2.4 Duplicated galaxies

In the light-cones built from multiple snapshots, it is possible for some galaxies to appear twice in the light-cone, or to never appear at all. This happens when a galaxy crosses the interface between two shells, as illustrated in Fig. 1. In the diagram on the left, the blue and red shaded regions indicate two shells in the light-cone, which are cut from neighbouring snapshots i and $i + 1$, at output times t_i and t_{i+1}. The points show the positions of galaxies at these two times. For the galaxy shown by the solid points, its position vector in the initial snapshot, $\boldsymbol{x}_i$, falls within the shell cut from snapshot i, so the galaxy appears in the light-cone. At time t_{i+1}, the galaxy has moved towards the observer, and its new position vector, $\boldsymbol{x}_{i+1}$, is inside the shell from snapshot $i + 1$. This galaxy appears twice in the light-cone, at each side of the boundary between shells. The opposite happens for the galaxy shown by the open circles. Its position at t_i falls outside of the first shell. The galaxy then moves away from the observer, and its new position at t_{i+1} is also outside of the second shell. This galaxy never appears in the light-cone. In reality, both galaxies should appear once in the light-cone. We can see from this figure that galaxies are only duplicated if they cross the boundary while travelling towards the observer, and missing galaxies always travel away from the observer. For the galaxies that appear twice, their pair separation is simply the distance travelled between the two snapshots.

The space–time diagram on the right of Fig. 1 illustrates the same two cases, with position on the x-axis and time on the y-axis. The galaxy that is moving towards the observer and appears twice in the light-cone is shown by the solid circles. Within each snapshot, the position of the galaxy is kept constant, which is indicated by the vertical lines. A galaxy in the shell from snapshot i will appear in the light-cone at a time in the range $(t_{i-1} + t_i)/2 < t < (t_i + t_{i+1})/2$, but its position is fixed at $\boldsymbol{x}_i$. At the boundary, the galaxy jumps instantaneously to its new position, $\boldsymbol{x}_{i+1}$, travelling faster than the speed of light. In this example, the galaxy crosses the light-cone in snapshot i, then jumps back over the light-cone, crossing a second time in snapshot $i + 1$. If the galaxy was interpolated, it would follow a smooth trajectory, indicated by the dashed line, crossing the light-cone once. Similarly, the open circles show the galaxy that never appears, since it instantaneously jumps over the light-cone at the interface where the shells are joined together.

2.5 Galaxy sample

We cut the mock to a volume-limited sample of $z < 0.2$ and $M_r < -20$, where the number density of galaxies is constant with redshift. The wide redshift range is covered by a total of nine snapshots, allowing us to investigate the effect of making light-cones from different numbers of shells. We also cut to central galaxies only, and do not consider the satellites. On small scales, most galaxy pairs are from satellite galaxies within the same halo, which will reduce the effect of duplicated galaxies on the clustering measurements. We focus on the central galaxies, where these effects will be strongest.

In this paper, we focus on a volume-limited galaxy sample, but we have checked that our results and conclusions remain unchanged for an apparent magnitude threshold galaxy sample.

3 COMPARING LIGHT-CONES WITH DIFFERENT NUMBERS OF SNAPSHOTS

In this section, we compare the light-cones constructed from different numbers of snapshots. Section 3.1 compares the two-point clustering

