The Halo-Galaxy Connection from the Large Scale Structure of the Universe

by

Sergio Adrián Rodríguez Torres

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Theoretical Physics

Supervised by

Prof. Francisco Prada & Prof. Gustavo Yepes

Departamento de Física Teórica
Instituto de Física Teórica

Dedicated to my parents,
my support on this wonderful
journey we call life

Contents

Acknowledgements

I would like to thank my supervisors, Prof. Francisco Prada and Prof. Gustavo Yepes, for the guidance and advice they have provided throughout my time as their student, also for the opportunity to take part in many interesting projects. Special thanks also to Prof. Uros Seljak for hosting and helping me during my visit to Berkeley. I must express my gratitude to Prof. Anatoly Klypin for his personal and scientific support and his insightful discussions and suggestions.

I have greatly benefited from the SDSS-III and SDSS-IV collaborations, which allowed me to be part of many exciting projects. Specially, I wish to thank Dr. Florian Bleuter, Dr. Joel R. Brownstein, Dr. Etienne Burtin, Dr. Adreu Font , Dr. Hector Gil-Marín, Dr. Hong Guo, Dr. Claudia Maraston, Dr. Ashley Ross, Dr. Jeremy Tinker and Dr. Rita Tojeiro for their invaluable support on this period. I gratefully acknowledge the funding received towards my PhD from the Campus de Excelencia Internacional and the Departamento de Física Teórica de la Universidad Autónoma de Madrid.

My sincere thanks also goes to Dr. Joahn Comparat and all the MD-PATHY team (Dr. Francisco Kitaura, Dr. Chia-Hsun Chuang, Cheng Zhao, Prof. Francisco Prada and Prof. Gustavo Yepes) for sharing their knowledge and experience in different projects during my PhD.

Unsurprisingly, I am infinitely grateful with my family, none of this would have been possible without them. My parents, with love, support each step in my life and help me to grow at each moment. They are the reason of my success. My parents and my sisters have made me feel supported at all times despite the distance. Thanks to my father for our conversations every day, which made distance disappear, even if we have an ocean between us.

I have had the support and encouragement of the Adenisios family. Sylvie, Carlos Mario and Daniel have been very welcoming since the first day we met and they have made me feel as another member of the family. I owe my deepest gratitude to Nicole, she has become one of the most important people in my life. We have shared wonderful moments, travels and laughs, all of which I will never forget. I thank her for the support, love and patience during all these years.

Last, but never least, thanks to Javi, Santi, Antonio, Ginevra, Doris, Franco and Fernando for all the useful discussions and specially for sharing great moments.

Abstract

Large-scale structure is one of the most important fields in cosmology. It allows us to study the evolution of the Universe using the distribution of galaxies on the sky. Baryon acoustic oscillation (BAO) and redshift-space distortions (RSD) analyses provide valuable information about the physical processes that produce the Universe we see today. The Sloan Digital Sky Survey (SDSS) has been observing the local universe to study these phenomena. In the cosmological precision era, this experiment has measured more than a million galaxies which allow us to have a more precise vision of the evolution of structures.

In this thesis I present a study of the clustering of Luminous Red Galaxies (LRG) and Quasars (QSO) in the Baryon Oscillation Spectroscopic Survey (BOSS) of the SDSS-III and the SDSS-IV extended Baryon Oscillation Spectroscopic Survey (eBOSS) respectively. Clustering results from the CMASS LRG sample are compared with predictions of the Halo Abundance Matching (HAM) scheme. This model is applied to the BigMultiDark N-body simulation using a flat ΛCDM model. We construct high fidelity mocks including the evolution of the dark matter field in light-cones and observational effects such as incompleteness, geometry, veto masks and fibre collisions. These catalogues are a proof that the ΛCDM model, which describes the CMB (Planck1 cosmological parameters), can also predict the LRG clustering at $z \sim 0.5$. The two-point correlation functions and power spectrum are in agreement with observations within 1σ for all scales from 0.5 to 150 h^{-1}Mpc. Moreover, the three-point correlation function and the stellar-to-halo mass relation also present a good agreement with observations. Combining the potential of our model and the PATCHY code, we show the different steps of the construction of galaxy mocks of the BOSS Final Data Release (DR12). These mocks are used to construct the covariance matrices for the analysis of BAO and RSD. We provide a large set of catalogues which represent a step forward due to their excellent agreement with observations. This large set of ∼12 thousands mocks is the largest ever simulated volume corresponding to $\sim 192,000$ $[h^{-}\text{Gpc}]^3$.

Quasars also are playing a key role in the study of the Universe at high redshift. Following the LRG analysis, we model the eBOSS first year sample of spectroscopically confirmed quasars in the redshift range $0.9 < z < 2.2$. In this case, we use a modified halo abundance matching model applied to the BigMultiDark simulation to generate QSO high fidelity mocks. Our model reproduces the two-point statistics with good agreement. The typical quasar halo mass found is $\sim 10^{12.7} M_{\odot}$. There are still open questions about the distribution of quasars inside the halo, which new data and this kind of models will help us understand.

The different methods to populate dark matter halos with galaxies and quasars explained in this work successfully describe the observational data, and combined with best prediction of the cosmological parameters provide a proof of the validity of the ΛCDM model. Our methodology gives excellent results when applied to N-body simulations or simulations constructed with Lagragian perturbation theory. These models can be used to describe other populations such as ELG and can allow us to make predictions for future surveys (e.g. DESI, EUCLID).

Resumen

La estructura a gran escala es uno de los campos más importantes en Cosmología. Ésta nos permite estudiar la evolución del Universo a partir de la distribución de galaxias. El análisis de las oscilaciones acústicas bariónicas (BAO) o de las distorsiones en el espacio por el corrimiento al rojo (RSD) proporcionan información muy valiosa sobre los procesos físicos que produjeron el Universo que vemos hoy. El Sloan Digital Sky Survey (SDSS) ha estado observando el universo local para estudiar estos fenómenos. En la era de la precisión cosmológica, este experimento ha medido la distancia a más de un millón de galaxias permitiendo construir una visión más precisa de la evolución de las estructuras en la distribución de galaxias.

En esta tesis, se presenta un estudio del agrupamiento de las galaxias luminosas rojas (LRG) y de los cuásares (QSO) en el Baryon Oscillation Spectroscopic Survey (BOSS) del SDSS-II y extended Baryon Oscillation Spectroscopic Survey (eBOSS) del SDSS-IV respectivamente. Los resultados de la muestra de LRG del CMASS de BOSS son comparados con las predicciones del esquema *Halo Abundance Matching* (HAM). Este modelo es aplicado a la simulación de N-cuerpos BigMultiDark, la cual es construida en base a un modelo cosmológico ΛCDM plano con los parámetros medidos por la colaboración Planck. Se producen catálogos simulados de alta precisión incluyendo la evolución del campo de materia oscura en conos de luz que incluyen efectos observacionales como la incompletitud, la geometría del cartografiado, las mascaras veto y las colisiones de fibras. Estos catálogos prueban que el modelo ΛCDM, el cual describe el CMB (con los parámetros cosmológicos Planck1), también puede predecir el agrupamiento de las LRG para $z \sim 0.5$. La función de correlación de dos puntos y del espectro de potencias reproducen las observaciones dentro de errores de 1σ para escalas entre 0.5 y 150 h^{-1}Mpc. Adicionalmente, la función de correlación de tres puntos y la relación entre masa estelar y masa del halo presentan un buen acuerdo con las observaciones. Combinando el potencial de nuestro modelo y el código PATCHY, presentamos los diferentes pasos en la producción masiva de los catálogos de galaxias simuladas para la entrega final de datos de BOSS, los cuales son utilizados para construir las matrices de covarianza en los análisis de BAO y RSD. Nuestro productos representan un paso adelante debido a la

excelente precisión con la que reproducen las observaciones. Este conjunto de unos 12 mil catálogos cubren el volumen más grande simulado hasta hoy, unos $\sim 192,000$ $[h^{-}\mathrm{Gpc}]^3$.

Los QSO están jugando un papel importante en el estudio del universo a alto z. Siguiendo el análisis de las LRG, hemos modelado los datos correspondientes al primer año de los QSO espectroscópicos de eBOSS en el rango $0.9 < z < 2.2$. En este caso, hemos utilizado una versión modificada del modelo HAM, la cual hemos aplicado a la simulación BigMultiDark para generar los catálogos QSO de alta precisión. Nuestro modelo reproduce las estadísticas de dos puntos. La masa típica de los halos que alojan QSO es $\sim 10^{12.7} M_{\odot}$. Existen preguntas abiertas sobre la distribución de los QSO en los halos, para las cuales son necesarios nuevos datos y este tipo de modelos ayudarán a responderlas en el futuro.

Los diferentes métodos explicados en este trabajo, utilizados para conectar halos de materia oscura con galaxias y cuásares, describen exitosamente los datos observacionales, y combinados con las mejores estimaciones de los parámetros cosmológicos, proporcionan una prueba de la validez del modelo ΛCDM. Nuestra metodología da excelentes resultados al ser aplicada a simulaciones de N-cuerpos y a simulaciones de teoría de perturbaciones lagrangianas. Estos modelos pueden ser usados para describir otras poblaciones de galaxias como las ELG y permitirnos hacer predicciones del agrupamiento de galaxias para futuros cartografiados (e.g. DESI, EUCLID).

CHAPTER 1

Introduction

Modern cosmology has become one of the most exciting research fields in Physics. In the last decades, new technology has allowed us to measure relics of the early universe such as the Cosmic Microwave Background (CMB) anisotropies (Smoot et al., 1992; Hinshaw et al., 2013; Planck Collaboration et al., 2014). Furthermore, we are also making up maps of the galaxy distribution up to 7 billion years back in time using large-scale redshift surveys(e.g. Cole et al., 2005; Drinkwater et al., 2010; Frieman and Dark Energy Survey Collaboration, 2013; Dawson et al., 2013). These experiments and the coming ones are introducing us in the era of precision cosmology. Using all these observational data, we are constraining our theoretical models in order to better understand the formation and evolution of structures in the Universe.

Einstein's general relativity is the most successful theory explaining the gravitational physics behind a large set of observables which covers a large range of scales (solar system, galaxy, cosmological distances). The assumption that our Universe is described by the Riemannian geometry and the "cosmological principle" (the Universe is homogeneous and isotropic at large scales) lead to the Friedmann-Lemaître-Robertson-Walker (FLRW) model, which provides the most accepted explanation for the formation and evolution of the Universe (e.g. Peebles, 1980; Dodelson, 2003). However, one of the controversial parts of this model comes from the mass-energy content of the Universe. An analysis of the current data within general relativity implies the existence of an unknown form of matter, that we call Cold Dark Matter (CDM), which represents $\sim$26 % of the total mass-energy content of the Universe. Additionally, recent measurements show an accelerated expansion of our Universe that is taken into account by introducing a cosmological constant Λ in the FLRW model. This new energy (called dark energy), which exerts a "force" opposite to gravity, can be understood as a fluid with negative pressure which represents the $\sim$69% of the total matter-energy of the Universe.

These components lead to the ΛCDM framework which is currently known as the standard cosmological model.

There is still a relevant set of unsolved questions which current and future surveys are trying to answer. In the next decades, we will try to comprehend the nature of dark energy and dark matter, as well as the large-scale structure of the Universe at it is seen today. A better knowledge of the dark components will enable us to better understand their relationship with normal matter (baryonic matter) and consolidate the standard model. This relation is of utmost importance to understand the role of dark matter in the formation and evolution of galaxies, which is another open question in modern cosmology. Furthermore, new data will allow us to grasp the physical mechanisms of the primordial universe which produced the distribution and properties of the galaxies we see today.

The Standard Cosmological Model

Observations support the idea of a universe expanding from a singularity called Big Bang. Approximately 14 billion years ago, the space-time began its expansion and the Universe moved from a hot and dense plasma to the cooled state we see today. In the first moments, the Universe was a soup of matter and energy which created and annihilated particle-antiparticle pairs in a fast process. With the expansion, the temperature became lower and the rate of creation-annihilation decreased. At this point, the remaining antiparticles-particles were annihilated in a process known as baryogenesis. At the end of this process, we found an excess of particles which allowed the formation of galaxies millions of years after. Physicists are still working on some unresolved questions about this imbalance between the amount of matter and antimatter in the early universe.

The protons and neutrons created in the baryogenesis combined into light elements as hydrogen, helium and lithium. This process is known as nucleosynthesis and happened in the first 10 minutes. In that era of the early universe, photons and baryons interacted very fast. This prevented the formation of neutral atoms and light was coupled in the plasma making the Universe opaque. After ∼400,000 years, in the recombination era, the Universe became big enough to decouple light and baryons forming the first neutral atoms. At that moment, the Universe was transparent and photons could travel freely producing the first information of the primordial Universe that we can measure today, the Cosmic Microwave Background (CMB; e.g. Smoot et al., 1992; Hinshaw et al., 2013; Planck Collaboration et al., 2014).

The CMB shows an image of the primordial distribution of matter in the Universe. We assume that these initial density fluctuations grew by gravitational instability forming the structure we see today. However, the CMB raises new questions. It shows a homogeneous and isotropic Universe with very tiny fluctuations, even if different regions were causally disconnected. This could be explained by a period of inflation (e.g. Liddle, 1998), where the space-time expanded extremely fast in the first 10^{-34} seconds.

The standard cosmological model has been successful in the prediction of the amount of primordial elements and can give an explanation for the large-scale structure seen today. In this framework, we can relate the primordial quantum fluctuation of the Universe with the distribution of galaxies. However, there are still many unanswered questions about the physical processes of galaxy formation and the connection between galaxies and their dark matter halos. For these reasons, different large surveys have been designed in the last years to go back in time and try to understand the processes that generate the abundance and distribution of galaxies we see today.

The Large-Scale Structure of the Universe

Galaxies in the Universe are homogeneously distributed at very large-scales. However, looking at smaller regions we can see that they are not uniformly distributed. Galaxies are residing in groups and clusters at scales smaller than 3 Mpc. These structures are connected by filaments longer than 10 Mpc. This excess of clustering in some parts also creates regions with very few number of galaxies known as voids. All these observed structures depend on the physics which made the initial perturbations grow (cosmological model) and the galaxy formation processes.

The study of the large-scale structure of the Universe requires a statistical analysis of a set of galaxies (Peebles, 1980). In order to understand the evolution of the fluctuations in the primordial density field, it is useful to translate the galaxy density, $\rho(\mathbf{x})$, in a dimensionless density contrast

$$\delta(\mathbf{x},t) = \frac{\rho(\mathbf{x},t) - \bar{\rho}(\mathbf{x},t)}{\bar{\rho}(\mathbf{x},t)} \tag{1.1}$$

where $\bar{\rho}(\mathbf{x})$ is the expected galaxy mean density. $\delta(\mathbf{x})$ is close to a Gaussian distribution at large scales and at early times.

One of the most used quantities to describe the clustering of galaxies is the two-point correlation function $\xi(r)$. This quantity includes the effects of clusters, voids and filaments, providing the measurements of the evolution of the Universe at different spacial scales. The two-point correlation function (2PCF) is defined as the excess probability, above a random Poissonian distribution, of finding a galaxy in a volume dV separated a distance r from another galaxy,

$$dP = n[1 + \xi(r)]dV, \tag{1.2}$$

where n is the mean number of galaxies per unit volume. Thus, $\xi(r)$ is defined as the average of the density contrast in two different points, $\xi(r_1, r_2) = \langle \delta(r_1)\delta(r_2) \rangle$. Clustering measurements can be also analysed in Fourier-space using the power spectrum,

$$P(k) = \int \xi(r) e^{i[k]\cdot \mathbf{r}} d^3 r. \tag{1.3}$$

Galaxy surveys give us information about the evolution of structures at different times of the Universe. Combining CMB and galaxy surveys we can have a direct proof of the formation history of the structure we see today. One of the most powerful studies can be made via the baryon acoustic oscillations (BAO). As we mentioned above, before recombination (∼400,000 years, z ∼1000) our Universe was ionised (hot and dense) and photons were providing pressure and restoring force. The combination of gravity and restoring force generated perturbation which oscillated as acoustic waves at that time. Once recombination was over, photons had a long mean free path to decouple from this plasma and the Universe became neutral and the acoustic oscillations also froze leaving a characteristic acoustic length scale (e.g. Hu and Sugiyama, 1996; Eisenstein and Hu, 1998),

$$r_{\text{BAO}} = \int_{z_{rec}}^{\infty} \frac{c_s(z)dz}{H(z)} \tag{1.4}$$

where c_s is the speed of the sound and $H(z)$ the Hubble function at redshift z given by

$$H(z) = H_0 \sqrt{\Omega_m^0 (1+z)^3 + \Omega_k^0 (1+z)^2 + \Omega_\Lambda^0}. \tag{1.5}$$

In this expression H_0 represents the Hubble constant, Ω_m^0, Ω_k^0 and Ω_Λ^0 are the mass, curvature and dark energy density parameters all of them at the present epoch. The curvature component Ω_k is equal to zero in a flat universe. Since recombination, perturbations have

been growing by gravitational instability imprinting the BAO scale also in the clustering of galaxies seen today. For this reason, once $r_{\rm BAO}$ is found in the CMB, it can be used as a standard ruler to measure cosmological parameters. In the two-point correlation function, the BAO length in the line-of-sight direction is given by

$$r_{\parallel} = \frac{cz}{H(z)}. \tag{1.6}$$

While in the transverse direction, related to an angular size $\Delta\theta$,

$$r_{\perp} = (1+z) d_A(z) \Delta\theta, \tag{1.7}$$

where $d_A(z)$ is the angular diameter distance,

$$d_A(z) = \frac{1}{1+z} \int_0^z \frac{c dz'}{H(z')}. \tag{1.8}$$

The BAO scale used as a standard ruler has become one of the most important observations in modern cosmology (Beutler et al., 2011; Blake et al., 2011; Alam et al., 2016; Ata et al., 2017). This quantity is very useful to constrain dark energy models as their interpretation requires a model to describe the formation and evolution of structures. Despite the high level of non-linear processes related to galaxy formation and the structures they form, the measurements of the BAO size provide geometrical information with a very low level of systematics and they are complementary to other measurements such as weak lensing or CMB.

On the other hand, the large-scale structure provides information about the distribution of galaxies and the underlying dark matter field, just as gravitational lensing measurements do. Due to gravitational effects, dark matter and baryons are following approximately the same structures. Today we use galaxies as biased tracers of dark matter. Clustering differences between both components can be parametrised by a scale dependent bias which is related to the two point correlation function by

$$b(r) = \sqrt{\xi_{gal}(r)/\xi_{DM}(r)}. \tag{1.9}$$

Contrary to small scales, where the effects of non-linear physics become larger, especially in the one halo term region ($<$1 Mpc), this relation is easily described in the linear regime.

Spectroscopic Surveys

In cosmology, one of the most important observational challenges is to measure the galaxy clustering with high precision, and then obtain information about the underlying matter distribution behind this signal. Measuring angular positions is a relatively easy process and it can be used for extracting the angular clustering from a galaxy sample. However, a big part of the information resides in the third spacial component, so it is not included in an angular only analysis. The radial distances from galaxies can be computed through the spectrum of the light coming from these objects. Spectroscopic surveys (e.g. CfA, 2dFGRS, SDSS) obtain information about the emission and absorption lines of galaxies and these spectra can be compared with rest-frame models in order to compute their redshifts and thus, their distances,

$$s(z) = \int_0^z \frac{c dz'}{H_0 \sqrt{\Omega_m (1+z')^3 + \Omega_\Lambda}}. \tag{1.10}$$

Photometric surveys (e.g. Frieman and Dark Energy Survey Collaboration, 2013) are another widely used method to extract the redshift of galaxies, in this case fitting the broadband colours to some template. The precision of this method depends on the quality of the photometry, as well as the number of bands. Extraction of redshifts from photometric surveys is faster and easier than with spectroscopic ones, but their errors are of the order of $0.05(1+z)$, whereas spectroscopic surveys allow us to estimate the redshift with a typical error between $\sim 0.001(1+z)$ and $0.0001(1+z)$. Spectroscopy allows for very precise measurements of the spatial distribution of galaxies and therefore a better estimation of the clustering signal along the line of sight.