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

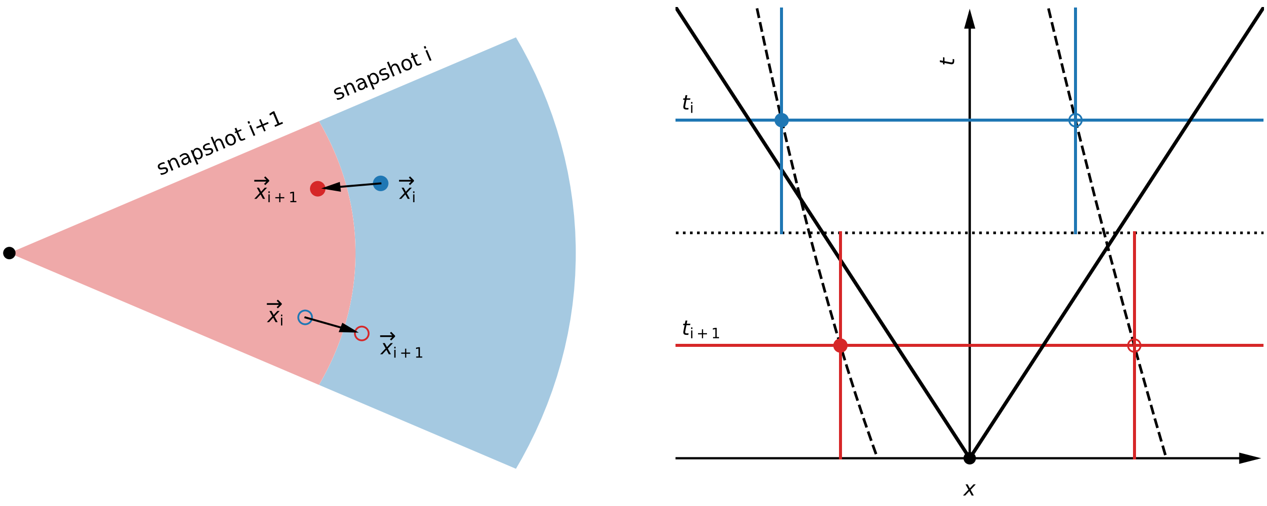

Figure 1. Left: diagram depicting galaxy positions in a two-dimensional slice through a light-cone, illustrating how the same galaxy can appear twice, or never appear. The blue and red shaded regions are shells cut from two neighbouring snapshots, i and $i + 1$, respectively. The circles show the positions of galaxies at the two snapshot output times, which cross the boundary. The galaxy indicated by the filled circle moves towards the observer, appearing in both shells. The galaxy indicated by the open circle moves away from the observer, appearing in neither. Right: a depiction of the same two cases in a space–time diagram, where time is shown along the vertical axis, the red and blue horizontal lines are the times corresponding the two simulation snapshots, and the diagonal lines represent the observer's light-cone. Within each snapshot, the position of each galaxy is kept fixed (vertical coloured lines), with an instantaneous jump between snapshots. Consequently, the galaxy depicted by the solid symbol, which is moving towards the observer, crosses the observer's past light-cone twice, while the one depicted by the open symbol does not cross it at all. If the galaxy trajectories had instead been interpolated (dashed lines) they would have each crossed the light-cone once. In this figure, the galaxy velocities have been greatly exaggerated. In reality, the dashed lines would be close to vertical and only galaxies very close to the light-cone at the interface between shells would be subject to these problems.

statistics, where the snapshots are carved into shells based on the real-space or redshift-space galaxy positions. Section 3.2 quantifies the distances that galaxies travel between snapshots, which sets the scales at which the clustering measurements are affected by duplicated galaxies. We assess the impact of removing duplicated galaxies on the two-point clustering statistics in Section 3.3.

3.1 Galaxy clustering

We measure the correlation function, $\xi(s, \mu)$, from the light-cones, where μ is the cosine of the angle between the line-of-sight direction and pair separation vector. This is then decomposed into Legendre multipoles,

$$\xi_l(s) = \frac{2l+1}{2} \int_{-1}^{1} \xi(s, \mu) P_l(\mu)\, \mathrm{d}\mu, \qquad (1)$$

where $P_l(\mu)$ is the lth order Legendre polynomial. In redshift space, the first two even multipoles are the monopole, $\xi_0(s)$, and quadrupole, $\xi_2(s)$, which are non-zero in linear theory. In real space, the monopole is equivalent to the real-space correlation function, $\xi(r)$, while the quadrupole is on average zero.

Fig. 2 shows the two-point correlation function in real space of the four light-cones, where the snapshots are cut into shells based on the real-space position of each galaxy. The monopole and quadrupole are shown in the upper panel, where the blue shaded region indicates the jackknife error from 100 jackknife regions. In real space, the quadrupole should be zero, and the signal measured in the mocks comes from cosmic variance in the finite volume. Since all four mocks are constructed from the same simulation, with the observer positioned at the same location, they all share the same large-scale structure, so the correlation function measurements have the same shape. The middle panel shows the difference in $\xi_0(r)$, relative to the mock built from a single snapshot, which is scaled by a factor of r to highlight any differences on large scales. We use the one-snapshot light-cone as the reference and differences from this are artefacts due to duplicated galaxies. Since the snapshot used is at the median redshift of the galaxy sample, the clustering measurements are a good approximation of a true light-cone. The grey shaded region is the jackknife error in $r\Delta\xi_0$, which provides an estimate of the noise when comparing simulations with the same initial conditions (see equation 14 of Grove et al. 2022). This error is calculated for all pairs of light-cones, and the average is plotted. On large scales, this noise is much smaller than the uncertainty due to cosmic variance, which is estimated with the standard jackknife error. On large scales, all four light-cones show good agreement in the monopole, and they remain in good agreement down to $\sim 1\ h^{-1}$ Mpc. Below this, the mocks built from multiple snapshots peel off, since their clustering is boosted by pairs of galaxies that are duplicated at the interfaces between shells. The scale at which this occurs is the smallest for the mock with nine snapshots, but for this light-cone, the effect on the monopole is also the strongest. The bottom panel shows the differences in $\xi_2(r)$. As with the monopole, all the mocks are in good agreement on large scales, but there is a non-zero signal below $1\ h^{-1}$ Mpc for the light-cones built from multiple snapshots, which again is the strongest and pushed to the smallest scales for the light-cone that uses all nine snapshots. The increase in the quadrupole indicates that the pair separation of duplicated galaxies is preferentially directed along the line of sight. This is because galaxies that travel along the observer's line of sight are more likely to cross a boundary between snapshots, and appear twice in the light-cone.

The left-hand panel of Fig. 3 shows the clustering in redshift space. Here, the snapshots were cut into shells using their redshift-space positions. Including the effect of velocities smooths out the effect of duplicated galaxies on the monopole. The clustering is still

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

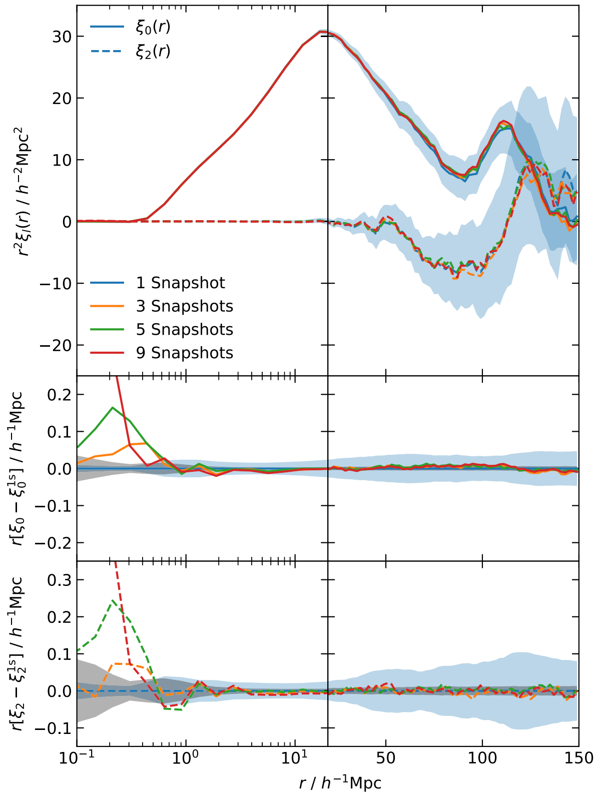

Figure 2. Correlation function multipoles in real space, scaled by r^2, for the light-cones built from one snapshot (blue), three snapshots (orange), five snapshots (green), and nine snapshots (red). The snapshots are cut into shells based on the galaxy positions in real space. The upper panel shows the monopole (solid lines) and quadrupole (dashed), transitioning from a logarithmic scale to a linear scale on the x-axis at 20 h^{-1} Mpc. The blue shaded region is the jackknife error, calculated using 100 jackknife samples. The grey shaded region indicates the jackknife error in $r\Delta\xi$, averaged over all pairs of mocks. The middle and lower panels show the differences in the monopole and quadrupole, respectively, compared to the mock built from a single snapshot, and scaled by r.

boosted for the light-cone with nine snapshots, but the increase in the clustering is smaller than in real space. However, this difference extends to larger scales, around 4 h^{-1} Mpc. A large effect is still seen in the quadrupole, which is also shifted to larger scales compared to in real space. A clear monotonic trend can be seen, where the strength of the quadrupole increases as more snapshots are included. Above ~10 h^{-1} Mpc, all the light-cones remain in good agreement.

When making the light-cones in redshift space, the shells were previously cut based on the redshift-space position of the galaxies in each snapshot. Alternatively, the cuts can be done based on the positions in real space, with the effect of velocities added to the mocks afterwards. The redshift-space clustering is shown in the right-hand panel of Fig. 3 for the light-cones where the cuts were applied in real space. The effect of applying the velocities at the end greatly reduces the effect that duplicated galaxies have on the clustering. For the monopole, the clustering measurements for all the light-cones are in good agreement down to very small scales of ~0.2 h^{-1} Mpc. Below this, a small boost in the clustering can only be seen for the mock made of nine snapshots. There is still an excess in the quadrupole at scales of ~1 h^{-1} Mpc, which is strongest for the nine-snapshot light-cone, but this is much smaller than when the snapshots were joined in redshift space.

3.2 Distance separation of duplicated galaxies

On small scales, the two-point clustering statistics are boosted due to galaxies that appear twice in the light-cone, at each side of the interface where two snapshots are joined together. The separation between the galaxies in each duplicated pair is simply the distance that the galaxy travelled in the time between the two snapshots. Therefore, the distribution of these distances will provide insight into how the clustering measurements are affected by the number of snapshots used to build the light-cones.

Fig. 4 shows the normalized distribution of distances that a central galaxy (i.e. a halo) travels between each of the simulation snapshots used to make the light-cone. This distribution is calculated from the full snapshots, using all central galaxies brighter than the magnitude threshold of $M_r < -20$, allowing the distributions to be measured smoothly. The distance that each galaxy travels is calculated by multiplying its velocity by the time interval, Δt, between the adjacent snapshots used to build the light-cone. The amplitude of the distribution is then scaled by a factor of r^2_{com}, where r_{com} is the comoving distance from the observer to the boundary in the light-cone, to take into account the differences in area (boundaries at high redshift cover a larger area, and hence there will be more galaxies that are duplicated).

The upper panel of Fig. 4 shows the distribution of the distances that galaxies travel between pairs of snapshots (i.e. the separations between duplicated galaxies) for the five-snapshot light-cone. The coloured curves show this distribution for each consecutive pair of snapshots, where the redshift at each boundary in the light-cone is indicated in the legend. The black curve is the total distribution (the sum of these). All the curves have been normalized to that the area under the sum (the black curve) is 1.

For the five-snapshot light-cone, the total distribution peaks at ~0.25 h^{-1} Mpc. Most duplicated pairs come from the highest redshift interface, at $z = 0.176$, since it covers a larger area in the light-cone. The number of duplicates is smaller at low redshifts, and the peak of the distribution also shifts to smaller scales. This is because the simulation snapshots are not evenly spaced, while the streaming velocities of haloes only evolve weakly with redshift. At low redshifts, the snapshots are spaced closer together, so there is less time in which the galaxies are able to travel.

The lower panel of Fig. 4 shows the total distributions (the sum of the coloured curves in the upper panel), comparing the light-cones that were made using three, five, and nine snapshots. When only three snapshots are used, the distribution peaks at ~0.4 h^{-1} Mpc, with a long tail extending to 1 h^{-1} Mpc. As the number of snapshots used is increased, the peak of the distribution shifts to smaller scales, since the Δt between snapshots at each boundary is smaller. When all nine snapshots are used, the peak shifts down to ~0.15 h^{-1} Mpc, with almost no pairs with separation above 0.5 h^{-1} Mpc. The total number of duplicated galaxies stays approximately constant when different numbers of snapshots are used. This is because if the number of interfaces is doubled, half as many galaxies cross at each one (since the Δt between snapshots halves, so each galaxy travels half the distance).

These observations are consistent with the clustering measurements in real space in Fig. 2. The scale at which the distributions peak (~0.4 and ~0.25 h^{-1} Mpc for three and five snapshots, respectively) is also where we see the largest difference in the monopole, compared to the single-snapshot light-cone. For the nine-snapshot light-cone, most duplicated pairs have separation less than ~0.4 h^{-1} Mpc, which is the same scale where the monopole peels upwards.