A second complexity when extracting the real-space distribution of galaxies is that their movements due to peculiar velocities introduce redshift-space distortions. These effects increase significantly near clusters or groups of galaxies. However, the physical processes responsible for these distortions are known and they can be modelled to add more constraints to the cosmological parameters or to recover the real-space clustering. Redshift-space distortions depend on two main processes. The peculiar random velocities of galaxies inside a cluster ($\lesssim 1\ h^{-1}$Mpc) introduce a Doppler shift which elongates the structures along the line-of-sight (Fingers-of-God; Jackson, 1972). The second effect is due to galaxies falling onto structures that are still in formation. This causes a contraction of the structures at scales $\gtrsim 1\ h^{-1}$Mpc (Kaiser effect; Kaiser, 1987).

In the last twenty years, spectroscopic surveys have become key experiments in astrophysics and cosmology. The multi-object spectrographs on ground-based telescopes have allowed different Collaborations to obtain more precise maps of the distribution of galaxies in the Universe. In the low redshift regime, the two-degree Field Galaxy Redshift Survey (2dFGRS; Cole et al., 2005) and the Sloan Digital Sky Survey (SDSS; Eisenstein et al., 2005) were able to measure the peak of baryon acoustic oscillations for the first time in the local universe. Following the legacy of its predecessor, the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al., 2013), part of the SDSS-III project (Eisenstein et al., 2011), improved the accuracy of BAO measurements extending the number density and the redshift range ($z < 0.75$) to more than a million new galaxies on 10,000 square degrees of the sky.

BOSS was designed to continue the study of the LRG sample in a larger redshift range compared to SDSS-I/II. For this purpose, they defined two redshift ranges with different colour cuts to improve the selection of LRG, LOWZ ("Low redshift") increasing the number of galaxies from SDSS-I/II in the redshift range 0.15 to 0.43 and a new sample, CMASS ("constant mass"), covering the range $0.43 < z < 0.75$. Just as in other experiments, there are different factors in multi-object spectroscopy that can contaminate the spectra or make it impossible to recover information from some of the galaxies. In order to account for these problems, BOSS corrects the clustering signal using weights for each galaxy. The final weight for each galaxy is given by (Ross et al., 2012)

$$w_g = w_{sys}(w_{zf} + w_{cp} - 1), \tag{1.11}$$

where w_{zf} denotes the redshift failure weight and w_{cp} represents the close pair weight. $w_{sys} = w_{see}w_{star}$ accounts for the observed systematic relationships between the number density of observed galaxies and stellar density and seeing (weights w_{star} and w_{see}, respectively). In the CMASS sample, weights take into account the observed systematic relationships between the number density of observed galaxies and stellar density and also the seeing using systematic weights, w_{sys}. Additionally, if a galaxy has a nearest neighbour (of the same target class) with a redshift failure, w_{zf} increases by one. A feature of the fibre-fed spectrograph is that the finite size of the fibre housing makes it impossible to place fibres within 62 arcsec of each other in the same plate. This causes a number of galaxies not to have an assigned fibre and hence, there is no measurement of their redshift. Similar to redshift failures, if a galaxy has a nearest neighbour without redshift because of fiber collision, w_{cp} will increase by one.

As well as LRGs, quasars (QSO) and Emission Line Galaxies (ELG) are good tracers of

the dark matter field. For this reason, SDSS I/II/III provided samples of spectroscopically confirmed quasars (Pâris et al., 2014), as well as Lyman-α forest quasars. BOSS was able to measure the BAO scale using Lyα forest quasars (Font-Ribera et al., 2014; Delubac et al., 2015) at $z \sim 2.5$. Combining the potential of SDSS-III/BOSS and new photometric information to optimise target selection, the current extended Baryon Spectroscopic Survey (eBOSS) extends the BAO studies to higher redshift giving the first measurement of the BAO scale at $0.9 < z < 2.2$ (Ata et al., 2017). eBOSS will increase the sample of LRG and QSO and it will provide a new sample of ELG. In total, this survey will provide redshifts for 300,000 luminous red galaxies (LRG) in the redshift range $0.6 < z < 1.0$, a new sample of ∼200,000 emission line galaxies (ELG) at redshift $0.6 < z < 1.0$, more than 500,000 spectroscopically confirmed quasars at $0.9 < z < 2.2$ and ∼120 000 new Lyα forest quasars at redshift$z > 2.1$.

One of the next generation spectroscopic surveys will be the Dark Energy Spectroscopic Instrument (DESI; Schlegel et al., 2015). This ground-based telescope will be capable of measuring spectra from different galaxies simultaneously, thanks to a new multi-object spectrograph. DESI is one of the most ambitious ground-based experiments, with a large number of fibers that will give us much more information about the ELG sample. The ESA Euclid mission will be a space telescope which will provide imaging and spectroscopic data for ~ 50 millions of galaxies(Laureijs et al., 2011; Sartoris et al., 2016), making a more precise study of the formation and evolution of the structures in the Universe.

N-Body simulations

The CMB is the best picture of the early universe we have access to. Although we do not have a good understanding of the processes which happened before recombination, this picture allows us to confirm the homogeneity and isotropy with very small perturbation on the density field ($\delta \sim 10^{-5}$). These fluctuations are well described by the linear perturbation theory. However, when these initial perturbations start growing due to gravitational instability, physics in the most dense regions becomes highly non-linear and perturbation theory cannot describe these processes.

Cosmological simulations are the way to solve the equations that govern the gravitational evolution of the Universe, including non-lineal effects. In these simulations the matter field is described by a fixed number of particles in a given comoving volume. These particles evolve with different time-steps until redshift zero. Cosmological simulations provide the

phase-space outputs at different redshifts (snapshots).

Ideal simulations should reproduce observations for baryons and dark matter. However, the physical processes involved in the evolution of baryons are more complex and increase the computational cost of simulations. Additionally, there is still a lack of knowledge in the galaxy formation physics, making it difficult to compare observations and simulations since differences can come from either the gravity model or the galaxy physics processes. This hard work is done by hydrodynamical simulations, which combine gravity and baryonic physics to understand galaxy formation physics.

Hydrodynamical simulations have to solve the fluid equations including gravity and hydrodynamics. However, this is not enough to reproduce the observed galaxies population. Thus, these simulations include additional processes such as star formation feedback, black holes or AGN feedback. One of the disadvantages of these simulations is the large amount of computational resources they require. This forces us to use relatively small volumes in order to have enough resolution to resolve galaxy physics. Hydrosimulations go from single halo scale to a few hundreds Mpc. These volumes are very small compared to the current surveys that study the large scale structure of the Universe. EAGLE[1] (Schaye et al., 2015) and ILLUSTRIS[2] (Genel et al., 2014) are two of the most recent simulations including these physical processes. These simulations can resolve small galaxies ($\sim 10^8 M_\odot$) producing observables such as stellar mass functions, size relations or the abundance of early- and late-type galaxies which are in agreement with observations. However, they have very small volumes, [100 Mpc]^3 and [106.5 Mpc]^3 respectively, compared to observational projects such as BOSS ($\sim 10^{10}$ Mpc^3). This make it impossible to do a fair comparison between this kind of simulations and the observational data.

Since recombination, large-scale structure growth has been dominated by gravity and the effects of dark energy, while baryonic physics has affected small scales. Taking it into account, we can assume that all the matter in the Universe is dark and can include baryonic processes of the early Universe (such as BAO) in the initial conditions. This gives us a reasonable explanation of the structures we see today. These N-body simulations are computationally less expensive and can cover volumes of a few Gpc and they allow us to simulate volume comparable to the current surveys.

Primordial fluctuations in the Universe are well studied using the CMB and can be described

[1] http://icc.dur.ac.uk/Eagle/
[2] http://www.illustris-project.org

by a Gaussian random field, that is defined as a white noise convolved by a transfer function. There exist two commonly used codes which are capable of computing the transfer function of the initial density fluctuations, CAMB (Lewis et al., 2000) and CLASS (Lesgourgues, 2011). Once the initial conditions are set, the dark matter field can be described by collisionless particles which can be evolved using non relativistic Newtonian dynamics. Even assuming only gravity physics, those processes are computationally expensive because of the force computation in these simulations scales on time like N^2 which makes them very slow. Tree algorithms (e.g. Springel, 2005) and Particle Mesh codes (e.g. Klypin and Holtzman, 1997) deal with this problem by reducing the number of operations by time step significantly.

Galaxies supposedly live in dark matter halos. These objects can be found in N-body simulations using halo finder codes (for a comparison between different halo finders see Knebe et al., 2011), which can define halos/subhalos and their properties, such as halo masses, concentrations or circular velocities. Additionally, merger tree schema can be applied to these halos to trace their history along all the snapshots (for a comparison between a variety of merger tree codes see Avila et al., 2014). In most of the present work we use the BigMultiDark simulation, which is part of the MultiDark suite of simulations[3](Klypin et al., 2016), and has a 2.5 h^{-1}Mpc box size with 3840^3 particles. BigMultiDark was run with a flat ΛCDM model consistent with Planck1 cosmological parameters (Planck Collaboration et al., 2014). The Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement halo finder (ROCKSTAR Behroozi et al., 2013a) is used in this simulation. This code identifies spherical dark matter halos and subhalos using an approach based on adaptive hierarchical refinement of friends-of-friends groups in six phase-space dimensions and one-time dimension. ROCKSTAR creates particle-based merger trees. The merger trees algorithm (Behroozi et al., 2013b) is used to estimate different quantities along the history of each halo.

Connecting Dark Matter Halos and Galaxies

Information of the large-scale structure of the Universe comes to us from light sources such as galaxies or quasars. Thus, we have to be able to extract information about our cosmological model without seeing the dark matter component directly. In order to compare simulations and observations, we have to include galaxies in our model. As mentioned above, hydrodynamical simulations present problems when describing large-scale structures due to their

[3] http://www.cosmosim.org/

small boxes. Other methods such as Semi-Analytical Models (for a review see Baugh, 2006; White and Frenk, 1991; Somerville and Primack, 1999; Lacey et al., 2016; Henriques et al., 2015) try to include galaxies in dark matter simulations using the evolution of halos and modelling some galaxy formation physics. However, in terms of clustering they are not close enough to observations. An alternative is statistical methods which assume that galaxies are biased tracers of dark matter. These models include galaxies in the simulation without modelling the stellar physics. They connect galaxies to halos using some observables to reproduce observations. There are two widely used methods that we describe below.

i) Halo Occupation Distribution

The Halo Occupation distribution model (HOD; e.g. Jing et al., 1998; Peacock and Smith, 2000; Scoccimarro et al., 2001; Berlind and Weinberg, 2002; Cooray and Sheth, 2002; Zheng et al., 2005) uses the probability of having N galaxies of a given type in a dark matter halo with mass M, $P(N|M)$ to assign galaxies. This probability is commonly described by a five parameters formulation and is the sum of two components,

$$\langle N(M)\rangle = \langle N_{cen}(M)\rangle + \langle N_{sat}(M)\rangle\,. \tag{1.12}$$

The central contribution is given by (Zheng et al., 2007)

$$\langle N_{cen}(M)\rangle = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log M - \log M_{min}}{\sigma_{\log M}}\right)\right]. \tag{1.13}$$

M_{min}, is the minimum mass of halos that can host a central galaxy and $\sigma_{\log M}$ the width of the cutoff profile. In order to place satellite galaxies there are different formulations. One of them is given by

$$\langle N_{sat}(M)\rangle = \langle N_{cen}(M)\rangle \left(\frac{M - M_0}{M_1}\right)^{\alpha}, \tag{1.14}$$

where M_0 is the mass scale of the drop, M_1 characterises the amplitude and α is the asymptotic slope at high halo mass. Satellite galaxies can be located following the mass profile of the halo or using subhalos. This process can require additional information about the velocities of the galaxies in the catalogue.

ii) Halo Abundance Matching

Another widely used method is the Halo Abundance Matching technique (Kravtsov et al., 2004; Conroy et al., 2006; Behroozi et al., 2010; Guo et al., 2010; Trujillo-Gomez et al., 2011; Nuza et al., 2013; Reddick et al., 2013). This is a simple but very successful method for modelling the clustering of galaxies. The basic assumption of the model is that most massive halos host the most massive galaxies. However, this relation is not one to one, this rank ordering is mediated by a scatter between both populations, which is related to the bias of the galaxy sample. The intrinsic scatter is a quantity that could be extracted from observation using the circular velocity/velocity dispersion to stellar mass relation. However, it is very difficult to extract a value from these relations due to different sources of errors and large uncertainties in these measurements. For this reason the scatter is fixed by the clustering signal. In this work, we are using the two-point correlation function.

There are different implementation of the HAM model depending on the proxies used for galaxies and halos. Reddick et al. (2013) make a comparison between the different halo proxies, finding better results for the maximum circular velocity in the whole history of the halo, (V_{peak}). In the following sections we use this quantity as a proxy for halos and stellar mass for galaxies. In order to implement the model, one can define a new variable

$$V_{peak}^{scat} = [1 + \mathcal{N}(0, \sigma_{\rm HAM})]V_{peak}, \tag{1.15}$$

where $\mathcal{N}$ is a random number coming from a Gaussian distribution with mean 0 and standard deviation $\sigma_{\rm HAM}$. Using this variable, the stellar mass of galaxies (coming from the stellar mass function) and V_{peak}^{scat} from halos are rank ordered and linked in a one to one assignment.

As we already discussed, redshift-space distortions are present in galaxy surveys. In order to make a comparison between observations and simulations, we have to translate the simulated galaxies from real- to redshift-space using

$$\mathbf{s} = \mathbf{r}_c + \frac{\mathbf{v} \cdot \hat{\mathbf{r}}}{aH(z_{real})}, \tag{1.16}$$

where $\mathbf{v}$ represents the peculiar velocities of the galaxies, $\hat{\mathbf{r}}$ is the unitary line-of-sight vector and $\mathbf{r}$ the comoving distance of the galaxy. Velocities of the simulated galaxies can be added in different ways for HOD or HAM. In the present work we assume that galaxies have the same velocity as their (sub)halo.

Thesis Overview

This thesis is presented as a compendium of three publications and is organised as follows: In the last part of this Introduction, I present a briefly description of my contribution in each paper and an additional list of publications in which I co-authored during my PhD. Chapter 2 provides a summary of the main results and discussions of the three papers. The first paper is shown in Chapter 3 presenting a study of the BOSS LRG clustering. The MultiDark-Patchy (MD-PATCHY) mocks for BOSS are discussed in Paper II which is included in Chapter 4. In Chapter 5, a study of the eBOSS QSO clustering is presented (Paper III). Finally, general conclusions and outlines are provided in the last part of this thesis.

Authorship papers

This thesis is presented as a compendium of three major publications. Here is a summary of my contribution in each of the papers. Additionally, a list of publications in which I was also involved is included.

Paper I: MNRAS, 460, 1173-1187 (2016)

In this paper, I was the leading author doing all the analyses of the results. I compiled the contributions from all authors writing the present paper. In addition, the model and the catalogues were produced with the SUrvey GenerAtoR code (SUGAR), which I developed myself for this research. Additionally, I computed all the two-point correlation functions and produced a special set of mock catalogues in order to estimate the errors in the measurements. SUGAR includes the reading algorithms for the simulations and the construction of the survey geometry (including the survey masks). This code also is designed to implement the HAM method explained in this paper. SUGAR will be publicly available soon.

Paper II: MNRAS, 456, 4156-4173 (2016)

In this paper, we present the different steps used to construct the MultiDark-Patchy mocks for BOSS DR12. As is explained in the paper, all this work is based on three complementary

codes: PATCHY (Kitaura et al., 2014), HADRON (Zhao et al., 2015) and SUGAR. I was involved in two important steps of this project. I generated the reference boxes producing the phase-space parameters for the PATCHY bias model. These boxes were constructed from the BigMultiDark simulation and had to reproduce the clustering and number density of the BOSS sample for different redshift ranges. I used a variation of the model presented in Paper I to generate ten boxes at different redshift which modelled the LOWZ and CMASS samples. Once PATCHY and HADRON produced the different outputs of the simulation, it was necessary to put them together, constructing light-cones and to include observational effects such as the radial and angular selection functions, fiber collisions, survey masks, stellar masses, etc. I implemented all these processes using a variation of the SUGAR code. I also computed and analysed the two-point statistics in configuration space for the different wedges, redshift bins and stellar mass thresholds. I participated in the discussion of the general results and I was involved in the writing of the paper where I included all my contributions.

Paper III: MNRAS, 468, 728–740 (2017)

Just as in Paper I, I carried out all the analysis of the results, compiling results from other authors and writing the paper. I implemented the modified HAM in my code, as well as new observational effects. All the catalogues used were made using my SUGAR code. A special set of GLAM mocks (Klypin and Prada, 2017) was produced in order to reproduce the small scales of the simulated quasar catalogue.

Publications

The work done for the three major papers presented in this thesis allowed me to participate in the analysis and discussion of different projects within the SDSS-III and SDSS-IV collaborations. In addition, I was involved in the analysis of ELG samples and the proposal of the LOw Redshift survey at Calar Alto (LORCA). The following list presents 40 papers (including the three major papers) in which I was involved. To this date, these papers have 856 citations with an average citation of 32.4 each one[4].

[4] https://ui.adsabs.harvard.edu

Alam, S., Albareti, F. D., Allende Prieto, C., Anders, F., Anderson, S. F., Anderton, T., et al. (2015). The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. ApJS, 219:12.

Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., et al. (2016). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. preprint, (arXiv:1607.03155).

Ata, M., Baumgarten, F., Bautista, J., Beutler, F., Bizyaev, D., Blanton, M. R., et al. (2017a). The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: First measurement of Baryon Acoustic Oscillations between redshift 0.8 and 2.2. preprint, (arXiv:1705.06373).

Ata, M., Kitaura, F.-S., Chuang, C.-H., Rodríguez-Torres, S., Angulo, R. E., Ferraro, S., et al. (2017b). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmic flows and cosmic web from luminous red galaxies. MNRAS, 467:3993–4014.

Beutler, F., Seo, H.-J., Ross, A. J., McDonald, P., Saito, S., Bolton, A. S., et al. (2017a). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Fourier space. MNRAS, 464:3409–3430.

Beutler, F., Seo, H.-J., Saito, S., Chuang, C.-H., Cuesta, A. J., Eisenstein, D. J., et al. (2017b). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: anisotropic galaxy clustering in Fourier space. MNRAS, 466:2242–2260.

Blanton, M. R., Bershady, M. A., Abolfathi, B., Albareti, F. D., Allende Prieto, C., Almeida, A., et al. (2017). Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies and the Distant Universe. preprint, (arXiv:1703.00052).

Chuang, C.-H., Pellejero-Ibanez, M., Rodríguez-Torres, S., Ross, A. J., Zhao, G.-b., Wang, Y., et al. (2016a). The Clustering of Galaxies in the Completed SDSS-III Baryon Oscillation Spectroscopic Survey: single-probe measurements from DR12 galaxy clustering – towards an accurate model. preprint, (arXiv:1607.03151).