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

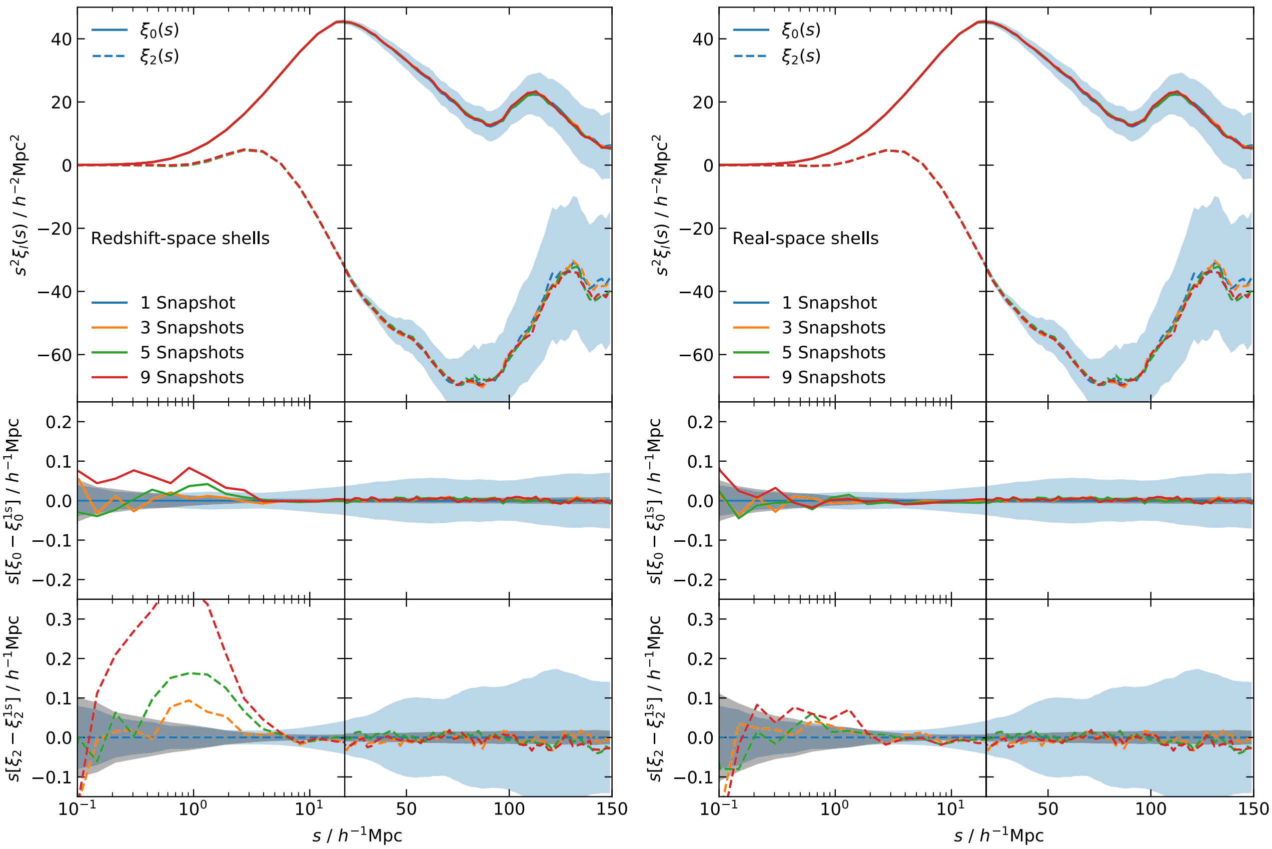

Figure 3. As Fig. 2, showing the correlation function multipoles in redshift space. The light-cones are constructed by cutting the snapshots in shells in redshift space (left) and in real space (right).

3.3 Removing duplicated galaxies

We now test the effect of removing the duplicated galaxies in the light-cone on the two-point clustering measurements. Removing these galaxies will lower the excess clustering signal on small scales, but it is not guaranteed that this will produce the correct clustering. This is because for every duplicated galaxy, there is also a galaxy that does not exist in the light-cone (see Fig. 1), which could potentially also have an effect on the measured clustering statistics.

We test the effect of removing the duplicated galaxies on the clustering measurements using the light-cone built from nine snapshots. Pairs of central galaxies in the sample with real-space separation $r < 0.3\ h^{-1}$ Mpc are identified, and one galaxy in each pair is randomly removed. Because of halo exclusion effects, there should be no pairs of central galaxies with these small separations, and all pairs are due to galaxies being duplicated in the light-cone.[1]

The ratio of the $n(z)$ of the galaxy sample is shown in Fig. 5, showing the fraction of galaxies that are removed, for the light-cone built from nine snapshots. We denote the redshifts in real and redshift space as $z_{\rm cos}$ (cosmological redshift) and $z_{\rm obs}$ (observed redshift), respectively. When the shells in the light-cone are cut based on the real-space galaxy positions, the reduction in the real space $n(z_{\rm cos})$ can only be seen in very narrow bins at the redshift of each boundary (the black curve). In redshift space, the dips in $n(z_{\rm obs})$ are broadened by the effect of velocities. The effect is the strongest when the shells are cut based on the redshift-space positions of galaxies (in blue), where the dips in $n(z_{\rm cos})$ are much deeper and broader than in real space. If the snapshots are instead cut into shells based on the real-space galaxy positions, with the velocities applied afterwards, the effect is much smaller (in red). In this case, the dips are also blueshifted to lower redshifts than the boundaries in the light-cone, since the velocity of the duplicated galaxies is always towards the observer (see Fig. 1).

After randomly removing one member of each pair, we recalculate the galaxy clustering. The random catalogue is generated by randomly sampling redshifts from the galaxies in the light-cone, and assigning random right ascension and declination coordinates that are uniformly distributed on the sky. A new random catalogue is generated after removing the duplicated galaxies, in order to take into account the small change in the $n(z)$.[2]

The small-scale clustering is shown in Fig. 6, for the light-cones with different numbers of snapshots. These measurements are the same as shown in Fig. 2, but without rescaling ξ by any factors of r to better see the differences between the curves on small scales. For the nine-snapshot mock, we show the clustering with and without duplicated galaxies included (in red and purple, respectively). The monopole is shown in the top left-hand panel, with the difference in the bottom left-hand panel, relative to the mock constructed from a single snapshot. Below $\sim 0.3\ h^{-1}$ Mpc, the monopole approaches $\xi_0 = -1$ for the mock built from one snapshot, showing that there

[1] We identified all close pairs in the light-cone, but this could be made faster by only considering galaxies close to the boundary.

[2] We have checked that if the original random catalogue is used, the effect on the clustering measurements is small.

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

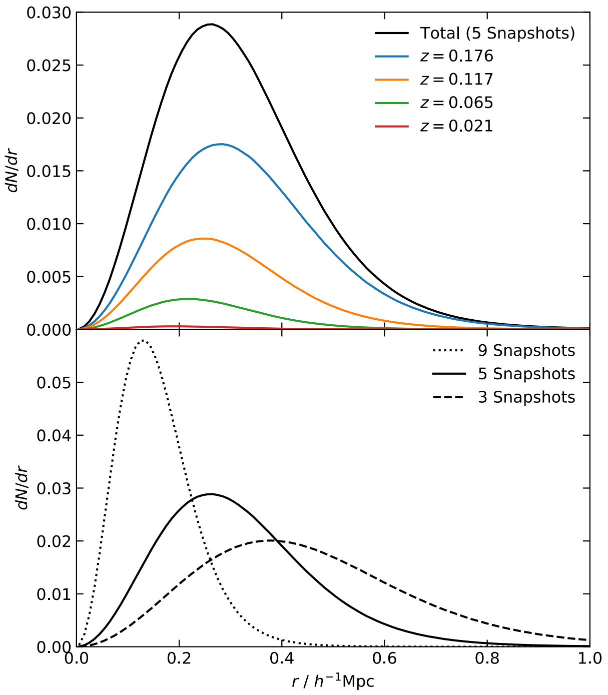

Figure 4. Top: normalized distribution of the distances, r, travelled by central galaxies at the boundaries between shells, for the light-cone constructed from five snapshots. The coloured curves show the distributions at the four boundaries, as indicated in the legend. The black curve is the total distribution (the sum of the coloured curves). The coloured curves are normalized so that the area under the black curve is 1. Bottom: total distributions (sum of the coloured curves in the top panel) for mocks built from nine snapshots (dotted curve), five snapshots (solid curve, which is the same as in the upper panel), and three snapshots (dashed curve).

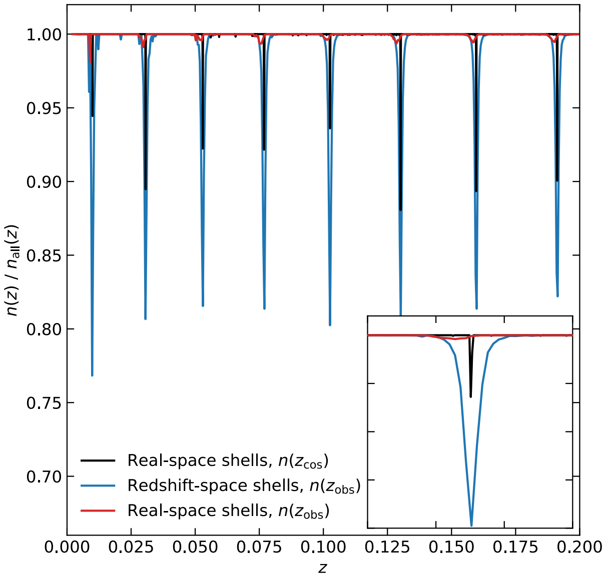

Figure 5. Ratio of the $n(z)$ of the galaxy sample, before and after the removal of duplicated galaxies, showing the fraction of galaxies that are removed. The black line is the ratio of the $n(z_{\rm cos})$ (i.e. in real space) from the light-cone with shells cut in real space. The blue and red lines show the ratios of the $n(z_{\rm obs})$ (i.e. in redshift space), from the light-cones with shells cut in redshift and real space, respectively. The smaller panel shows the dip in the $n(z)$ ratios at $z = 0.1$, but zoomed in on the x-axis.

are almost no pairs on these scales, due to halo exclusion effects. As more snapshots are included, the clustering gets stronger and stronger on small scales, and is strongest for the mock built using all nine snapshots. When the duplicated galaxies are removed, this reduces the clustering signal below 0.3 h^{-1} Mpc, bringing the monopole into agreement with the single-snapshot light-cone. The right-hand panel of Fig. 6 shows the quadrupole. On small scales, there is a non-zero signal due to the duplicated pairs, which is the strongest for the mock with nine snapshots. Removing the duplicates also removes this clustering signal, bringing it consistent with zero.

The clustering in redshift space is shown in the top of Fig. 7. This is for a light-cone where the snapshots were cut into shells based on the position of each galaxy in redshift space, including the effect of velocities. For the monopole in the left-hand panel, we see a similar trend as in real space, where the clustering on small scales is strongest for the light-cone with nine snapshots, but by a smaller amount than in real space. Removing duplicates reduces the clustering, bringing it into better agreement with the single snapshot. The same is seen in the quadrupole in the right-hand panel.