Chuang, C.-H., Prada, F., Pellejero-Ibanez, M., Beutler, F., Cuesta, A. J., Eisenstein, D. J., et al. (2016b). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: single-probe measurements from CMASS anisotropic galaxy clustering. MNRAS, 461:3781–3793.

Comparat, J., Chuang, C.-H., Rodríguez-Torres, S., Pellejero-Ibanez, M., Prada, F., Yepes, G., et al. (2016). The Low Redshift survey at Calar Alto (LoRCA). MNRAS, 458:2940–2952.

Dawson, K. S., Kneib, J.-P., Percival, W. J., Alam, S., Albareti, F. D., Anderson, S. F., et al. (2016). The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. AJ, 151:44.

Favole, G., Comparat, J., Prada, F., Yepes, G., Jullo, E., Niemiec, A., et al. (2016a). Clustering properties of g-selected galaxies at z $\sim$ 0.8. MNRAS, 461:3421–3431.

Favole, G., Rodríguez-Torres, S. A., Comparat, J., Prada, F., Guo, H., Klypin, A., et al. (2016b). Galaxy clustering dependence on the [OII] emission line luminosity in the local Universe. preprint, (arXiv:1611.05457).

Gil-Marín, H., Percival, W. J., Brownstein, J. R., Chuang, C.-H., Grieb, J. N., Ho, S., et al. (2016a). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the LOS-dependent power spectrum of DR12 BOSS galaxies. MNRAS, 460:4188–4209.

Gil-Marín, H., Percival, W. J., Cuesta, A. J., Brownstein, J. R., Chuang, C.-H., Ho, S., et al. (2016b). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: BAO measurement from the LOS-dependent power spectrum of DR12 BOSS galaxies. MNRAS, 460:4210–4219.

Gil-Marín, H., Percival, W. J., Verde, L., Brownstein, J. R., Chuang, C.-H., Kitaura, F.-S., et al. (2017). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies. MNRAS, 465:1757–1788.

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., Scoccimarro, R., Crocce, M., Dalla Vecchia, C., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the Fourier space wedges of the final sample. MNRAS, 467:2085–2112.

Guo, H., Zheng, Z., Behroozi, P. S., Zehavi, I., Chuang, C.-H., Comparat, J., et al. (2016a). Modelling galaxy clustering: halo occupation distribution versus subhalo matching. MNRAS, 459:3040–3058.

Guo, H., Zheng, Z., Behroozi, P. S., Zehavi, I., Comparat, J., Favole, G., et al. (2016b). Galaxy Three-point Correlation Functions and Halo/Subhalo Models. ApJ, 831:3.

Hahn, C., Scoccimarro, R., Blanton, M. R., Tinker, J. L., and Rodríguez-Torres, S. A. (2017). The Effect of Fiber Collisions on the Galaxy Power Spectrum Multipoles. MNRAS, 467:1940–1956.

Kitaura, F.-S., Ata, M., Angulo, R. E., Chuang, C.-H., Rodríguez-Torres, S., Monteagudo, C. H., et al. (2016a). Bayesian redshift-space distortions correction from galaxy redshift surveys. MNRAS, 457:L113–L117.

Kitaura, F.-S., Chuang, C.-H., Liang, Y., Zhao, C., Tao, C., Rodríguez-Torres, S., et al. (2016b). Signatures of the Primordial Universe from Its Emptiness: Measurement of Baryon Acoustic Oscillations from Minima of the Density Field. Physical Review Letters, 116(17):171301.

Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., Zhao, C., Prada, F., Gil-Marín, H., et al. (2016c). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: mock galaxy catalogues for the BOSS Final Data Release. MNRAS, 456:4156–4173.

Leauthaud, A., Saito, S., Hilbert, S., Barreira, A., More, S., White, M., et al. (2017). Lensing is low: cosmology, galaxy formation or new physics? MNRAS, 467:3024–3047.

Montero-Dorta, A. D., Bolton, A. S., Brownstein, J. R., Swanson, M., Dawson, K., Prada, F., et al. (2016). The high-mass end of the red sequence at z $\sim$ 0.55 from SDSS-III/BOSS: completeness, bimodality and luminosity function. MNRAS, 461:1131–1153.

Montero-Dorta, A. D., Perez, E., Prada, F., Rodriguez-Torres, S., Favole, G., Klypin, A., et al. (2017). Observational evidence of galaxy assembly bias. preprint, (arXiv:1705.00013).

Pellejero-Ibanez, M., Chuang, C.-H., Rubiño-Martín, J. A., Cuesta, A. J., Wang, Y., Zhao, G., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: towards a computationally efficient analysis without informative priors. MNRAS, 468:4116–4133.

Reid, B., Ho, S., Padmanabhan, N., Percival, W. J., Tinker, J., Tojeiro, R., et al. (2016). SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: galaxy target selection and large-scale structure catalogues. MNRAS, 455:1553–1573.

Rodríguez-Torres, S. A., Chuang, C.-H., Prada, F., Guo, H., Klypin, A., Behroozi, P., et al. (2016). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: modelling the clustering and halo occupation distribution of BOSS CMASS galaxies in the Final Data Release. MNRAS, 460:1173–1187.

Rodríguez-Torres, S. A., Comparat, J., Prada, F., Yepes, G., Burtin, E., Zarrouk, P., et al. (2017). Clustering of quasars in the first year of the SDSS-IV eBOSS survey: interpretation and halo occupation distribution. MNRAS, 468:728–740.

Ross, A. J., Beutler, F., Chuang, C.-H., Pellejero-Ibanez, M., Seo, H.-J., Vargas-Magaña, M., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: observational systematics and baryon acoustic oscillations in the correlation function. MNRAS, 464:1168–1191.

Salazar-Albornoz, S., Sánchez, A. G., Grieb, J. N., Crocce, M., Scoccimarro, R., Alam, S., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: angular clustering tomography and its cosmological implications. MNRAS, 468:2938–2956.

Sánchez, A. G., Grieb, J. N., Salazar-Albornoz, S., Alam, S., Beutler, F., Ross, A. J., et al. (2017a). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: combining correlated Gaussian posterior distributions. MNRAS, 464:1493–1501.

Sánchez, A. G., Scoccimarro, R., Crocce, M., Grieb, J. N., Salazar-Albornoz, S., Dalla Vecchia, C., et al. (2017b). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the configuration-space clustering wedges. MNRAS, 464:1640–1658.

SDSS Collaboration, Albareti, F. D., Allende Prieto, C., Almeida, A., Anders, F., Anderson, S., et al. (2016). The Thirteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-IV Survey MApping Nearby Galaxies at Apache Point Observatory. preprint, (arXiv:1608.02013).

Slepian, Z., Eisenstein, D. J., Beutler, F., Chuang, C.-H., Cuesta, A. J., Ge, J., et al. (2017). The large-scale three-point correlation function of the SDSS BOSS DR12 CMASS galaxies. MNRAS, 468:1070–1083.

Vargas-Magaña, M., Ho, S., Cuesta, A. J., O'Connell, R., Ross, A. J., Eisenstein, D. J., et al. (2016). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: theoretical systematics and Baryon Acoustic Oscillations in the galaxy correlation function. preprint, (arXiv:1610.03506).

Wang, Y., Zhao, G.-B., Chuang, C.-H., Ross, A. J., Percival, W. J., Gil-Marín, H., et al. (2016). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: tomographic BAO analysis of DR12 combined sample in configuration space. preprint, (arXiv:1607.03154).

Zhao, G.-B., Raveri, M., Pogosian, L., Wang, Y., Crittenden, R. G., Handley, W. J., et al. (2017a). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Examining the observational evidence for dynamical dark energy. preprint, (arXiv:1701.08165).

Zhao, G.-B., Wang, Y., Saito, S., Wang, D., Ross, A. J., Beutler, F., et al. (2017b). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: tomographic BAO analysis of DR12 combined sample in Fourier space. MNRAS, 466:762–779.

CHAPTER 2

Modelling Luminous Red Galaxies and Quasar samples

2.1 Main Results

The main goal of this thesis is to study the methods to be used to connect galaxies or quasars with dark matter halos from spectroscopic surveys and N-Body simulations respectively. We perform this work considering the clustering of galaxies by using the two-point correlation function as the more basic observable to be reproduced. Furthermore, we include some of the most important observational effects, which are due mainly to colour selections used in observations, as well as unavoidable uncertainties from the instrumentation.

The three papers presented in this work were produced within the SDSS-III and the SDSS-IV programs. They were companion papers in the BOSS Data Release 12 (DR12) and the eBOSS First Year of Quasars (Y1Q). Papers I and II are closely linked as both of them were part of the same project within the BOSS collaboration. They present a study of the clustering of luminous red galaxies. Paper III is a follow-up of the research started in BOSS which seeks to build high-fidelity mocks reproducing most of the features of the clustering of the observed samples. In it, we propose a modified halo abundance matching model which can be applied to other galaxy samples such as emission line galaxies (Favole et al., 2016).

The following sections present the most relevant results obtained in the three papers. Although all the results are part of the same line of research, they are divided in two blocks, the simulated LRG catalogues and the quasars mocks. We split the results because each block uses a different model to study the observed sample. In both cases, LRG and quasars, we use the BigMultiDark Planck N-Body simulation which has a box size large enough to cover

the volume of the surveys and a sufficiently accurate numerical resolution to resolve the dark matter halos that host the different galaxy populations. This simulation allows us to make predictions about different properties of the observed samples such as their bias or their halo occupation distribution. Additionally, our LRG model was one of the essential pillars in the construction of mocks for the covariance matrices of the BOSS Final Data Release (e.g. Zhao et al., 2017; Beutler et al., 2017; Ross et al., 2017).

2.1.1 Luminous Red Galaxies – BOSS

Nuza et al. (2013) produced one of the first CMASS catalogues of simulated galaxies from the $1h^{-1}$Gpc MultiDark WMAP7 simulation ($\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$). They proposed a simple model to select the typical dark matter halos that host LRGs. Their model mainly depends on two parameters which are fixed by the number density and the two-point correlation function of the observed sample. Their final mock catalogue uses the whole volume of the simulation reproducing with reasonable accuracy the correlation function and the power spectrum of the data. Paper I continues the work started by Nuza et al. (2013). Our main contribution is the inclusion of a large number of observational effects, which help to produce a galaxy catalogue closer to the observed distribution.

The first step to model these observations is to include the survey geometry in the simulated catalogue. This implies the construction of light-cones from different snapshots of the simulation. For this purpose I developed the SUrvey GenerAtoR code (SUGAR) which was designed to produce mock catalogues from simulations that reproduce some of the features of the observed data. In the case of CMASS, the goal is to construct light-cones which reproduce the dependence of the number density with redshift, as well as the geometrical mask that includes the angular completeness of the survey. Unlike Nuza et al. (2013), in our model halos hosting LRG are selected using a single parameter because the number density is fixed by construction. Then, we select dark matter halos using the peak circular velocity, V_{peak} (see Equation (1.15)). For galaxies, the stellar mass is used as a proxy, we specifically use the Portsmouth SED-fit DR12 stellar mass catalogue (Maraston et al., 2013). An important ingredient in the assignment of galaxies to halos is the stellar mass function (Figure 3, Chapter 3). It enable us to assign stellar masses from a complete sample to all halos in the simulation.

Unlike previous halo abundance matching studies, we do no construct volume limited samples

from the observed data. One of the improvements of this work is that we describe all the observed galaxies in the CMASS sample, which implies to model the observed stellar mass incompleteness at different redshifts. Moreover, we include the fiber collision effects, since the resolution of the BigMultiDark simulation allows us to study the small scales clustering of the sample. We use the methodology proposed by Guo et al. (2012) to mimic the impact of fiber collision in our catalogue. Using the distribution of plates in the telescope, this method assigns fibers to some galaxies and randomly removes the redshift of others. Then, we correct the clustering of the simulated catalogue using close-pair weights, just as in the case of the observational data. If a galaxy does not have spectroscopic redshift, the one from its nearest neighbour is assigned. This process introduces an unphysical displacement of the galaxy in the radial coordinate. This effect can be used to compare our fiber collisions model with observations. We can compare the displacements of the galaxies in the simulation with those from the data as is shown in Hahn et al. (2017). Figure 7 of Chapter 3 shows an excellent agreement between our catalogues and observations.

Current observations do not provide a direct measurement of the scatter between dark matter halos and galaxies, so its value has to be fixed by indirect measurements. In this work, we use the impact of the scatter on the clustering of galaxies. We use the value that better reproduces the monopole of the two-point correlation function between $2h^{-1}$Mpc and $30h^{-1}$Mpc. This range is chosen in order to avoid possible disagreements due to fiber collisions at small scales or remaining systematics at large scales. Additionally, the effect of cosmic variance is not too large at these scales. Once the value of the scatter is fixed, the model can reproduce almost all scales of the monopole within 1σ errors. The largest differences are mainly due to remaining systematics in the data.

Our catalogue reproduces the projected correlation function at all scales within 1σ, just as the monopole of the 2PCF. However, we do not find the same agreement for the quadrupole of the correlation function. In this case, we find differences larger than 1σ at ~ 20 h^{-1}Mpc. The disagreement in the quadrupole and also a small deviation in the three point correlation function for angles ~ 0 and $\sim \pi$ could be related to the same physical processes. However, for all the other scales we find that our catalogue reproduces with good agreement the three point correlation function for different configurations.

We also find an excellent agreement with the monopole of the power spectrum from observation. We are able to reproduce scales at $k \sim 1$. Similarly, the prediction of the baryonic peaks are in excellent agreement with the observed data. Using the dark matter particles

from the simulation and our simulated galaxy catalogue, we compute the bias of the CMASS sample which is in agreement with previous results (e.g. White et al., 2011; Nuza et al., 2013). Additionally, we show our prediction of the halo mass to stellar mass relation that can give us information about the formation of galaxies in dark matter halos. This prediction is in good agreement with lensing measurements.

The model presented in Paper I constitutes the basis of Paper II, which describes the pipeline used in the construction of 12,288 mock galaxy catalogues for the Final Data Release of BOSS. For this analysis, we also include the LOWZ sample. These catalogues were designed on the purpose of having a better estimation of the covariance matrices, providing strong constraints on the measurements of the cosmological parameters. As in the case of Paper I, the basic idea is to produce catalogues that reproduce observation with a good precision (high fidelity mocks). For this reason, we include observational effects such as fiber collision, geometry of the survey, the radial selection function for each sample and the evolution of the clustering including 10 snapshots in the redshift range $0.15 < z < 0.75$. These catalogues allowed a robust analysis of BAOs and RSDs done in another studies (e.g. Grieb et al., 2017; Gil-Marín et al., 2017; Ross et al., 2017; Beutler et al., 2017).

The mock production is divided in different stages which are performed by three principal codes. Boxes at different redshifts are produced by the PATCHY CODE (Kitaura et al., 2014) which uses an Augmented Lagrangian Perturbation Theory (ALPT) to evolve the density field of dark matter. This method allows us to run simulations with a low computational cost compared with the N-body simulation. For this project, the PATCHY CODE provided the catalogue of galaxies directly by modelling the bias of the sample. In order to fix the bias, the PATCHY CODE needs galaxy catalogues from an N-Body simulation to calibrate the phase parameters of the bias model. Thus, we follow a similar procedure as exposed in Paper I. In this case we do not generate light-cones from the simulation, but ten boxes at different redshift reproducing the number density and the clustering of the observational data. Unfortunately, the PATCHY CODE only provides positions and velocities of halos. In order to include stellar masses for each mock a proxy for halo masses is needed. So the HAlo Distribution ReconstructiON code (HADRON Zhao et al., 2015) is used, assigning masses to the various objects.

In the last step, we construct a pipeline based on the SUGAR CODE in order to manage thousands of simulations with different snapshots and join them for the final version of the mocks. In this part, we construct light-cones and include all the observational effects such as

the geometry, the fiber collisions or the stellar mass of each galaxy. Finally, we provide 2048 mocks for each LOWZ, CMASS, and combined LOWZ+CMASS and northern and southern galactic cap. So far, these catalogues constitute the largest ever simulated volume ($\sim 192,000$ h^{-1}Gpc) and they are publicly available[5].

Our catalogues reproduce the two- and three-order clustering statistics with a good agreement, as well as the dependence of the clustering with the stellar mass, making them the most realistic option compared to the other available mock catalogues. In Paper II we show that MD-PATCHY BOSS DR12 mocks are in agreement within 1σ for the monopole, quadrupole and hexapole of the two point correlation function and power spectrum. Just as in the case of the BigMultiDark Planck simulation, the resulting mocks reproduce the three-point function in redshift and k space within 1σ for most of scales.

2.1.2 Quasar – eBOSS

For the last few years, the clustering of the LRG sample has been carefully studied with different methods (Zheng et al., 2007; White et al., 2011; Guo et al., 2014; Leauthaud et al., 2016; Montero-Dorta et al., 2016; Tinker et al., 2017). LRGs are the most massive galaxies, thus they fill the high-mass end of the stellar mass function. This feature simplifies their connection with dark matter halos using luminosity or stellar mass. In the case of CMASS, this connection is even cleaner than in other surveys, because the colour cuts were performed to prefer the selection of LRGs. However, ELG and quasars sample are more complex to model. Their halo masses are not well constrained and measuring the incompleteness of each sample is not easy.

Just as in Paper I, the goal of Paper III is to design a simple model which reproduces the observed sample and extracts information about the halos hosting quasars. The scatter used in the standard HAM and the stellar mass incompleteness of the LRG affect the distribution of halos hosting galaxies in a similar way. If we knew the final distribution of halos, we could select them from the simulation following this distribution without using the scatter and the incompleteness. This statement seems obvious but can be helpful for quasars. In that case, we do not know neither the intrinsic scatter between dark matter halos and quasars nor the incompleteness of the sample. Then, we assume that the final distribution of the halos hosting quasars is given by a Gaussian function that will depend on three parameters, the

[5] http://skyserver.sdss.org/

mean, the standard deviation and the fraction of quasars living in subhalos. Just as in the case of LRG, these parameters are fixed by the observed clustering.

Using this model, we construct simulated quasar catalogues that reproduce the two-point correlation function of ∼70,000 optical quasars from the eBOSS Y1Q CORE sample in the redshift range $0.9 < z < 2.2$. Because of the large redshift range, the model is implemented in light-cones constructed from the BigMultiDark Planck simulation, covering a comparable area to the eBOSS Y1Q sample. In the case of quasars, we use V_{max} as halo proxy. Current observations do not bear information on small-scale clustering. For this reason, we cannot constrain the fraction of satellites and we do not distinguish between host and subhalos when the selection is done. The final mock thus has the same fraction of satellites as the complete simulation in the mass range used.

The current data do not allow us to impose constraints on the width of the distribution, so we used a single parameter in order to model the clustering of the Y1Q. The width of the Gaussian distribution is fixed to 30 $\mathrm{km}s^{-1}$ and we only impose a value to the satellite fraction in the BigMDPL-QSO-NSAT light-cone while for the other light-cones we do not fix this parameter. We produce three kinds of light-cones, one including the evolution of the parameters with redshift (BigMDPL-QSOZ), another describing the whole redshift range with a single parameter (BigMDPL-QSO) and a third one fixing the satellite fraction to zero (BigMDPL-QSO-NSAT), where the mean halo masses are $10^{12.61}$, $10^{12.66}$ and $10^{12.70}$ $M_{\odot}$, respectively.

The prediction of our model is in a good agreement with the 2PCF and the monopole of the power spectrum of the Y1Q data. In our catalogue we include redshift errors given by Dawson et al. (2016). These errors improve the agreement between our model and the data. They provide a good description of the observed clustering on small scales, which is very sensitive to variations caused by these errors.

In order to compare all the light-cones, we use the Bayes factor, finding a strong evidence that the BigMDPL-QSOZ (four parameters) reproduces the data better than the BigMDPL-QSO (one parameter). However, we cannot make the same conclusion with the model without satellites, which reproduces the data with a similar agreement than the BigMDPL-QSOZ model. Finally, we describe the relation between dark matter halos and quasar computing the linear and second order bias. The mean linear bias of the Y1Q sample gives us 2.37 ± 0.12 and a second-order bias $b_2 = 0.314 \pm 0.030$.

2.2 Discussion

Luminous red galaxies have been studied in a wide redshift range. SDSS has provided a large number of galaxies which have been studied with many different methods in order to understand the more massive galaxies in the Universe. HOD and HAM methods are widely used to analyse the clustering of the LRGs. Other studies include more elaborated models taking into account effects such as the halo assembly bias (Saito et al., 2016). Our study shows that a simple HAM model can have a good agreement with the current data. However, future experiments will increase the accuracy of measurements and it will be possible to distinguish between more complex models.

The galaxy catalogue produced in this work combines a large number of observational features with a very accurate simulation built with the most precise measurements of the cosmological parameters. Our work improves the results found by Nuza et al. (2013) due mainly to the cosmological parameters used to run the simulation and the larger volume of the survey. The observational effects included in our catalogues ensure that the mean scatter between dark matter and galaxies has to be closer to true scatter in the Universe, since we remove the contribution of the incompleteness. Using a light-cone we also include the evolution of the dark matter field and the dependence of the number density with redshift. However, we model the incompleteness of the sample using a simple downsampling which can introduce small systematics effects in our catalogue, because not all the galaxies in that range of stellar masses are LRGs. Furthermore, the scatter value and the incompleteness of the sample have a strong dependence with the model used to compute the stellar mass of each galaxy, so each stellar mass model will provide differences in the clustering of the observed data, that is to say different values of the scatter.

Our HAM model for LRG reproduces most of the observational measurements within 1σ (Figure 9 and Figure 11, Chapter 3). However, some discrepancies are found at large scales in the monopole of the correlation function. It is possible to see that all the data points larger than a 100 h^{-1}Mpc are systematically boosted compared to the theory. This could be explained by remaining systematics that should be included in the observational data. A more important disagreement is found in the quadrupole of the correlation function at scales $\sim$20 h^{-1}Mpc, as well as in the three point correlation function for angles between ~ 0 and $\sim \pi$. We enumerate some of the possible causes of this disagreement. First, it is important to notice that we do some approximations in the mock building process such as the non-

evolution of the stellar mass function or the random selection of galaxies in order to account for incompleteness of the observational sample. It is also relevant to comment that our model works fine in the most complete range of the CMASS sample ($0.51 < z < 0.57$), showing that incompleteness effects could play a role in this discrepancy. From a theoretical point of view, this disagreement could be related to the velocity dispersion of the simulated galaxies inside the halo. We are assuming that galaxies are at the centre of the host halos and have their same velocities. Guo et al. (2016) solve this problem including a velocity bias for satellite galaxies. Another possible explanation is the effects due to the assembly bias of the halos as shown in Saito et al. (2016). Despite the differences in the quadrupole, our model reproduces observations with high precision. These excellent predictions are the result of combining the observational effects of the survey and a simulation with the best cosmological parameters measured from the CMB. This proves that the ΛCDM that best fit the CMB anisotropies at $z = 1100$ by Planck Collaboration et al. (2014) predicts with high accuracy the clustering of the LRG at $z \sim 0.5$.

The PATCHY CODE has been a successful method reproducing N-Body simulation at scales larger than few Mpc (Kitaura et al., 2014; Chuang et al., 2015b). Then, our model for LRG and the potential of the PATCHY CODE enable us to construct thousand of LRG mocks using few computational resources. Just as in the case of the BigMultiDark, quadrupole disagreements are also present in the PATCHY simulations. However, mock catalogues have to describe the observational results in the most accurate way. So, the discrepancy on the quadrupole was corrected increasing the Gaussian noise in the velocity distribution of the simulated galaxies. We obtained results with an excellent agreement for scales ~ 10 Mpc in real space and ~ 0.3 Mpc^{-1}h in Fourier Space. The excellent description of the observational data made by the MD-PATCHY DR12 mocks was very important to construct realistic covariance matrices in different studies of BOSS DR11/DR12. Our mocks have shown a better match to the data than the second set of mocks available for BOSS, Quick Particle Mesh mocks (QPM; White et al., 2014), in terms of two and three point statistics.

In the new era of precision cosmology, it is fundamental to improve the results of our numerical methods in order to have a more detailed description of observations. The quality of the simulations can have a direct impact on the constraints of the measurements and thus, in the future it will play a decisive role when testing different cosmological models. Over the next decades, surveys will provide a huge quantity of information and their analyses will be a challenge for numerical simulation, particularly for the different methods used to construct

the covariance matrices. New experiments will observe very large volumes in the sky, focusing on the study of ELG and quasars that live in halos with a wide mass range. BAO studies can still use the current methods to generate mock catalogues, because the linear regime can be well described by them. However, RSD analyses study the non-linear scales which current mock simulated catalogues cannot properly resolve. Of course, this happens if the simulated volume is very large and the computational resources are limited. In the future, we will have to solve these problems, which are already presented in large surveys as eBOSS, and they will be more critical for the next generation surveys like DESI or EUCLID.

Both emission line galaxies and quasars have been the principal targets of current and future observational projects. Different studies have shown that quasars are good biased tracers of the dark matter density field and due to their observational features they allow us to explore deeper regions in the Universe. Recent studies of the Lyman-α forecast of the quasars have provided measurements of the BAO scale at $z > 2$. In the coming years, eBOSS collaboration expects to measure the BAO peak using spectroscopically confirmed quasars in the range $0.9 < z < 2.2$, for the first time. These new results will provide important information about the formation and evolution of the structures and therefore of the formation of galaxies. All these studies require a better understanding of the distribution of the ELG and quasars in relation to the underlying dark matter field.

One of the biggest challenges of cosmological simulations is to increase the current volumes, making them comparable to the new surveys, and, at the same time, increase the resolution in order to resolve the halos hosting quasars or ELG. Different studies find that the typical mass of halos hosting quasars is $\sim 10^{12.5}\ M_{\odot}$. For the eBOSS first year of quasars, the BigMultiDark simulation provides a volume comparable to observations and it resolves the dark matter halos needed for this study. However, we are close to the resolution limit of the simulation and this does not allow us to increase the parameter space of our model.

It is not very common to find pairs of quasars living in the same halo host and this is also due to the small number density of those objects. Additionally, fiber collisions can also have some impact on small scales. So it is very difficult to understand the distribution of quasars inside dark matter halos and the fraction of quasars living in subhalos is an open question. Similar problems are found in the analysis of emission line galaxies. In the future, the cross-correlation between different populations could help understand what the distribution of these objects at small scales is. In the last year of eBOSS, we will have regions in the sky where LRG and ELG are overlapped, making it possible to measure the cross-correlation of

both galaxy samples. The simple methods for modelling the clustering of quasars or ELG can be a very useful tool in the construction of mock catalogues, as in the case of LRG. Mock catalogues as GLAMmocks (Klypin et al., 2016) o EZmocks (Chuang et al., 2015a) are using our models for fixing the bias of the eBOSS sample.

CHAPTER 3

Clustering of LRG in the BOSS-DR12

Publication: Monthly Notices of the Royal Astronomical Society, Volume 460, Issue 2, p.1173-1187

Motivation

The SDSS I/II programs provided a large amount of data which allowed us to measure the BAO scale in the local universe for the first time. These projects observed a large number of the most massive galaxies in the Universe. Many studies of this sample have increased our knowledge of these galaxies. The next generation of the SDSS project, BOSS increased the number and the volume of the sample at $z \sim 0.5$. It allowed us to have more precise measurements. This way, we need more accurate models to study the distribution of LRG at different redshift. So we propose to combine one of the most successful models, an N-body simulation (with the best prediction of the cosmological parameters) and observational effects from the survey in order to have the most realistic description of the galaxy clustering of the survey.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **460,** 1173–1187 (2016) doi:10.1093/mnras/stw1014
Advance Access publication 2016 April 29

The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: modelling the clustering and halo occupation distribution of BOSS CMASS galaxies in the Final Data Release

Sergio A. Rodríguez-Torres,[1,2,3★†] Chia-Hsun Chuang,[1,4‡] Francisco Prada,[1,2,5,6] Hong Guo,[7,8] Anatoly Klypin,[9,10] Peter Behroozi,[11] Chang Hoon Hahn,[12] Johan Comparat,[1,3] Gustavo Yepes,[3] Antonio D. Montero-Dorta,[8] Joel R. Brownstein,[8] Claudia Maraston,[13] Cameron K. McBride,[14] Jeremy Tinker,[12] Stefan Gottlöber,[4] Ginevra Favole,[1,2] Yiping Shu,[8] Francisco-Shu Kitaura,[4] Adam Bolton,[8] Román Scoccimarro,[12] Lado Samushia,[13,15,16] David Schlegel,[5] Donald P. Schneider[17,18] and Daniel Thomas[13]

Affiliations are listed at the end of the paper

Accepted 2016 April 27. Received 2016 April 27; in original form 2015 September 29

ABSTRACT

We present a study of the clustering and halo occupation distribution of Baryon Oscillation Spectroscopic Survey (BOSS) CMASS galaxies in the redshift range $0.43 < z < 0.7$ drawn from the Final SDSS-III Data Release. We compare the BOSS results with the predictions of a halo abundance matching (HAM) clustering model that assigns galaxies to dark matter haloes selected from the large BigMultiDark *N*-body simulation of a flat Λ cold dark matter Planck cosmology. We compare the observational data with the simulated ones on a light cone constructed from 20 subsequent outputs of the simulation. Observational effects such as incompleteness, geometry, veto masks and fibre collisions are included in the model, which reproduces within 1σ errors the observed monopole of the two-point correlation function at all relevant scales: from the smallest scales, $0.5\,h^{-1}$ Mpc, up to scales beyond the baryon acoustic oscillation feature. This model also agrees remarkably well with the BOSS galaxy power spectrum (up to $k \sim 1\ h\ \mathrm{Mpc}^{-1}$), and the three-point correlation function. The quadrupole of the correlation function presents some tensions with observations. We discuss possible causes that can explain this disagreement, including target selection effects. Overall, the standard HAM model describes remarkably well the clustering statistics of the CMASS sample. We compare the stellar-to-halo mass relation for the CMASS sample measured using weak lensing in the Canada–France–Hawaii Telescope Stripe 82 Survey with the prediction of our clustering model, and find a good agreement within 1σ. The BigMD-BOSS light cone including properties of BOSS galaxies and halo properties is made publicly available.

Key words: methods: numerical – galaxies: abundances – galaxies: haloes – large-scale structure of Universe.

1 INTRODUCTION

One of the major goals in cosmology is to explain the formation of the large-scale structure (LSS) of the Universe. However, the main ingredient that drives this evolution – the dark matter – can only be probed using the distribution of galaxies, and galaxies are biased tracers of the matter field. This makes this study challenging. In the last 20 years, vast amounts of observational data have been obtained, improving each time the precision of the LSS measurements and demanding ever more accurate theoretical models. In fact, one of the strongest arguments that we understand how the LSS forms and evolves is our ability to reproduce the galaxy clustering through cosmic time, starting from the primordial Gaussian perturbations.

★E-mail: sergio.rodriguez@uam.es
† Campus de Excelencia Internacional UAM/CSIC Scholar.
‡ MultiDark Fellow.

During the last decade, surveys such as the Sloan Digital Sky Survey (SDSS-I/II/III; York et al. 2000; Eisenstein et al. 2011) have made it possible to determine the clustering of galaxy populations at scales out to tens of Mpc and beyond with reasonable accuracy.

The Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013) Data Release 12 (DR12; Alam et al. 2015) provides redshift of 1.5 million massive galaxies in 10 000 deg^2 area of the sky and for redshifts in the range 0.15–0.75. BOSS DR12 has an effective volume seven times larger than that of the SDSS-I/II project. These data provide us with a sufficiently statistical sample to examine our theoretical predictions over a range of scales.

In order to compare the Λ cold dark matter (ΛCDM) model and the observational data, it is necessary to link the galaxy and the dark matter distributions. There are a number of methods to assign galaxies to the dark matter. State-of-the-art hydrodynamical simulations, which include detailed galaxy formation descriptions, are computationally unaffordable for the volumes considered here (e.g. Vogelsberger et al. 2014; Schaye et al. 2015), and indeed, there are no large samples of simulated galaxies that can be used to match BOSS. Semi-analytic models are less computationally consuming methods to populate dark matter haloes with galaxies (e.g. Knebe et al. 2015). These models incorporate some physics of galaxy formation.

The most popular models are based on the statistical relations between galaxies and dark matter haloes. One of the most used models is the halo occupation distribution (HOD; e.g. Jing, Mo & Börner 1998; Peacock & Smith 2000; Berlind & Weinberg 2002; Zheng et al. 2005; Leauthaud et al. 2012; Guo et al. 2014). The main component of the HOD is the probability, $P(N|M_{\rm halo})$, that a halo of virial mass $M_{\rm halo}$ hosts N galaxies with some specified properties. These models have several parameters which allow one to match the observed clustering.

The model known as the halo abundance matching (HAM; Kravtsov et al. 2004; Conroy, Wechsler & Kravtsov 2006; Behroozi, Conroy & Wechsler 2010; Guo et al. 2010; Trujillo-Gomez et al. 2011; Nuza et al. 2013; Reddick et al. 2013) connects observed galaxies to simulated dark matter haloes and subhaloes by requiring a correspondence between the luminosity or stellar mass and a halo property. The assumption of this model is that more luminous (massive) galaxies are hosted by more massive haloes. However, this relation is not a one-to-one relation because there is a physically motivated scatter between galaxies and dark matter haloes (e.g. Shu et al. 2012). By construction, the method reproduces the observed luminosity function, LF (or stellar mass function, SMF). HAM relates the LF (SMF) of an observed sample with the distribution of haloes in an N-body simulation. The implemented assignment requires that one works with complete samples in luminosity (stellar mass) or have precise knowledge of the incompleteness as a function of the luminosity (stellar mass) of the galaxy sample. Luminous red galaxies (LRGs) are the most massive galaxies in the Universe, and they represent the high-mass end of the SMF. This feature makes this population of galaxies an excellent group to be reproduced with the abundance matching.

In this paper, we compare the clustering of the BOSS CMASS DR12 sample with predictions from N-body simulations. We use an abundance matching to populate the dark matter haloes of the BigMultiDark Planck simulation (BigMDPL; Klypin et al. 2016). In order to include systematic effects from the survey, as well as the proper evolution of the clustering, we construct light cones which reproduce the angular selection function, the radial selection function and the clustering of the monopole in configuration space. To generate these catalogues, we developed the SUrvey GenerAtoR (SUGAR) code. Once the HAM and the light cone are applied, we compute the predictions of our model for two-point statistics and the three-point correlation function (3PCF). We also present the prediction of the stellar-to-halo mass relation and its intrinsic scatter compared to lensing measurements. The HAM, the BigMDPL and the methodology to produce light cone played a key role in the construction of the MultiDark PATCHY BOSS DR12 mocks (MD-PATCHY mocks; Kitaura et al. 2016, companion paper).

In order to have a good estimation of the uncertainties in this work, we use 100 MD-PATCHY mocks. These mocks are produced using five boxes at different redshifts that are created with the PATCHY code (Kitaura, Yepes & Prada 2014). The PATCHY code can be decomposed into two parts: (1) computing approximate dark matter density field and (2) populating galaxies from dark matter density field with the biasing model. The dark matter density field is estimated using augmented Lagrangian perturbation theory (Kitaura & Heß 2013) which combines the second-order perturbation theory (see e.g. Buchert 1994; Bouchet et al. 1995; Catelan 1995) and spherical collapse approximation (see Bernardeau 1994; Mohayaee et al. 2006; Neyrinck 2013). The biasing model includes deterministic bias and stochastic bias (for details see Kitaura et al. 2014). The velocity field is constructed based on the displacement field of dark matter particles. The modelling of finger-of-god has also been taken into account statistically. The MD-PATCHY mocks are constructed based on the BigMD simulation with the same cosmology used in this work. The mocks match the clustering of the galaxy catalogues for each redshift bin (see Kitaura et al. 2016, companion paper, for details). The BigMultiDark light-cone catalogues of BOSS CMASS galaxies in the Final DR12 (hereafter BigMD-BOSS light cone) presented in this work are publicly available.

This paper is structured as follows: Sections 2 and 3 describe the SDSS-III/BOSS CMASS galaxy sample and the BigMDPL N-body cosmological simulations used in this work. In Section 4, we provide details on different observational effects and briefly describe the SUGAR code. Section 4.1 presents the main ingredients of the HAM modelling of the CMASS galaxy clustering. A comparison of our results to observation is shown in Section 5. Subsequently, we discuss the principal results in Section 6. Finally, in Section 7, we present a summary of our work. For all results in this work, we use the cosmological parameters $\Omega_{\rm m} = 0.307$, $\Omega_{\rm B} = 0.048$, $\Omega_\Lambda = 0.693$.

2 SDSS-III/BOSS CMASS SAMPLE

The Baryon Oscillation Spectroscopic Survey[1] (BOSS; Bolton et al. 2012; Dawson et al. 2013) is part of the SDSS-III programme (Eisenstein et al. 2011). The project used the 2.5 m aperture Sloan Foundation Telescope at Apache Point Observatory (Gunn et al. 2006). The telescope used a drift-scanning mosaic CCD camera (Gunn et al. 1998) with five colour bands, u, g, r, i, z (Fukugita et al. 1996). Spectra are obtained using the double-armed BOSS spectrographs, which are significantly upgraded from those used by SDSS I/II, covering the wavelength range 3600–10 000 Å with a resolving power of 1500–2600 (Smee et al. 2013). BOSS provides redshift for 1.5 million galaxies in 10 000 deg^2 divided into two samples: LOWZ and CMASS. The LOWZ galaxies are selected to be the brightest and reddest of the low-redshift galaxy population ($z \lesssim 0.4$), extending the SDSS I/II LRGs. The CMASS target

[1] http://skyserver.sdss.org/dr12/en/home.aspx

selection is designed to isolate galaxies at higher redshift ($z \gtrsim 0.4$), most of them being also LRGs.