The bottom of Fig. 7 shows the clustering in redshift space again, but for the light-cones where the snapshots were cut into shells in real space, and the effect of velocities was applied after making the light-cone. As was also seen in Section 3.1, the clustering of the nine-snapshot light-cone is in much better agreement with the one-snapshot mock, compared to in real space, or compared to when the light-cones were constructed in redshift space. The boost in the monopole is only seen on very small scales (∼0.1 h^{-1} Mpc). The quadrupole also shows better agreement, but with a small excess on scales below ∼1 h^{-1} Mpc. As seen in the other mocks, removing the duplicated galaxies reduces the excess clustering signals, bringing the measurements into better agreement with the light-cone constructed from a single snapshot.

4 CONCLUSIONS

In the analysis of large galaxy surveys, it is essential to rely on realistic mock catalogues in order to validate theoretical models, and understand how the measurements are affected by systematics. As current and future galaxy surveys get larger, it is increasingly important to make the mocks as accurate as possible, creating light-cones that include redshift evolution. Ideally, light-cone mocks would be constructed from the light-cone output of an N-body simulation, but for many simulations, only snapshots outputs in the cubic box are available at discrete times. A common method of making approximate light-cones from the snapshots outputs is to join them in spherical shells. Making light-cones this way is computationally easy to do, but has the issue that there are discontinuities at the boundaries where two snapshots are joined. It is possible for a galaxy to appear twice, and be repeated at either side of one of the interfaces, or to not appear in the light-cone at all.

We test the accuracy of light-cone mocks constructed from snapshots using four all-sky light-cones constructed from the MXXL simulation. The galaxies in these light-cones are assigned r-band magnitudes, to match the evolving luminosity function measured from the SDSS and GAMA surveys. The light-cones we use are created using one, three, five, and nine snapshots, where the snapshots are cut in shells in either real space or redshift space. We measure the two-point clustering statistics of central galaxies in a volume-limited sample with $z < 0.2$ and absolute magnitude $M_r < -20$.

There is a boost in the monopole on small scales, due to galaxies that are duplicated at the boundaries between snapshots. In real space, this effect is larger as more snapshots are included, but is also shifted

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

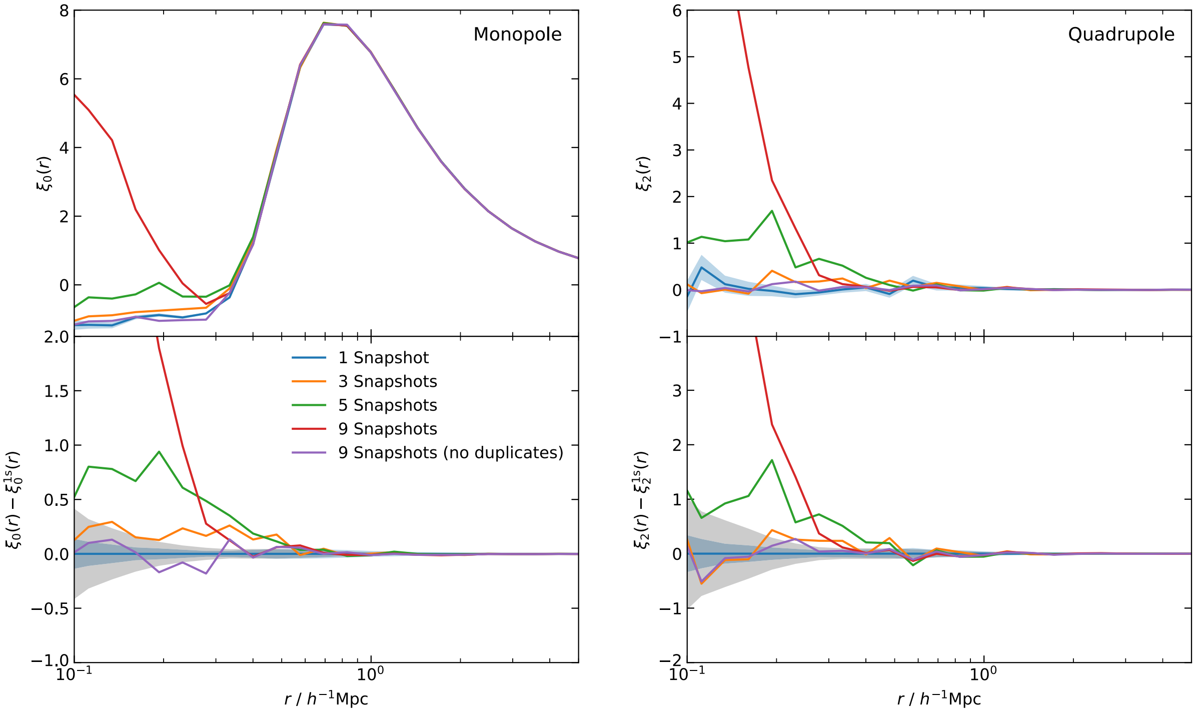

Figure 6. Left: real-space monopole on small scales, for mocks built from one snapshot (blue), three snapshots (orange), five snapshots (green), nine snapshots (red), and nine snapshots with duplicated galaxies removed (purple). The upper panel shows the monopole, while the lower panel is the difference relative to the mock built from a single snapshot. The blue shaded region is the jackknife error, using 100 jackknife regions, and the grey shaded region is the jackknife error in $\Delta\xi$. Right: like the left-hand plot, but showing the correlation function quadrupole.

to smaller scales. In redshift space, this effect is smoothed out by including the velocities. It is the smallest in the case that the snapshots are cut into shells in real space, with the effect of velocities applied afterwards. Similar effects are also seen in the quadrupole on small scales, which is also boosted by the inclusion of duplicated galaxies. The clustering is boosted on physical scales $\lesssim 1\ h^{-1}$ Mpc, and this scale depends on the distance that galaxies travel between the two snapshots.