In the present paper, we focus on the CMASS DR12 North Galactic Cap (NGC) sample. Galaxies are selected from SDSS DR8 imaging (Aihara et al. 2011) according to a series of colour cuts designed to obtain a sample with approximately 'constant stellar mass' (Reid et al. 2016). The following photometric cuts are applied:

$$17.5 < i_{\rm cmod} < 19.9 \quad (1)$$

$$r_{\rm mod} - i_{\rm mod} < 2 \quad (2)$$

$$d_{\perp} > 0.55 \quad (3)$$

$$i_{\rm fib2} < 21.5 \quad (4)$$

$$i_{\rm cmod} < 19.86 + 1.6(d_{\perp} - 0.8), \quad (5)$$

where i and r indicate magnitudes, and $i_{\rm fib2}$ is the i-band magnitude within a 2 arcsec aperture. All magnitudes are corrected for Galactic extinction [via the Schlegel et al. (1998) dust maps]. The subscript 'mod' denotes the 'model' magnitudes and the subscript 'cmod' refers to the 'cmodel' magnitudes. The model magnitudes represent the best fit of the DeVaucouleurs and exponential profile in the r band (Stoughton et al. 2002) and the cmodel magnitudes denote the best-fitting linear combination of the exponential and DeVaucouleurs models (Abazajian et al. 2004). $d_{\perp}$ is defined as

$$d_{\perp} = r_{\rm mod} - i_{\rm mod} - (g_{\rm mod} - r_{\rm mod})/8.0. \quad (6)$$

Star–galaxy separation is performed on the CMASS targets via

$$i_{\rm psf} - i_{\rm mod} > 0.2 + 0.2(20.0 - i_{\rm mod}) \quad (7)$$

$$z_{\rm psf} - z_{\rm mod} > 9.125 - 0.46 z_{\rm mod}. \quad (8)$$

The subscript 'psf' refers to point spread function magnitudes. CMASS sample contains galaxies with redshift $z > 0.4$, having the peak of the number density at $z \approx 0.5$. We will concentrate our analysis in the redshift range $0.43 < z < 0.7$ for this sample.

BOSS sample is corrected for redshift failures and fibre collisions. In the following sections, we will use the same weights given in Anderson et al. (2014) in order to correct the clustering signal affected by these systematics (Ross et al. 2012). The total weight for a galaxy is given by

$$w_{\rm g} = w_{\rm star} w_{\rm see} (w_{\rm zf} + w_{\rm cp} - 1). \quad (9)$$

In this equation, $w_{\rm zf}$ denotes the redshift failure weight and $w_{\rm cp}$ represents the close pair weight. Both quantities start with unit weight. If a galaxy has a nearest neighbour (of the same target class) with a redshift failure ($w_{\rm zf}$) or its redshift was not obtained because it was in a close pair ($w_{\rm cp}$), we increase $w_{\rm zf}$ or $w_{\rm cp}$ by one. As found in Ross et al. (2012), the impact of this effect is very small for the CMASS sample; for this reason, we do not model the redshift failures in this study. For CMASS, additional weights are applied to account for the observed systematic relationships between the number density of observed galaxies and stellar density and seeing (weights $w_{\rm star}$ and $w_{\rm see}$, respectively).

3 BIGMULTIDARK SIMULATION

The BigMDPL is one of the MultiDark[2] N-body simulation described in Klypin et al. (2016). The BigMDPL was performed with GADGET-2 code (Springel 2005). This simulation was created in a box of 2.5 h^{-1} Gpc on a side, with 3840^3 dark matter particles. The mass resolution is $2.4 \times 10^{10}\ h^{-1}\ {\rm M}_{\odot}$. The initial conditions, based on initial Gaussian fluctuations, are generated with Zeldovich approximation at $z_{\rm init} = 100$. The suite of BigMultiDark is constituted of four simulations with different sets of cosmological parameters. In this study, we adopt a flat ΛCDM model with the Planck cosmological parameters: $\Omega_{\rm m} = 0.307$, $\Omega_{\rm B} = 0.048$, $\Omega_{\Lambda} = 0.693$, $\sigma_8 = 0.829$, $n_{\rm s} = 0.96$ and a dimensionless Hubble parameter $h = 0.678$ (Klypin et al. 2016). The simulation provides 20 redshift outputs (snapshots) within the redshift range $0.43 < z < 0.7$.

For the present analysis, we use the ROCKSTAR (Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement) halo finder (Behroozi, Wechsler & Wu 2013a). Spherical dark matter haloes and subhaloes are identified using an approach based on adaptive hierarchical refinement of friends-of-friends groups in six phase-space dimensions and one time dimension. ROCKSTAR computes halo mass using spherical overdensities of a virial structure. Before calculating halo masses and circular velocities, the halo finder performs a procedure which removes unbound particles from the final mass of the halo. ROCKSTAR creates particle-based merger trees. The merger trees algorithm (Behroozi et al. 2013b) was used to estimate the peak circular velocity over the history of the halo, $V_{\rm peak}$, which we use to perform the abundance matching.

4 METHODOLOGY: THE SUGAR CODE

We construct light-cone catalogues from the BigMDPL simulation which reproduce the clustering measured in the monopole of the redshift-space correlation function from the BOSS CMASS DR12 sample. For this purpose, we developed the SUGAR code which implements the HAM technique to generate galaxy catalogues from a dark matter simulation. The code can apply the geometric features of the survey and selection effects, including stellar mass incompleteness and fibre collision effects. All the available outputs (snapshots) of the BigMDPL simulation are used, so that the light cone has the proper evolution of the clustering.

In the following subsections, we present the ingredients used to produce the BigMD-BOSS light cone, which is shown in Figs 1 and 2. We present the HAM method and the SMF adopted in this work. The light-cone production, the fibre collision assignment and the modelling of the stellar mass incompleteness are also shown.

4.1 HAM procedure

We use a HAM technique to populate dark matter haloes with galaxies (see e.g. Nuza et al. 2013). This physically motivated method produces mock galaxy catalogues that in the past gave good representations of large galaxy samples (see for SDSS, e.g. Trujillo-Gomez et al. 2011; Reddick et al. 2013). The basic assumption of this method is that massive haloes host massive galaxies. This allows one to generate a rank-ordered relation between dark matter haloes and galaxies. However, observations show that this assignment cannot be a one-to-one relation (Shu et al. 2012). In order to create a more realistic approach, it is necessary to include scatter in

[2] http://www.multidark.org/

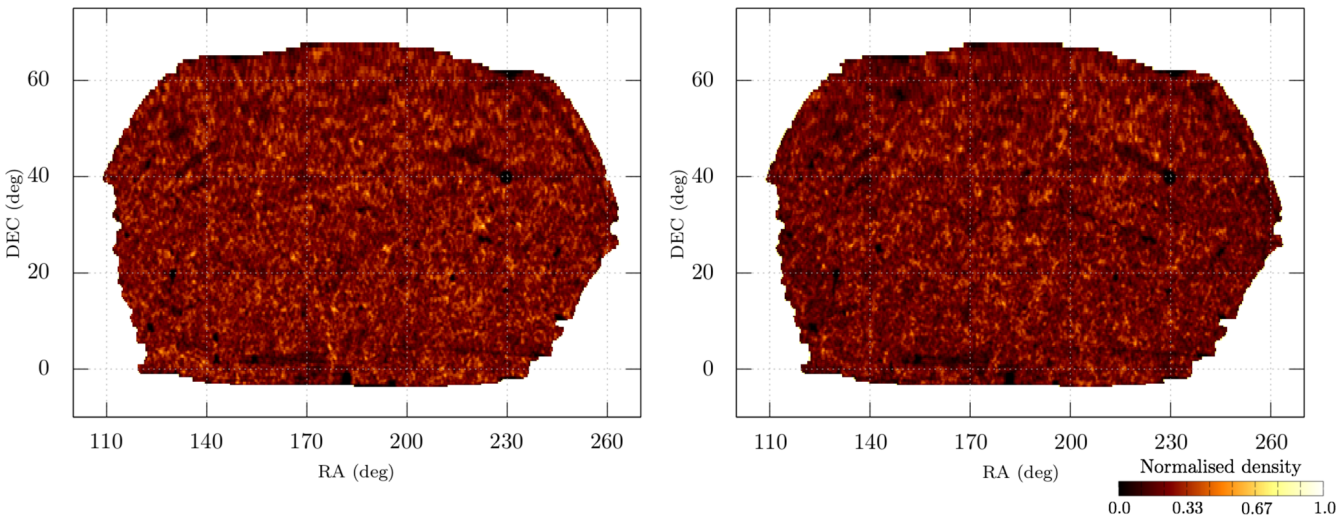


Figure 1. Left-hand panel: sky area covered by the BigMD-BOSS light cone. This region includes the BOSS CMASS DR12 geometry and veto masks. Right-hand panel: sky area covered by the BOSS CMASS DR12 sample. Colours indicate the angular number density, which is normalized by the most dense pixel. Each pixel has an angular area of 1 deg^2. BigMD-BOSS light cone uses the same mask as the BOSS CMASS DR12, including angular completeness and veto masks.

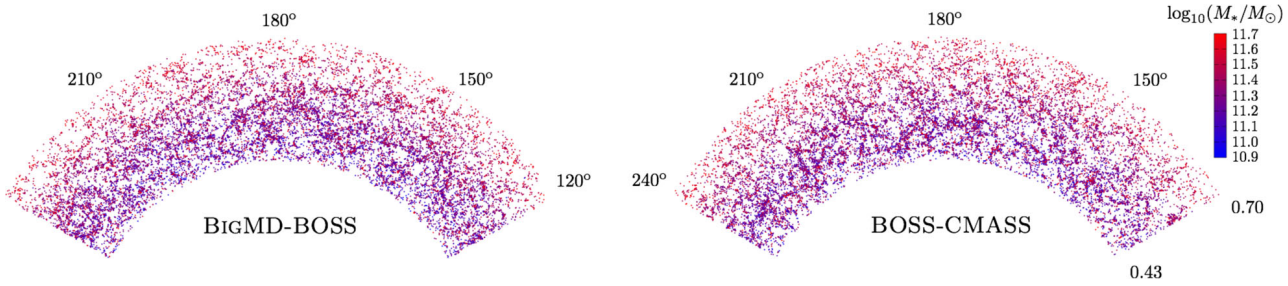


Figure 2. Pie plot of the BigMD-BOSS light cone (left-hand panel) and the BOSS CMASS DR12 data (right-hand panel). Both figures were made with 2 deg of thickness (Dec. coordinate).

this matching. The HAM can relate galaxy luminosities or stellar mass from galaxies to a halo property. In this paper, we use the peak value of the circular velocity over the history of the halo ($V_{\rm peak}$), which has advantages compared to the halo mass ($M_{\rm halo}$). $M_{\rm halo}$ is well defined for host haloes, but its definition becomes ambiguous for subhaloes. The subhalo mass also depends on the halo finder used (Trujillo-Gomez et al. 2011; Reddick et al. 2013). In addition to $M_{\rm halo}$ and $V_{\rm peak}$, HAM can be performed using other quantities such as the maximum circular velocity of the halo ($V_{\rm max}$), the maximum circular velocity of the halo at time of accretion ($V_{\rm acc}$) or the halo mass at time of accretion ($M_{\rm acc}$). Other studies present the effect of the halo property in the HAM (e.g. Reddick et al. 2013; Guo et al. 2015a).

We adopt a modified version of the scatter proposed in Nuza et al. (2013). Our implementation of the abundance matching can be briefly summarized in the following steps.

(i) For the dark matter haloes, we define a scattered $V_{\rm peak}$, which is used only to assign stellar mass to the haloes. This scattered quantity is defined by

$$V_{\rm peak}^{\rm scat} = (1 + \mathcal{N}(0, \sigma_{\rm HAM}))V_{\rm peak}, \tag{10}$$

where $\mathcal{N}$ is a random number, produced from a Gaussian distribution with mean 0 and standard deviation $\sigma_{\rm HAM}(V_{\rm peak}|M_*)$.

(ii) Sort the catalogue by $V_{\rm peak}^{\rm scat}$, starting from the object with the largest velocity and continuing down until reaching all the available objects. Use this catalogue to construct the cumulative number density of the haloes as a function of $V_{\rm peak}^{\rm scat}$.

(iii) Compute the cumulative number density of galaxies as a function of the stellar mass using the adopted SMF (see Section 4.2).

(iv) Finally, construct a monotonic relation between the cumulative number density functions from steps (ii) and (iii) such as

$$n_{\rm gal}(> M_*^i) = n_{\rm halo}(> V_{{\rm peak},i}^{\rm scat}). \tag{11}$$

This relation implies that a halo with $V_{{\rm peak},i}^{\rm scat}$ will contain a galaxy with stellar mass M_*^i.

This assignment is monotonic between $V_{\rm peak}^{\rm scat}$ and M_*, but not between $V_{\rm peak}$ and M_*. The relation of these two quantities is mediated by the scatter parameter, $\sigma_{\rm HAM}(V_{\rm peak}|M_*)$.

4.2 Stellar mass function

We employ the Portsmouth SED-fit DR12 stellar mass catalogue (Maraston et al. 2013) with the Kroupa initial mass function (Kroupa 2001) to estimate the SMF. The CMASS LSS catalogue does not include the stellar mass information. For that reason, we matched the BOSS and the LEGACY stellar mass catalogues with the LSS BOSS CMASS catalogue. In order to identify an SDSS spectrum in the different catalogues, there are three numbers that determine each galaxy: PLATE, MJD and FIBERID. We use these three quantities

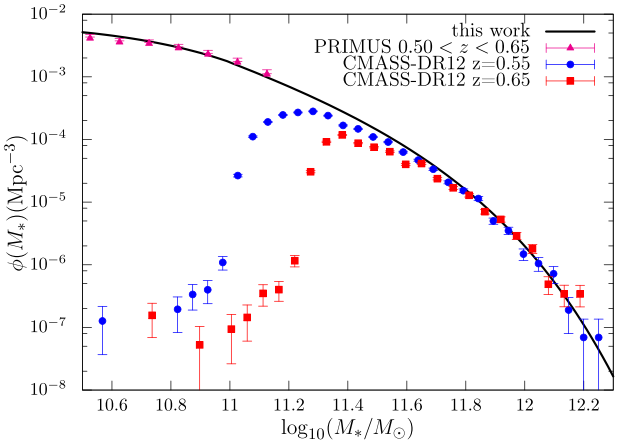


Figure 3. SMF from BOSS CMASS DR12 sample. Circles and squares show the stellar mass distribution for two redshift bins from the Portsmouth DR12 catalogue. Poissonian errors are included. The solid line shows the estimate of the SMF for this work, which is constructed combining the high-mass end of the BOSS sample and Guo et al. (2010) for the low-mass range ($\log_{10} M_* < 11.0$). In order to compare with a complete sample in the redshift range 0.5–0.65, we include the PRIMUS SMF (triangles) in the low-mass regime.

to match the stellar mass catalogues (LEGACY and BOSS) and the LSS BOSS CMASS catalogue. Once the stellar masses of the observed sample are assigned, we need to construct an SMF which describes the mass distribution.

The Portsmouth DR12 catalogue has the SMF that is different from SMF of previous surveys (Maraston et al. 2013). Fig. 3 shows the mass distribution of the CMASS DR12 for two different redshift regions. A detailed study of the Portsmouth catalogues and other stellar mass catalogues was reported by Maraston et al. (2013).

Due to the selection function in the BOSS data, we do not have the information on the shape of the SMF at low masses. There are different ways of handling this problem. For example, Leauthaud et al. (2016) use the stripe 82 massive galaxy catalogue to compute the SMF of the BOSS data. We use a different approach: for the high-mass end, we use the Portsmouth stellar masses and we combine them with Guo et al. (2010) results to describe the low-mass regime. Specifically, to compute the SMF for masses larger than $3.2 \times 10^{10}\,\mathrm{M}_\odot$ (which is the mass range used in the CMASS sample).

In order to construct the SMF, we select galaxies in the redshift range $0.55 < z < 0.65$, because this is the most complete range for the CMASS sample (see Montero-Dorta et al. 2014). We combine the CMASS sample for masses larger than $2.5 \times 10^{11}\,\mathrm{M}_\odot$ and the SMF from Guo et al. (2010) for low masses. We fit both results using a double Press–Schechter mass function (Press & Schechter 1974) with the parameters given in Table 1.

Fig. 3 presents the SMF used in this work. We also add in Fig. 3 the PRIMUS SMF (Moustakas et al. 2013) in the redshift range $0.5 < z < 0.65$ with the purpose of comparing the low-mass range of our SMF with a complete sample in the same redshift and mass ranges. A detailed comparison of the Portsmouth catalogues and other stellar mass catalogues is presented in Maraston et al. (2013).

In our analysis, we do not include redshift evolution of the SMF. This approximation agrees with results of the PRIMUS survey (Moustakas et al. 2013), which is a complete survey in the redshift range we study. Moustakas et al. (2013) show that there is only a small evolution of the SMF in the CMASS redshift range.

Table 1. Parameters of the double Press–Schechter SMF for this work.

Mass range ($\mathrm{M}_\odot$)	ϕ_* ($\mathrm{Mpc}^3\,\log_{10}\,\mathrm{M}_\odot^{-1}$)	α	$\log_{10} M_*$ ($\mathrm{M}_\odot$)
$\log_{10} M_* \leq 11.00$	4.002×10^{-3}	−0.938	10.76
$\log_{10} M_* > 11.00$	2.663×10^{-4}	−2.447	11.42

4.3 Production of light-cones

We implement a method to generate light cones from snapshots of cosmological simulations. This method has been implemented previously (see e.g. Blaizot et al. 2005; Kitzbichler & White 2007). The SUGAR code works with cubic boxes using positions and velocities of dark matter haloes as inputs. We will now describe the procedure which we use to construct mocks for the CMASS sample.

BigMD-BOSS light cones are constructed from the BigMDPL simulation which is large enough ($2.5\,h^{-1}$ Gpc) to map the CMASS NGC. We use the periodic boundary conditions to maximize the use of the volume (Manera et al. 2013), but we do not reuse any region of the box. So there are no duplicated structures in our light cone.

The first step in the construction of the light cone is to locate the observer ($z = 0$) and transform from comoving Cartesian coordinates to equatorial coordinates (RA,Dec.) and redshift. To include the effects of galaxy peculiar velocities in the redshift measurements, we transform the coordinates of the haloes to redshift space using

$$\boldsymbol{s} = \boldsymbol{r}_\mathrm{c} + \frac{\boldsymbol{v}\cdot\hat{\boldsymbol{r}}}{aH(z_\mathrm{real})}, \tag{12}$$

where $\boldsymbol{r}_\mathrm{c}$ is the comoving distance in real space, $\boldsymbol{v}$ is the velocity of the object with respect to Hubble flow, $\hat{\boldsymbol{r}}$ is the line-of-sight direction, a is the scale factor and H is the Hubble constant at z_real, which is the redshift corresponding to r_c, and is computed from

$$r_\mathrm{c}(z_\mathrm{real}) = \int_0^{z_\mathrm{real}} \frac{c\,\mathrm{d}z}{H_0\sqrt{\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda}}, \tag{13}$$

where c is light speed and H_0 is the Hubble constant in $\mathrm{s}^{-1}\,\mathrm{Mpc}^{-1}$ km. Using equations (12) and (13), it is possible to compute $s(z_\mathrm{obs})$, where z_obs is the observed redshift. The next step is to select objects from each snapshot to construct shells for the light cone. Thus, an object with redshift z_obs, which comes from a snapshot at $z = z_i$, will be selected if $(z_i + z_{i-1})/2 < z_\mathrm{obs} \leq (z_i + z_{i+1})/2$. We repeat this process for all objects in snapshots between $z = 0.43$ and 0.7. We fix the number density in each shell following the radial selection function of the BOSS CMASS sample. Fig. 4 shows the comparison between the radial selection function of the observed data and the one obtained on the BigMD-BOSS light cone.

Finally, we apply the angular CMASS NGC mask to match the area of the observed sample. The angular completeness is taken into account by downsampling the regions where it is smaller than one. As was done in the BOSS CMASS catalogue, we select regions in the sky with completeness weight larger than 0.7. Due to the presence of random numbers in the selection process, the observed radial selection function can have variations of ∼4 per cent. Fig. 4 presents the standard deviation from 100 MD-PATCHY mocks to examine the effect of different seeds on the random generator.

Fig. 1 shows the angular distribution of the BigMD-BOSS light cone. In order to reproduce the angular distribution, we applied the BOSS CMASS DR12 NGC geometry, and, in addition, we applied veto mask to exclude exactly the same regions removed in the observed data. Fig. 2 presents a 2D comparison of the spatial

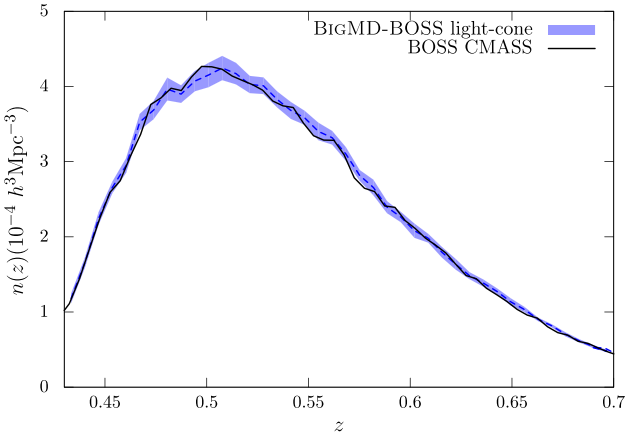


Figure 4. The comoving number density of BOSS CMASS DR12 NGC (black line) compared to the comoving number density of the BigMD-BOSS light cone (dashed line). Shaded area comes from 100 MD-PATCHY mocks.

galaxy distribution between the BigMD-BOSS light cone and the BOSS CMASS data.

4.4 Stellar mass incompleteness

This paper focuses in the production of mocks which can describe the full CMASS DR12 sample. Instead of extracting a subsample which has better completeness in terms of stellar mass, we 'model' the observed stellar mass incompleteness. This model not only accounts for the incompleteness at small masses (presented across the complete redshift range), but also incompleteness in the high-mass end, which is important for $z \lesssim 0.45$. Fig. 5 compares the results of our modelling in the BigMD-BOSS light cone to the observed data for three different redshifts.

In order to reproduce the observed stellar mass distribution, we construct a continuous function by interpolation. Once the abundance matching is applied and galaxies are assigned to dark matter haloes, we select galaxies by downsampling based on the observed stellar mass distribution. This process is repeated for 20 different redshifts (corresponding to the snapshots of the simulation). Then, in order to construct the observed stellar mass distribution corresponding to snapshot at $z = z_i$, a galaxy with redshift z_g in the stellar mass catalogue will be selected if $(z_i + z_{i-1})/2 < z_g \leq (z_i + z_{i+1})/2$. This model has an important impact on the scatter applied to the abundance matching. Since bias is as a function of stellar mass, incompleteness that varies as a function of stellar mass will affect the overall bias as well. This effect reduces the amplitude of the clustering, which implies that a smaller scatter is required to reproduce the signal of the observed clustering. If we ignore the incompleteness effect, we can still reproduce the clustering in the two-point correlation function (2PCF). However, this scatter is not the intrinsic one, and the final stellar mass distribution will not match the observed sample. Favole et al. (2015a) show a similar model to reproduce the incompleteness of the Emission Line Galaxies population from the BOSS sample.

Most galaxies in the CMASS sample are red galaxies. However, there is also a fraction of blue galaxies in the data. In addition, the blue sample is less complete than the red one (Montero-Dorta et al. 2014).The random downsampling of galaxies in the BigMD-BOSS light cone does not distinguish between both populations, which can produce potential systematics due to the different completeness of both samples. In this study, we reproduce the observed stellar mass distribution by downsampling galaxies from a no-evolving SMF. However, SMF evolves with redshift, which can produce underestimation of the incompleteness for some ranges of stellar mass and overestimation for other ranges.

4.5 Fibre collisions

A feature of the BOSS fibre-fed spectrograph is that the finite size of the fibre housing makes impossible to place fibres within 62 arcsec of each other in the same plate. This causes a number of galaxies to not have a fibre assigned and hence, there is no measurement of their redshift. We model the effect of fibre collisions as follows. A total of 5 per cent of the CMASS targets could not been observed due to the fibre collisions. These objects have an important effect at scales $\lesssim 10\ h^{-1}$ Mpc. In this paper, we model the fibre collision effect by adopting the method described in Guo, Zehavi & Zheng (2012).

The first step is to find the maximum number of galaxies that could be assigned fibres. This decollided sample (D_1) is a set of galaxies which are not angularly collided with other galaxies in this subsample. The second population (D_2) are the potentially collided galaxies. Each galaxy in this subsample is within the fibre collision scale of a galaxy in population 1. We must determine from the observed sample the fraction of collided galaxies (D_2') in the D_2

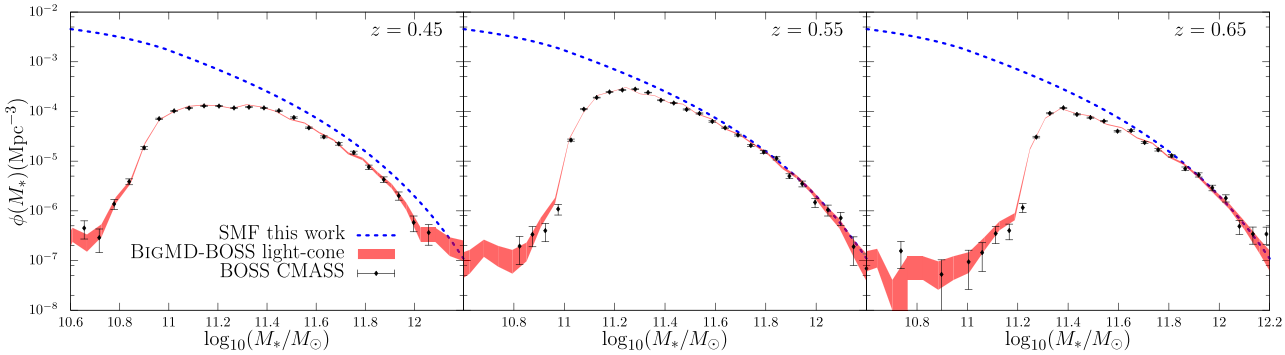


Figure 5. Incompleteness modelling for three different redshift bins. Shaded area shows the BigMD-BOSS light cone; dots are the measurements from the CMASS Portsmouth catalogue. In both cases, Poissonian errors are used. Dashed line represents the SMF adopted in this work. We select three bins as an example to show the results of the incompleteness modelling implemented in this work. Stellar mass distribution in the BigMD-BOSS light cone is produced by downsampling galaxies from the SMF adopted. Left-hand panel shows the incompleteness at low redshift in the high mass of the SMF.

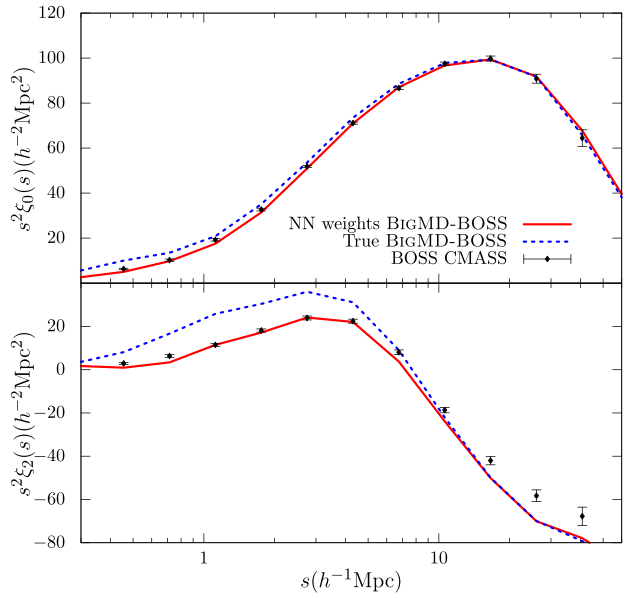


Figure 6. Monopole (top panel) and quadrupole (bottom panel) of the redshift-space correlation function for the BigMD-BOSS light cone before and after applying fibre collisions. Fibre collisions are corrected using nearest-neighbour (NN) weights. The effects of the fibre collisions are stronger in the quadrupole, with important differences for scales $s \lesssim 7\ h^{-1}$ Mpc. The impact on the monopole is smaller. The fibre collision assignment is an approximative method which can introduce systematic effects. In order to avoid these effects, we select the range 2–30 h^{-1} Mpc to fit the monopole with the scatter parameter, $\sigma_{\mathrm{HAM}}(V_{\mathrm{peak}}|M_*)$.

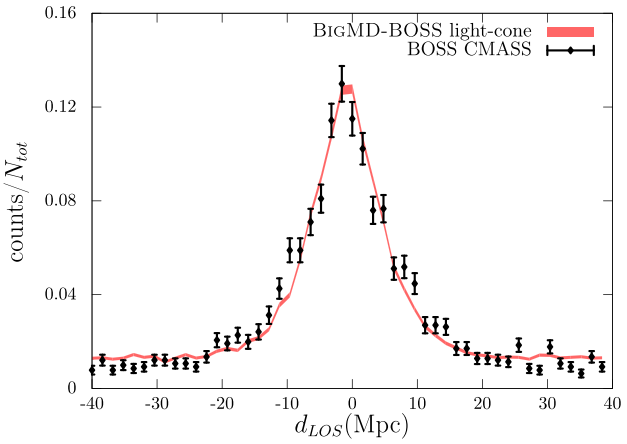


Figure 7. Line-of-sight displacement of a collided galaxy due to the fibre collision. The figure shows the number of counts per bin divided by the total number of collided galaxies. Uncertainties were computed using Poissonian errors.

group (i.e. D_2''/D_2) for sectors covered by different numbers of tiles. Finally, we randomly select the fraction D_2''/D_2 to the D_2 galaxies in the mocks to be collided galaxies.

Fig. 6 displays the impact of the fibre collisions on the correlation function in redshift space. The effect in the monopole becomes very important for scales smaller than 1 h^{-1} Mpc. However, the quadrupole is more sensitive to this effect, with big impact for scales smaller than 10 h^{-1} Mpc. The assignment of fibre collisions has an important impact on the fraction of satellites. Before fibre collisions the satellite fraction of the light cone is 11.8 per cent, and after the assignment is equal to 10.5 per cent. This effect reduces the central–satellite pairs, which have a strong impact on the quadrupole.

Unlike Guo et al. (2012), we only use nearest-neighbour weights for both samples. Our goal is to compare the results of the abundance matching with data, so that we implement the same fibre collision correction to our light cone as observed data.

When nearest-neighbour weights are applied, a collided galaxy will be 'moved' from its original coordinates to the position of its nearest neighbour. Fig. 7 presents the line-of-sight displacement of those collided galaxies from their original positions.

The displacement for the simulation shown in Fig. 7 is computed using the old and new positions of the collided galaxies. In CMASS data, the displacement is calculated using the overlapping tiled regions of the survey where the spectroscopic redshifts of both galaxies within the fibre collision angular scale are resolved. Fig. 7 demonstrates an excellent agreement between our model and the observed data, suggesting that the combination between the clustering at small scales of the simulation and the fibre collision model used in the mock has a reasonable agreement with observations.

5 MODELLING BOSS CMASS CLUSTERING

The clustering signal in the abundance matching is determined by two quantities: the number density and the scatter in the M_*-V_{peak} relation. The number density is fixed by the radial selection function of the observed sample. In order to find a scatter value that reproduces the clustering of the CMASS sample, we fit the monopole of the correlation function in redshift space. The following sections present the results of this monopole fitting, and the prediction of our model of the quadrupole in redshift space, projected correlation function, monopole in Fourier space and the 3PCF.

BigMD-BOSS light cone covers the same volume as CMASS sample between redshift $z = 0.43$ and 0.7. In order to have a good estimation of the uncertainties in our measurements, we use 100 MD-PATCHY mocks (Kitaura et al. 2016, companion paper). These mocks are produced using five boxes at different redshifts that are created with the PATCHY code (Kitaura et al. 2014). This code matches the clustering of the galaxy catalogues for each redshift bin. The MD-PATCHY mocks are based on the BigMDPL simulation, and they are produced with the same cosmology used in this work. To compute errors, we use the square root of the diagonal terms of the covariance matrix defined as

$$C_{ii} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2, \tag{14}$$

where N is the number of mock catalogues and X is the statistical quantity measured.

5.1 Two-point clustering: result from model and observations

In order to compute the correlation function for our light cone and the observed data, we use a Landy & Szalay estimator (Landy & Szalay 1993). The correlation function is defined by

$$\xi(r) = \frac{DD - 2DR + RR}{RR} \tag{15}$$

where DD, DR and RR represent the normalized data–data, data–random and random–random pair counts, respectively, for the distance range $[r - \Delta r/2, r + \Delta r/2]$.

In this paper, we use random catalogues 20 times larger than the data catalogues. In order to estimate the projected correlation function and the multipoles of the correlation function, we use the

2D correlation function, $\xi(r_\mathrm{p}, \pi)$, where $s = \sqrt{r_\mathrm{p}^2 + \pi^2}$, r_p is the perpendicular component to the line of sight and π represents the parallel component. The correlation function of the BigMD-BOSS light cone is computed using close pair weights and FKP weights (Feldman, Kaiser & Peacock 1994),

$$w_\mathrm{FKP} = \frac{1}{1 + n(z)P_\mathrm{FKP}}, \tag{16}$$

where $n(z)$ is the number density at redshift z and $P_\mathrm{FKP} = 20\,000\ h^{-3}\ \mathrm{Mpc}^3$. We use the FKP weights to optimally weight regions with different number densities. In the case of the BOSS CMASS sample, we use the galaxy weights given in equation (9) and in addition the FKP weights. The total weights for the data used in our analysis are $w_\mathrm{tot} = w_\mathrm{FKP} w_\mathrm{g}$.

Note that P_FKP is chosen to minimize the variance of power spectrum measurements. For the correlation function measurements, one should use the optimal weight from Hamilton (1993),

$$w_\mathrm{H} = 1/(1 + n(z)J_w), \tag{17}$$

where

$$J_w = \int_0^r \xi(r)\,\mathrm{d}V. \tag{18}$$

However, since we are fixing w_FKP or w_H to be a constant to simplify the computation, we expect that w_H should be similar to w_FKP. In any case, the choice of optimal weight will not bias the measurements.

5.1.1 *Redshift-space correlation function*

Previous works demonstrated the impact of the scatter in the clustering signal of a mock generated with the abundance matching (e.g. Reddick et al. 2013). In this study, we search for a scatter parameter ($\sigma_\mathrm{HAM}(V_\mathrm{peak}|M_*)$) which reproduces the monopole of the correlation function and provides the prediction for other quantities. The multipoles of the 2PCF, in redshift space, are defined by

$$\xi_l(s) = \frac{2l+1}{2} \int_{-1}^{1} \xi(r_\mathrm{p}, \pi) P_l(\mu)\,\mathrm{d}\mu, \tag{19}$$

where

$$\mu = \frac{\pi}{\sqrt{r_\mathrm{p}^2 + \pi^2}} \tag{20}$$

and $P_l(\mu)$ is the Legendre polynomial. We will present results for the monopole ($l = 0$) and the quadrupole ($l = 2$).

To find the best value, we fit the clustering using the monopole in the redshift space for the range 2–30 h^{-1} Mpc. The top panel in Fig. 8 shows the results of the fitting compared to the CMASS DR12 data. Errors in Figs 8 and in 9 are computed using 100 MD-PATCHY mocks (Kitaura et al. 2016, companion paper). The parameter that best reproduces the clustering in the monopole is $\sigma_\mathrm{HAM}(V_\mathrm{peak}|M_*) = 0.31$. This result is in agreement with previous works on abundance matching (Trujillo-Gomez et al. 2011; Nuza et al. 2013; Reddick et al. 2013).

The simulation provides a good agreement with data in the monopole for scales smaller than 50 h^{-1} Mpc. However, the bottom panel in Fig. 8 shows a disagreement in the quadrupole for scales smaller than 0.7 h^{-1} Mpc, which can be due to the method used to assign the fibre collisions in the BigMD-BOSS light cone; for this reason, we do not analyse these scales. An additional disagreement is found at scales larger than 6 h^{-1} Mpc, which will be commented in the last section of this work. Nuza et al. (2013) use

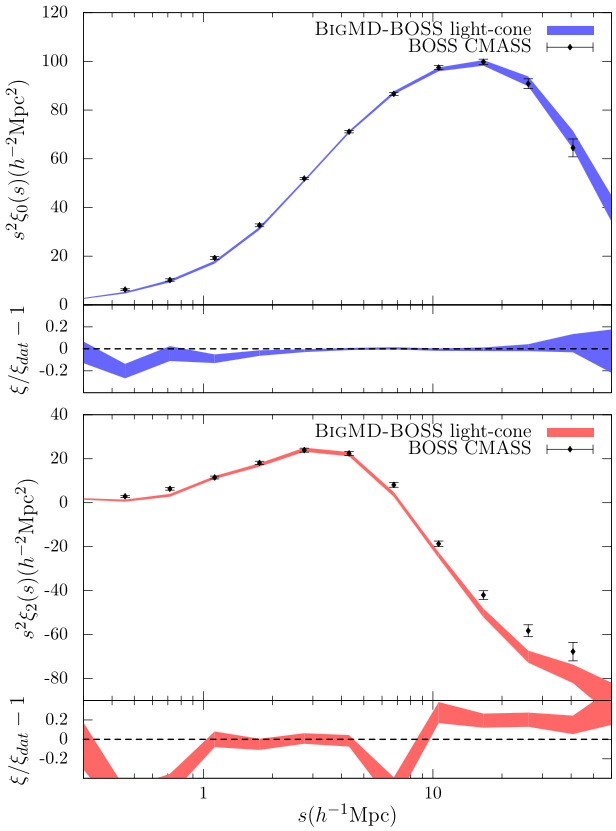


Figure 8. Top panel: monopole in redshift space from CMASS DR12 sample (black points). The shaded area represents the modelling of the monopole using the BigMD-BOSS light cone. Bottom panel: quadrupole in redshift space from CMASS DR12 sample compared with the theoretical prediction from the BigMD-BOSS light cone. Error bars were computed using MD-PATCHY mocks. Small panels show the ratio between the model and the observed data. Fitting of the monopole is performed between 2 and 30 h^{-1} Mpc. The observed monopole is in good agreement with our model for scales larger than 2 h^{-1} Mpc. However, the quadrupole shows tensions with observations for scales <1 h^{-1} Mpc and $5 >$ h^{-1} Mpc.

the MultiDark simulation with $\Omega_\mathrm{m} = 0.27$. Comparing their results for the monopole, we obtain a better agreement for scales larger than 10 h^{-1} Mpc, mainly due to the difference in cosmologies used in this work.

Fig. 9 shows the prediction of the monopole and quadrupole for large scales compared to the observed data. Discrepancies for some values between the model and the data at scales larger than 60 h^{-1} Mpc could not be due only to the cosmic variance. Differences at the baryon acoustic oscillation (BAO) scales are of the order of 1σ errors while for large scales differences can be of the order of 2σ or 3σ. In Fig. 9, we can see that the BOSS CMASS correlation function at large scales is systematically shifted. This excess of power in the correlation function monopole could be due to the potential photometric calibration systematics which only affect very large scales. Huterer, Cunha & Fang (2013) make a detailed study about the photometric calibration errors and their implication in the measurements of clustering and demonstrate that calibration uncertainties generically lead to large-scale power.