We test the effect of randomly removing duplicated galaxies in the nine-snapshot light-cone on the two-point clustering measurements. This is done by identifying all pairs with a real-space separation $r < 0.3\ h^{-1}$ Mpc, and randomly removing one galaxy in each pair. On these scales there are no genuine pairs, due to halo exclusion effects. Both in real space and redshift space, this is able to reduce the excess small-scale clustering signal.

In this paper, we focus on central galaxies only, where the effect of duplicated galaxies is the strongest. Including satellites will reduce this, since the one-halo term dominates on small scales, and most pairs come from satellites within the same halo. However, there will also be some satellites that are duplicated at the boundaries in the light-cone. The impact of including satellitesx depends a lot on the galaxy sample, e.g. luminous red galaxies (LRGs) contain very few satellites. For the galaxy sample used in this paper, which has a 28 per cent satellite fraction, we have checked the impact of including satellites in real space. While the effect is smaller than when only centrals are used, there is still some excess small-scale clustering that is at a level greater than 1σ. When satellites are included, the same method can be used as before to identify duplicated central galaxies, but the randomly removed central galaxy would also have its satellites removed.

To summarize, in order to create light-cone mocks by joining snapshots in spherical shells, we propose using all available snapshots, joining them together in shells based on the real-space positions of galaxies. Galaxies that appear twice in the light-cone can be removed by identifying close pairs of centrals galaxies at the boundaries. However this does not take into account the galaxies that are missing in the light-cone. If each shell in the light-cone is made slightly wider, so that there is a small overlap in the volume at each boundary, the missing galaxies would be included in the light-cone. This would have the effect of increasing the number of spurious duplicates, but the same method we have employed can be used to remove them.

This only affects very small-scale clustering statistics, below $\sim 1\ h^{-1}$ Mpc. Large scales are unaffected, so for many applications of light-cone mocks, such as a BAO analysis, no correction is necessary. However for other applications where the small-scale clustering is important, such as a joint cosmology and HOD analysis, the results could potentially be affected by this systematic. Other applications, such as assessing the impact of fibre collisions, will also be affected by the spurious pairs of repeated galaxies.

ACKNOWLEDGEMENTS

AS would like to thank Francisco Prada, Anatoly Klypin, and Shadab Alam for helpful discussions. This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1, and ST/S002502/1, Durham University, and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