5.1.2 *Projected correlation function*

The projected correlation function is a quantity which is insensitive to the impact of the redshift-space distortion and provides an

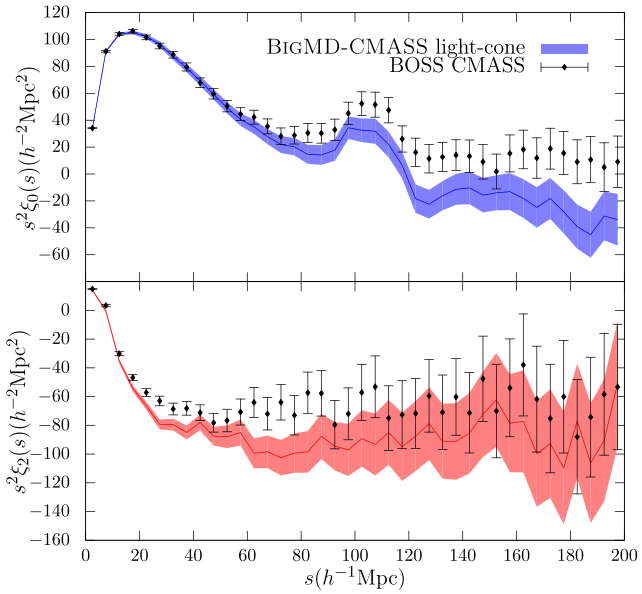


Figure 9. Monopole (top panel) and quadrupole (bottom panel) of the redshift-space correlation function. The shaded areas are the model predictions for large scales using a single light cone. Error bars were computed using MD-PATCHY mocks. Differences in the quadrupole are the same shown in Fig. 8. The monopole has a good agreement up to 100 h^{-1} Mpc. However, large scales present significant difference, but this can be due to the cosmic variance and remaining systematics in the data. These differences are within 2σ errors.

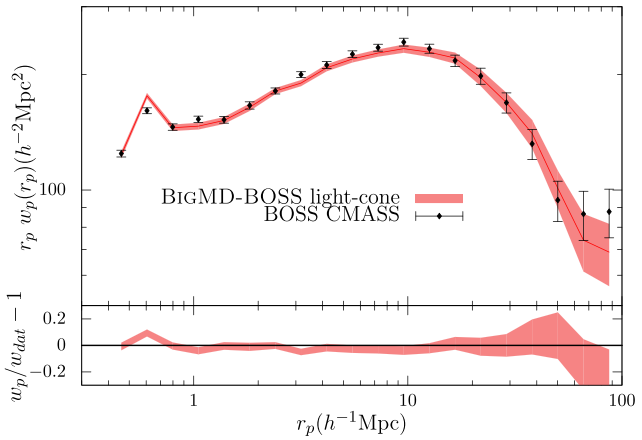


Figure 10. Projected correlation function prediction from the BigMD-BOSS light cone (shaded region) compared to the BOSS CMASS sample. The width of the shaded area represents 1σ errors, computed using MD-PATCHY mocks. Our model reproduces the clustering for all relevant scales. Scales $<0.6\,h^{-1}$ Mpc are dominated by fibre collision effects.

approximation to the real-space correlation function (Davis & Peebles 1983). The projected correlation function is defined as the integral of the 2D correlation function, $\xi(r_\mathrm{p}, \pi)$, over the line of sight:

$$w_\mathrm{p}(r_\mathrm{p}) = 2\int_0^\infty \xi(r_\mathrm{p}, \pi)\,\mathrm{d}\pi. \tag{21}$$

In order to compute $w_\mathrm{p}(r_\mathrm{p})$ from the discrete correlation function (equation 15), we use the estimator

$$w_\mathrm{p}(r_\mathrm{p}) = 2\sum_i^{\pi_\mathrm{max}} \xi(r_\mathrm{p}, \pi_i)\Delta\pi_i. \tag{22}$$

We adopt a linear binning in the light-of-sight direction, $\Delta\pi_i = \Delta\pi = 5\ h^{-1}$ Mpc. We selected $\pi_\mathrm{max} = 100\ h^{-1}$ Mpc. Nuza et al. (2013) find convergence of the projected correlation for this scale. Fig. 10 shows the results found for the BigMD-BOSS light cone compared to the CMASS data. Error bars were computed using 100 MD-PATCHY mocks.

Fig. 10 reveals a discrepancy at scales $\approx 3\ h^{-1}$ Mpc. However, results are in agreement at the 2σ level, so we can consider the data consistent with the prediction of our model. Scales below $0.5\,h^{-1}$ Mpc are dominated by fibre collision. Due to this effect, the clustering declines rapidly.

5.1.3 Fourier space

The power spectra for the BOSS CMASS sample with nearest angular neighbour upweighted weights and the BigMDPL are computed using the Feldman et al. (1994) power spectrum estimator modified to account for the systematic weights of the galaxies. In BOSS CMASS, each galaxy is assigned a systematics weight (equation 9), which is accounted for in the estimator. For the BigMD-BOSS light cone, we set $w_\mathrm{g} = w_\mathrm{cp}$, for the power spectrum using nearest-neighbour upweighted fibre collisions weights, and $w_\mathrm{g} = 1$ for the true power spectrum.

The power spectrum for the BOSS CMASS sample is computed using the method described in Hahn et al. (in preparation) in order to correct the effects of fibre collisions on smaller scales. The fibre collision correction method reconstructs the clustering of fibre-collided pairs by modelling the distribution of the line-of-sight displacements between them using pairs with measured redshifts. In addition, the method corrects fibre collisions in the shot-noise correction term of the power spectrum estimator. In simulated mock catalogues, the correction method successfully reproduces the true power spectrum with residuals $\lesssim$1 per cent at $k \sim 0.3\ h$ Mpc^{-1} and $<$10 per cent at $k \sim 0.9\ h$ Mpc^{-1}. The top panel of Fig. 11 compares the fibre collision and systematics corrected BOSS CMASS power spectrum to the true power spectrum of BigMD-BOSS light cone, showing remarkably good agreement between data and model. Figs 8 and 11 confirm that the standard HAM is accurate in the modelling of the clustering not only at large scales, but also in the one-halo term.

Monopoles from our model and the BOSS CMASS data using fibre collision weights are shown in the bottom panel of Fig. 11. Both power spectra agree for k smaller than $1\ h$ Mpc^{-1}. The BigMD-BOSS light cone and the observed data have a remarkably good agreement in the BAO region (inset panel Fig. 11), which is not seen in the correlation function (Fig. 9). This difference can be due to remaining systematics that have a bigger impact on the correlation function than in the power spectrum. The agreement between our model and the observed data, for the true power spectrum and the nearest-neighbour corrected power spectrum, demonstrates that the method used to assign fibre collisions in the BigMD-BOSS light cone is a good approach to simulate this effect.

As we discussed in Section 5.1.1, the disagreement between the model and the data in the correlation function monopole could be due to potential photometric calibration systematics. The effect on the power spectrum will be limited to very small k, so that it has less impact on the BAO scales. However, this excess of power does not have impact on BAO measurements from correlation functions when we marginalize the overall shape (see Chuang et al. 2013; Ross et al., in preparation).

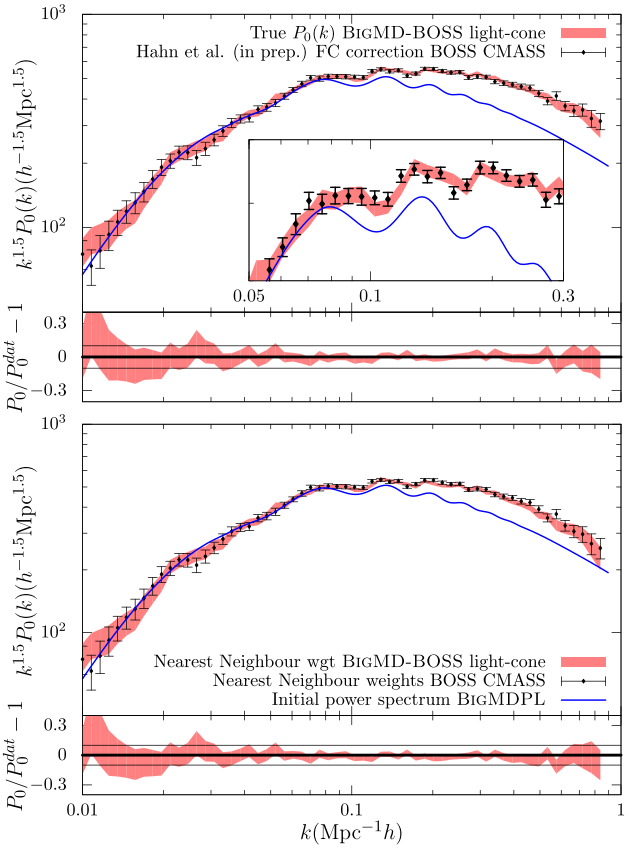


Figure 11. Monopole of power spectrum from the BigMD-BOSS light cone and the CMASS DR12 sample. Top panel: the true power spectrum for our light cone compared to the CMASS DR12 data corrected by fibre collisions using Hahn et al. (in preparation) method. Solid curve shows the initial matter power spectra of the BigMDPL simulation scaled to match the amplitude of fluctuations at long waves. A remarkable agreement between the data and the model is found for scales $k \lesssim 1\,h\,\mathrm{Mpc}^{-1}$. Bottom panel: the comparison between simulation and observed data using nearest-neighbour weights ($w_{\rm cp}$). In addition to $w_{\rm cp}$, observed measurements include systematics weights: $w_{\rm star}$, $w_{\rm zf}$ and $w_{\rm see}$. The agreement between the data and the model, in both panels, shows the good performance of the fibre collision assignment in the light cone. In bottom subpanels, dashed lines represent an accuracy level of 10 per cent.

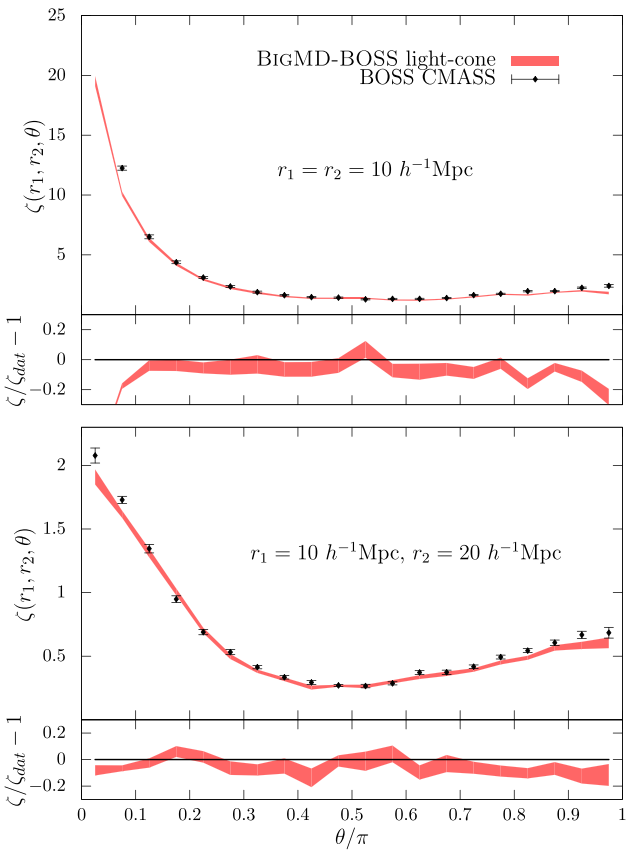


Figure 12. Top panel: BOSS CMASS DR12 3PCF compared with the model prediction of this work. Shaded area shows 1σ uncertainties, with limits $r_1 = 10\ h^{-1}$ Mpc and $r_2 = 20\ h^{-1}$ Mpc. Bottom panel: 3PCF for limits $r_1 = r_2 = 10\ h^{-1}$ Mpc. The BigMD-BOSS light cone can reproduce almost all scales between 2σ errors.

5.2 Three-point correlation function

We are also interested in comparing the prediction of the 3PCF using the HAM on the BigMDPL simulation with the observed data. The 3PCF provides a description of the probability of finding three objects in three different volumes. In the same manner as the 2PCF, the 3PCF is defined as

$$\zeta(r_{12}, r_{23}, r_{31}) = \langle \delta(r_1)\delta(r_2)\delta(r_3)\rangle, \tag{23}$$

where $\delta(r)$ is the dimensionless overdensity at the position r and $r_{ij} = r_i - r_j$. We use the Szapudi & Szalay estimator (Szapudi & Szalay 1998)

$$\zeta = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}. \tag{24}$$

Fig. 12 displays our prediction compared with the BOSS CMASS data. We see the results for two kinds of triangles: $r_1 = r_2 = 10\,h^{-1}$ Mpc and $r_1 = 10\,h^{-1}$ Mpc, $r_2 = 20\,h^{-1}$ Mpc, where θ is the angle between r_1 and r_2.

A good agreement in the shape of the 3PCF is seen in Fig. 12 between our prediction and the data. Most of the points are in agreement within 2σ errors for both configurations represented in Fig. 12. However, the BigMD-BOSS light cone is underestimating the 3PCF for $\theta \sim 0$ and $\theta \sim \pi$. Guo et al. (2015b) find similar discrepancies for those scales, which can be produced by velocity effects and can be corrected including a velocity bias. Therefore, the disagreement in the 3PCF and in the quadrupole of the correlation function can be caused by the same kind of effects.

5.3 Stellar-to-halo mass relation

The stellar-to-halo mass ratio (SHMR) is an important quantity to evaluate if the simulated light cone is providing a realistic halo occupation. In this way, we use results from weak lensing, which is one of the most powerful mechanisms to know the observational SHMR. Fig. 13 shows the SHMR predicted by the BigMD-BOSS light cone and measurements in the Canada–France–Hawaii Telescope (CFHT) Stripe 82 Survey (Shan et al. 2015). In order to ensure the convergence of the haloes in our prediction, we select haloes with masses larger than $5.2 \times 10^{12}\,\mathrm{M}_\odot$. This limit is 150 dark matter particles which give convergence for subhaloes (Klypin et al. 2015).

Predictions of the abundance matching are in agreement with the weak lensing data. In Fig. 13, shaded blue area shows the intrinsic scatter measured. The dependence between scatter and stellar mass

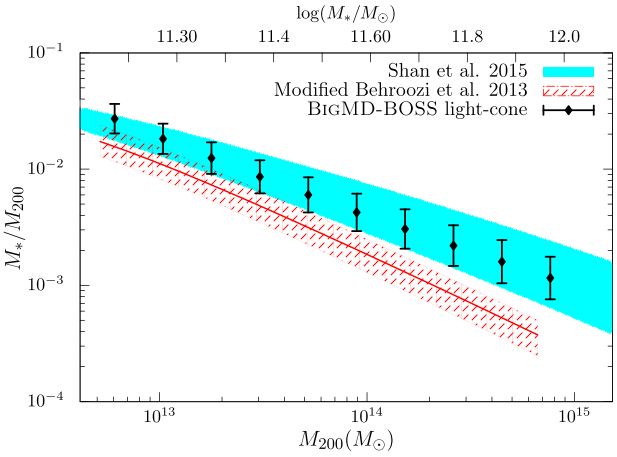


Figure 13. Stellar-to-halo mass ratio. The shaded blue area represents the best fit of the stellar-to-halo mass relation measured using weak lensing in the CFHT Stripe 82 Survey (Shan et al. 2015). The red area represents previous HAM result from Behroozi, Wechsler & Conroy (2013c). The analysis in Behroozi et al. (2013c) was modified using the Planck cosmology parameters and changing the definition of the halo mass. Black dots are the prediction from the HAM-BigMD-BOSS light cone. Differences between our model and Behroozi et al. (2013c) are mainly due to the SMF adopted in both works. Scatter between M_{200} and M_* is similar between the data and our model. We adopted constant scatter while observed data suggest a dependence of the scatter with the stellar mass.

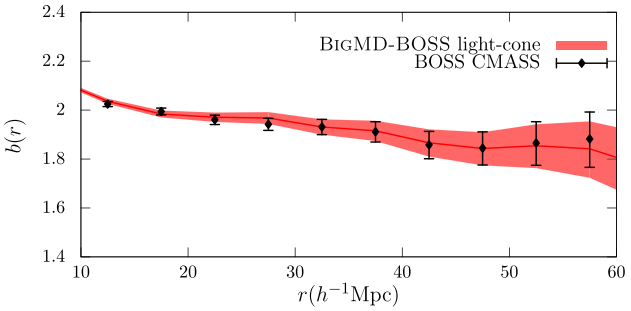


Figure 14. Scale-dependent galaxy bias from the model presented in this work. We measure the bias with respect to the correlation function of dark matter in the BigMDPL light cone for the data and the model. There is an excellent agreement between the CMASS observations and the predictions of the HAM-BigMD-BOSS light cone.

is clear. It is also shown in the abundance matching (e.g. Trujillo-Gomez et al. 2011; Reddick et al. 2013). However, our HAM model uses a constant scatter to reproduce the clustering. This approximation can generate the disagreement in the scatter between data and mock. The red area in Fig. 13 indicates the results from Behroozi et al. (2013c). We modify Behroozi et al. (2013c) in order to use the same definition of halo mass and implement the Planck cosmology in the analysis. The SMF assumptions can be one of the origins for the disagreement between both predictions. While we use the BOSS DR12 stellar mass catalogues to estimate the SMF, Behroozi et al. (2013c) use the PRIMUS SMF (Moustakas et al. 2013). The difference in how the stellar mass catalogues handle profile fitting produces a variation in the high-mass end of both SMFs. This effect causes important difference at large stellar mass between both predictions.

Shankar et al. (2014) present the stellar-to-halo mass relation assuming different mass functions and compare their results with recent models. They find differences between Behroozi et al. (2013c) and Maraston et al. (2013) similar to the one shown in our Fig. 13. Shankar et al. (2014) also find that an intrinsic scatter in stellar mass at fixed halo mass of 0.15 dex is needed to reproduce the BOSS clustering. This result is in agreement with our model, which predicts an intrinsic scatter in stellar mass of 0.14 dex at a fixed halo mass.

5.4 Bias prediction

Using the HAM-BigMD-BOSS light cone and its corresponding dark matter light cone, we can estimate the real-space bias, $b(r)$, solving the equation (Kaiser 1987; Hamilton 1992)

$$\xi(s) = \left(1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right) b(r)^2 \xi_{\rm DM}(r), \qquad (25)$$

where $\beta \approx f/b$ is the redshift-space parameter and $f(z = 0.55) = 0.77$ (Planck cosmology).

Fig. 14 shows the linear bias, which is in agreement with previous papers that reproduced the CMASS clustering (see Nuza et al. 2013). For the data and the model, we use the dark matter correlation function from the BigMD simulation. For the scales shown, the scale-dependent bias factor is in the range 1.8–2. We use the BigMD dark matter light cone to estimate the relative bias of the CMASS sample to this catalogue.

6 DISCUSSION

The BigMD-BOSS light cone is designed to reproduce the full BOSS CMASS sample between redshift 0.43 and 0.7, including observational effects. In order to recover the information at small scales, similar papers (e.g. Nuza et al. 2013; Guo et al. 2014, 2015c) correct the observed data by fibre collision (see Guo et al. 2012; Hahn et al., in preparation). In this work, we assign fibre collisions to galaxies in the light cone, and we use nearest-neighbour weights in the data and in the model. Our model can be useful to test methods that recover the clustering in the fibre collision region (Guo et al. 2012) or in the production of mocks for covariance matrices (Kitaura et al. 2016, companion paper). The fibre collision assignment adopted in this work can reproduce in a good way this observational effect (Fig. 11). However, this approach can introduce small systematics that we do not include in our modelling.