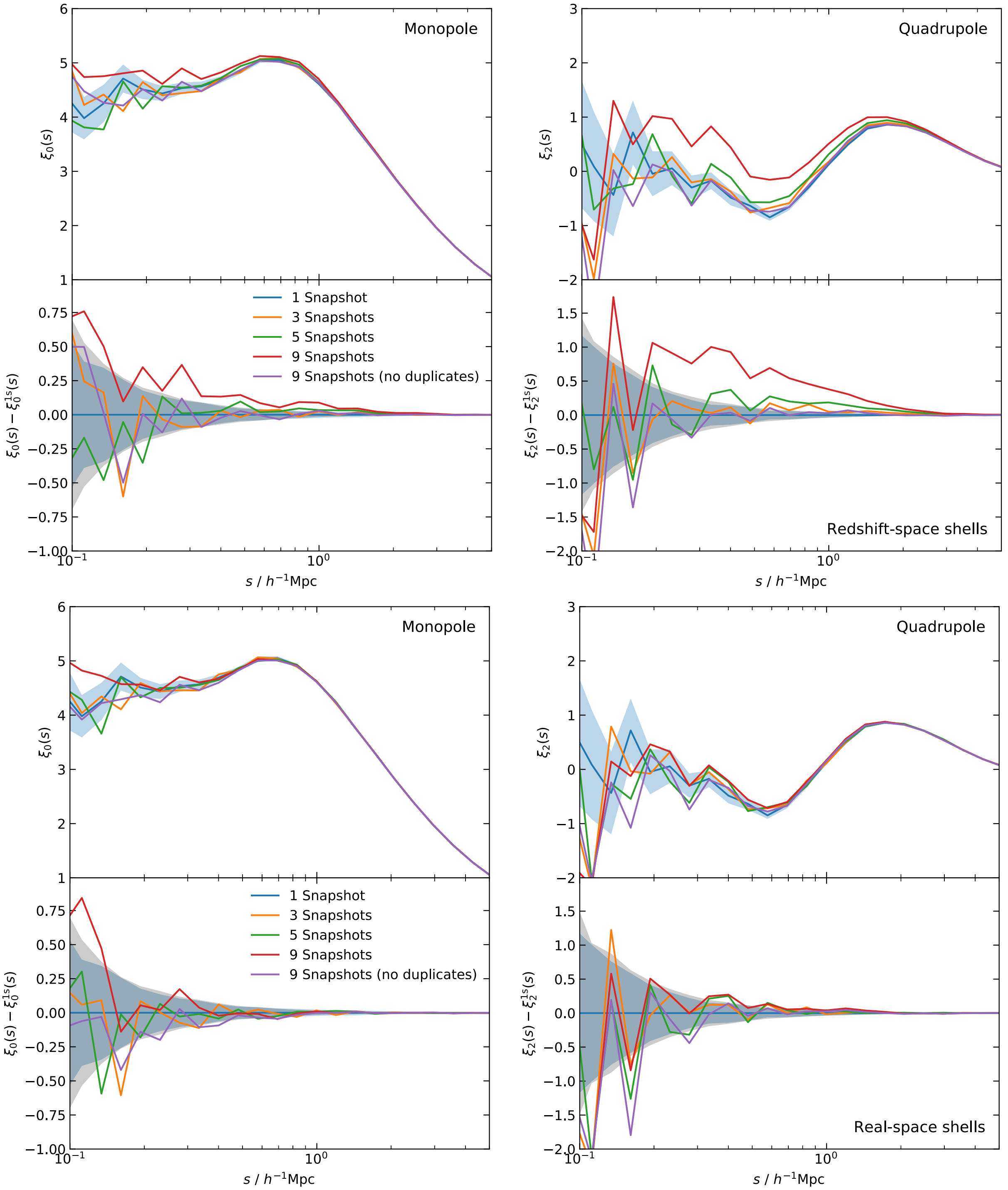

Figure 7. As Fig. 6, but showing the small-scale clustering in redshift space, where the spherical shells were joined in redshift space (top panels), and where the shells were joined in real space (bottom panels).

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

DATA AVAILABILITY

The light-cone mocks underlying this paper will be shared on reasonable request to the corresponding author.

REFERENCES

Abareshi B. et al., 2022, preprint (arXiv:2205.10939)

Alam S., Peacock J. A., Kraljic K., Ross A. J., Comparat J., 2020, MNRAS, 497, 581

Alam S. et al., 2021a, Phys. Rev. D, 103, 083533

Alam S. et al., 2021b, MNRAS, 504, 4667

Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, MNRAS, 426, 2046

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022

Avila S. et al., 2018, MNRAS, 479, 94

Benson A. J., Bower R., 2010, MNRAS, 405, 1573

Blanton M. R. et al., 2003, ApJ, 592, 819

Carbone C., Springel V., Baccigalupi C., Bartelmann M., Matarrese S., 2008, MNRAS, 388, 1618

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168

Cole S. et al., 2005, MNRAS, 362, 505

Comparat J. et al., 2019, MNRAS, 487, 2005

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371

Dawson K. S. et al., 2016, AJ, 151, 44

DeRose J. et al., 2022, Phys. Rev. D, 105, 123520

DESI Collaboration et al., 2016a, preprint (arXiv:1611.00036)

DESI Collaboration et al., 2016b, preprint (arXiv:1611.00037)

Dong-Páez C. A. et al., 2022, preprint (arXiv:2208.00540)

Eisenstein D. J. et al., 2005, ApJ, 633, 560

Evrard A. E. et al., 2002, ApJ, 573, 7

Fornasa M. et al., 2013, MNRAS, 429, 1529

Fosalba P., Gaztañaga E., Castander F. J., Manera M., 2008, MNRAS, 391, 435

Fosalba P., Gaztañaga E., Castander F. J., Crocce M., 2015, MNRAS, 447, 1319

Giocoli C. et al., 2016, MNRAS, 461, 209

Giocoli C. et al., 2017, MNRAS, 470, 3574

Grove C. et al., 2022, MNRAS, 515, 1854

Guzzo L. et al., 2008, Nature, 451, 541

Hadzhiyska B., Garrison L. H., Eisenstein D., Bose S., 2022, MNRAS, 509, 2194

Hilbert S. et al., 2020, MNRAS, 493, 305

Ivezić Ž. et al., 2019, ApJ, 873, 111

Izquierdo-Villalba D. et al., 2019, A&A, 631, A82

Kaiser N., 1987, MNRAS, 227, 1

Kitzbichler M. G., White S. D. M., 2007, MNRAS, 376, 2

Korytov D. et al., 2019, ApJS, 245, 26

Laureijs R. et al., 2011, preprint (arXiv:1110.3193)

Loveday J. et al., 2012, MNRAS, 420, 1239

Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, MNRAS, 508, 4017

Merson A. I. et al., 2013, MNRAS, 429, 556

Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493

Potter D., Stadel J., Teyssier R., 2017, Comput. Astrophys. Cosmol., 4, 2

Rodríguez-Torres S. A. et al., 2016, MNRAS, 460, 1173

Rossi G. et al., 2021, MNRAS, 505, 377

Safonova S., Norberg P., Cole S., 2021, MNRAS, 505, 325

Sgier R., Fluri J., Herbel J., Réfrégier A., Amara A., Kacprzak T., Nicola A., 2021, J. Cosmol. Astropart. Phys., 02, 047

Smith A., Cole S., Baugh C., Zheng Z., Angulo R., Norberg P., Zehavi I., 2017, MNRAS, 470, 4646

Smith A. et al., 2019, MNRAS, 484, 1285

Smith A. et al., 2020, MNRAS, 499, 269

Smith A., Cole S., Grove C., Norberg P., Zarrouk P., 2022, preprint (arXiv:2207.04902)

Spergel D. N. et al., 2003, ApJS, 148, 175

Spergel D. et al., 2015, preprint (arXiv:1503.03757)

Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726

Sugiyama S., Takada M., Kobayashi Y., Miyatake H., Shirasaki M., Nishimichi T., Park Y., 2020, Phys. Rev. D, 102, 083520

Wang Y. et al., 2022, ApJ, 928, 1

Zandanel F., Fornasa M., Prada F., Reiprich T. H., Pacaud F., Klypin A., 2018, MNRAS, 480, 987

Zavala J., Springel V., Boylan-Kolchin M., 2010, MNRAS, 405, 593

Zehavi I. et al., 2011, ApJ, 736, 59

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

Downloaded from https://academic.oup.com/mnras/article/516/1/1062/6657650 by University of Durham user on 17 November 2022