White et al. (2011) model the full CMASS clustering. They find a good fit of the HOD parameters to reproduce the observed data. However, they cannot describe the small scales because they only include close pair weights in the data measurements, which cannot recover the small-scale clustering (Guo et al. 2012). Nuza et al. (2013) also reproduce with a good agreement the CMASS data using a standard HAM model; they correct by fibre collision using the method explained in Guo et al. (2012). Our paper continues the work presented in Nuza et al. (2013), including light-cone effects, redshift evolution, radial selection function, etc. All these papers can reproduce the clustering of the full CMASS sample.

Recent papers show tensions between models and observed data when a most careful selection is done. Guo et al. (2015c) study a volume-limited LRG sample in the redshift range of $0.48 < z < 0.55$ of the CMASS sample. They need a galaxy velocity bias to describe the clustering of the most massive galaxies ($\sim 10^{13}$–$10^{14}\ h^{-1}\ \mathrm{M}_\odot$) using HOD. Saito et al. (2015) show an extension of the HAM to describe the colour dependence of the clustering for the CMASS sample. Guo et al. (2015a) present a comparison

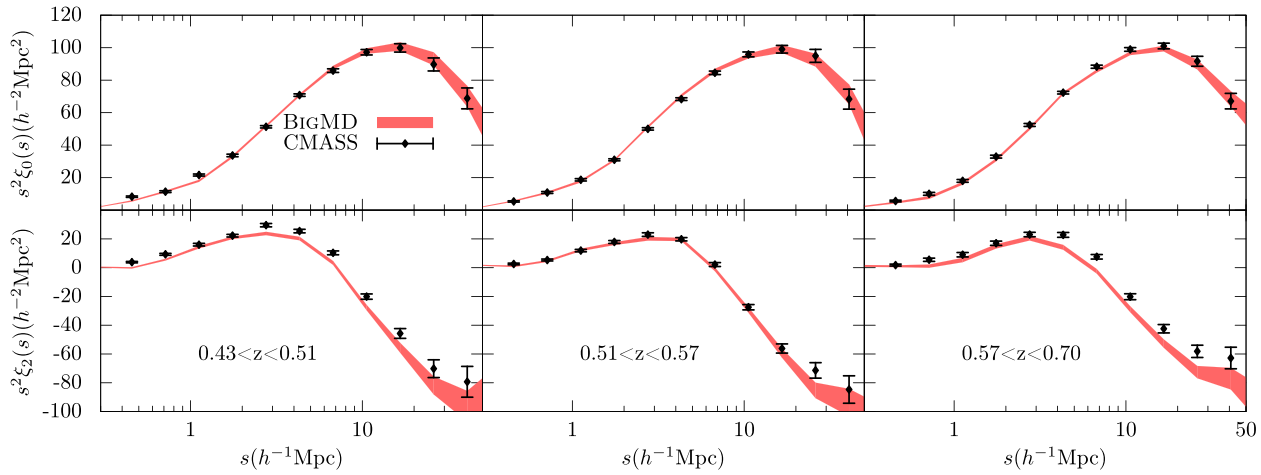


Figure 15. Monopole and quadrupole of the redshift-space correlation function of the CMASS DR12 sample compared to the HAM-BigMD-BOSS light cone for three redshift bins. The monopole is fitted for all redshift ranges. The middle bin is the most complete range in the CMASS sample, and also the best reproduced quadrupole. We perform a HAM with three different scatter parameters to fit each of the redshift bins. Differences at low and high redshift can be due to target selection effects we do not include in this study. Another source of discrepancy can be the relation between the scatter and the more massive galaxies (Saito et al. 2015).

between HOD and HAM models; they also modify the standard HAM model in order to reproduce clustering at different luminosity cuts. Favole et al. (2015b) present a study of the blue population properties compared to the red galaxies. They present a modified HOD which allows them to include both samples in the same mock catalogue. The clustering dependence on stellar mass (luminosity) is not implemented in our model, and we do not distinguish between blue and red galaxies. Our implementation of the HAM and stellar mass incompleteness is capable of reproducing the full CMASS sample, including a big amount of data in our analysis. Zu & Mandelbaum (2015) present a modified HOD in order to include the stellar mass incompleteness (iHOD). This model combines galaxy cluster and galaxy–galaxy lensing and allows one to increase ∼80 per cent the number of modelled galaxies than the traditional HOD models.

We find the largest discrepancy between our model and the data in the quadrupole measurements (Fig. 8). For scales larger than 10 h^{-1} Mpc, this difference is within the 3σ errors. The disagreement for $s < 1\ h^{-1}$ Mpc is larger than 20 per cent. However, this can be due to the uncertainties introduced by the fibre collisions at those scales and effects of the resolution of the simulation. Therefore, we will focus our attention at scales larger than 5 h^{-1} Mpc where the impact of fibre collision is smaller.

In order to study the clustering in different redshift bins using the HAM implemented in this work, we divide the full range into three bins. We select approximately the same number of galaxies in each redshift bin in order to have similar statistics in all of them. We perform an abundance matching (different scatter values that vary from 0.05 to 0.5) for each range to fit the monopole. Fig. 15 shows the monopole and quadrupole for the three different redshift bins. The discrepancy in the quadrupole can be due to one or more of the approaches used in this work. Possible causes of this discrepancy are enumerated below.

(i) Guo et al. (2015c) find similar discrepancies in the quadrupole in configuration space for scales >5 h^{-1} Mpc. They argue that the underestimation of the quadrupole on large scales is possible due to the correlated neighbouring bins in the covariance matrix. They obtain a reasonable χ^2, even with this feature of the predicted quadrupole.

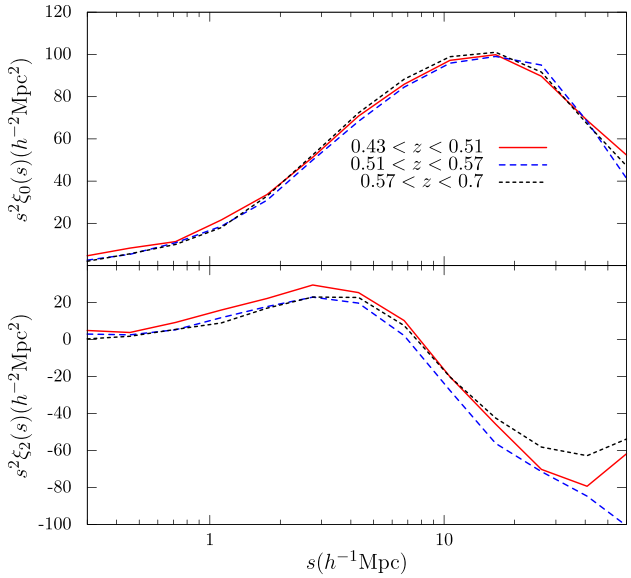


Figure 16. Correlation function for the CMASS sample in three redshift bins. Top panel: monopole with small variations in time. Bottom panel: the quadrupole for the selected ranges. In contrast with the monopole, the quadrupole shows larger variations for the different redshifts.

(ii) Montero-Dorta et al. (2014) show that the intermediate-redshift bin ($0.51 < z < 0.57$) is the most complete region in the CMASS sample. The standard HAM can reproduce monopole and quadrupole for this redshift bin (see Fig. 15), but cannot reproduce the quadrupole for the other two bins. The CMASS DR12 sample has small variations in the monopole. However, quadrupole changes and it becomes similar for the two redshift ranges where the incompleteness of the sample is larger (Fig. 16).

(iii) The values of scatter used to fit the monopole of the correlation function in the different redshift bins vary in a wide range. This can be due to the evolution of the number density in the CMASS sample and some approximations used in this work. Leauthaud et al. (2016) show a non-negligible evolution of the SMF at low redshift compared with the complete redshift range (0.43–0.7). Our

approximation of non-evolving SMF could overestimate the incompleteness in the low-redshift range (Fig. 5, left-hand panel), and then the necessary scatter to reproduce the observed correlation function will be smaller. We also assume a constant mean scatter, but indeed scatter depends on the stellar mass; it increases with the mass of the galaxies (Trujillo-Gomez et al. 2011; Reddick et al. 2013). This dependence can explain why the scatter needed to reproduce the clustering of the low-redshift range is smaller than the one used in the intermediate redshift. At low redshift, the number density is equal to $3.466 \times 10^{-4}\ h^3\ \mathrm{Mpc}^{-3}$, which is smaller than $3.942 \times 10^{-4}\ h^3\ \mathrm{Mpc}^{-3}$ for the middle redshift. If both samples were complete, we will expect a larger scatter in the first range. However, due to the large incompleteness in the high-mass end at low redshift, the mean mass of this sample is $1.86 \times 10^{11}\ \mathrm{M}_{\odot}$ compared to $2.04 \times 10^{11}\ \mathrm{M}_{\odot}$ for the second redshift range. For this reason, the scatter needed to reproduce the clustering is smaller in the low-redshift range. In the high-redshift bin, we can only see very massive galaxies (see Fig. 5, right-hand panel) compared to the whole population of galaxies in the CMASS sample. This range is complete in the high-mass end and, compared to the other two redshift ranges, has a number density very small ($1.534 \times 10^{-4}\ h^3\ \mathrm{Mpc}^{-3}$), which implies larger mean mass ($2.63 \times 10^{11}\ \mathrm{M}_{\odot}$) and scatter than for the other samples.

(iv) We have added a simple model for the stellar mass incompleteness in the CMASS sample. However, there can be other effects of the incompleteness in the target selection that cannot be modelled in this simple way. Although the selection is performed to select LRG, an incomplete blue cloud is in the sample and its fraction compared to the red sequence evolves with redshift (e.g. Guo et al. 2013; Montero-Dorta et al. 2014). Those two populations can live in different kinds of haloes, and therefore they should be described by different scatter values. The errors introduced by this effect can increase with redshift, because the fraction of blue galaxies increases as well. As opposed to the low-redshift bin, the high-redshift bin is complete in the high-mass end (Fig. 5, $z = 0.65$), but the fraction of blue galaxies is larger than the middle bin, which can affect the prediction of the quadrupole. The presence of a small fraction of the so-called 'green valley' can also introduce small errors in our modelling.

(v) The number density in the high-redshift bin ($0.57 < z < 0.70$) is very small compared to the middle redshift range. In this region, the fraction of small galaxies decreases and the impact of the most massive objects in the clustering becomes stronger. Guo et al. (2015a) and Saito et al. (2015) need modification of the HAM model when colour cuts are applied. In addition, Guo et al. (2015c) show the necessity to introduce a velocity bias in the HOD to reproduce the most massive galaxies. If the standard HAM does not describe the clustering of the most massive galaxies, HAM mocks, which model samples as the CMASS in the redshift range $0.57 < z < 0.70$, will not reproduce accurately the clustering of the observed data.

(vi) In addition, recent papers report results for LRG samples where the number of significant miscentral galaxies in haloes is larger than expected (e.g. Hoshino et al. 2015) or the presence of off-centring for central galaxies (e.g. Hikage et al. 2013). The implementation of these results in the construction of mocks reproducing LRG samples could also modify the quadrupole.

7 SUMMARY

We investigated the galaxy clustering of the BOSS CMASS DR12 sample using light cones constructed from the BigMDPL simulation. We perform a HAM to populate the dark matter haloes with galaxies using the Portsmouth DR12 stellar mass catalogue. In addition, the stellar mass distribution is modelled to take into account the incompleteness in stellar mass of the CMASS sample. Our study included features such as the survey geometry, veto masks and fibre collision. The combination of HAM and the BigMDPL simulation provides results in good agreement with the observed data. Our results show that the HAM is a method extremely useful in the study of the relation between dark matter haloes and galaxies, and can be very helpful in the production of mock catalogues (Kitaura et al. 2016, companion paper).

Our main results can be summarized as follows.

(i) We model the observed monopole in configuration space using HAM. Assuming a complete sample, the scatter parameter is very large compared to previous studies. The modelling of stellar mass incompleteness significantly decreases the value of scatter to $\sigma_{\mathrm{HAM}}(V_{\mathrm{peak}}|M_*) = 0.31$. Our model reproduces the observed monopole for nearly every scale.

(ii) The prediction of the quadrupole in configuration space appears to be in disagreement with the observed data. We present possible explanations of this disagreement. In future works, we will concentrate on reducing the possible systematics, in order to understand better the limits of our model.

(iii) We compute the projected correlation function and the 3PCF, finding good agreement between the model and the observed data within 1σ errors for most of the scales. For scales ~ 0 and $\sim \pi$, the differences are of the order of 2σ errors, which can be related to the same factors of the disagreement in the quadrupole. The monopole in k-space of the BigMD-BOSS light cone is in remarkable agreement with the measurement from the CMASS sample corrected by fibre collisions ($\sim$10 per cent of difference at $k = 0.9$). The same agreement is found when we use nearest-neighbour weights, which shows that the assignment of fibre collision in the light cone can reproduce the observed data.

(iv) We compare our prediction of the stellar-to-halo mass relation with lensing measurements. The results are in good agreement with the observed data. Our assumption of a constant scatter is reflected in the differences with observations. Lensing measurements suggest the need to include the stellar mass dependence in the scatter of the HAM.

The BigMD-BOSS light cone is publicly available. It can be found in the SDSS SkyServer.[3] The current version includes angular coordinates (RA, Dec.), redshift in real space and redshift space, peculiar velocity in the line of sight, M_{200}, V_{peak} and M_*. Properties of galaxies such as effective radius (R_{eff}), velocity dispersion (σ_{v}) and mass-to-light ratio (M/L) will be included in future updates.

ACKNOWLEDGEMENTS

SRT is grateful for support from the Campus de Excelencia Internacional UAM/CSIC. SRT also thanks Fernando Campos del Pozo for useful discussions and help while developing the SUGAR code.

The BigMultiDark simulations have been performed on the SuperMUC supercomputer at the Leibniz-Rechenzentrum (LRZ) in Munich, using the computing resources awarded to the PRACE project number 2012060963. The authors want to thank V. Springel for providing them with the optimized version of GADGET-2.

[3] http://skyserver.sdss.org/dr12/en/home.aspx

SRT, CC, FP, AK, FSK, GF and SG acknowledge support from the Spanish MICINNs Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064, MINECO Centro de Excelencia Severo Ochoa Programme under grant SEV-2012-0249 and grant AYA2014-60641-C2-1-P. GY acknowledges support from MINECO (Spain) under research grants AYA2012-31101 and FPA2012-34694 and Consolider Ingenio SyeC CSD2007-0050. FP wishes to thank the Lawrence Berkeley National Laboratory for the hospitality during the development of this work. FP also acknowledges the Spanish MEC 'Salvador de Madariaga' programme, Ref. PRX14/00444.

CH also wants to thank the Instituto de Física Teórica UAM/CSIC for the hospitality during his summer visit, where part of this work was completed. GF acknowledges financial support from the Ministerio de Educación y Ciencia of the Spanish Government through FPI grant AYA2010-2131-C02-01. FSK acknowledges the support of the Karl-Schwarzschild Program from the Leibniz Society.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation and the US Department of Energy Office of Science. The SDSS-III website is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington and Yale University.

REFERENCES

Abazajian K. et al., 2004, AJ, 128, 502
Aihara H. et al., 2011, ApJS, 193, 29
Alam S. et al., 2015, ApJS, 219, 12
Anderson L. et al., 2014, MNRAS, 441, 24
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013a, ApJ, 762, 109
Behroozi P. S., Wechsler R. H., Wu H.-Y., Busha M. T., Klypin A. A., Primack J. R., 2013b, ApJ, 763, 18
Behroozi P. S., Wechsler R. H., Conroy C., 2013c, ApJ, 770, 57
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Bernardeau F., 1994, ApJ, 427, 51
Blaizot J., Wadadekar Y., Guiderdoni B., Colombi S. T., Bertin E., Bouchet F. R., Devriendt J. E. G., Hatton S., 2005, MNRAS, 360, 159
Bolton A. S. et al., 2012, AJ, 144, 144
Bouchet F. R., Colombi S., Hivon E., Juszkiewicz R., 1995, A&A, 296, 575
Buchert T., 1994, MNRAS, 267, 811
Catelan P., 1995, MNRAS, 276, 115
Chuang C.-H. et al., 2013, preprint (arXiv:1312.4889)
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Davis M., Peebles P. J. E., 1983, ApJ, 267, 465
Dawson K. S. et al., 2013, AJ, 145, 10
Eisenstein D. J. et al., 2011, AJ, 142, 72
Favole G. et al., 2015a, preprint (arXiv:1507.04356)
Favole G., McBride C. K., Eisenstein D. J., Prada F., Swanson M. E., Chuang C.-H., Schneider D. P., 2015b, preprint (arXiv:1506.02044)
Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23
Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748
Gunn J. E. et al., 1998, AJ, 116, 3040
Gunn J. E. et al., 2006, AJ, 131, 2332
Guo Q., White S., Li C., Boylan-Kolchin M., 2010, MNRAS, 404, 1111
Guo H., Zehavi I., Zheng Z., 2012, ApJ, 756, 127
Guo H. et al., 2013, ApJ, 767, 122
Guo H. et al., 2014, MNRAS, 441, 2398
Guo H. et al., 2015a, MNRAS, 449, L95
Guo H. et al., 2015b, MNRAS, 449, L95
Guo H. et al., 2015c, MNRAS, 453, 4368
Hamilton A. J. S., 1992, ApJ, 385, L5
Hamilton A. J. S., 1993, ApJ, 417, 19
Hikage C., Mandelbaum R., Takada M., Spergel D. N., 2013, MNRAS, 435, 2345
Hoshino H. et al., 2015, MNRAS, 452, 998
Huterer D., Cunha C. E., Fang W., 2013, MNRAS, 432, 2945
Jing Y. P., Mo H. J., Börner G., 1998, ApJ, 494, 1
Kaiser N., 1987, MNRAS, 227, 1
Kitaura F.-S., Heß S., 2013, MNRAS, 435, L78
Kitaura F.-S., Yepes G., Prada F., 2014, MNRAS, 439, L21
Kitaura F.-S. et al., 2016, MNRAS, 456, 4156
Kitzbichler M. G., White S. D. M., 2007, MNRAS, 376, 2
Klypin A., Yepes G., Gottlober S., Prada F., Heß S., 2016, MNRAS, 457, 4340
Klypin A., Prada F., Yepes G., Heß S., Gottlöber S., 2015, MNRAS, 447, 3693
Knebe A. et al., 2015, MNRAS, 451, 4029
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Kroupa P., 2001, MNRAS, 322, 231
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Leauthaud A. et al., 2012, ApJ, 744, 159
Leauthaud A. et al., 2016, MNRAS, 457, 4021
Manera M. et al., 2013, MNRAS, 428, 1036
Maraston C. et al., 2013, MNRAS, 435, 2764
Mohayaee R., Mathis H., Colombi S., Silk J., 2006, MNRAS, 365, 939
Montero-Dorta A. D. et al., 2014, preprint (arXiv:1410.5854)
Moustakas J. et al., 2013, ApJ, 767, 50
Neyrinck M. C., 2013, MNRAS, 428, 141
Nuza S. E. et al., 2013, MNRAS, 432, 743
Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
Press W. H., Schechter P., 1974, ApJ, 187, 425
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Reid B. et al., 2016, MNRAS, 455, 1553
Ross A. J. et al., 2012, MNRAS, 424, 564
Saito S. et al., 2015, preprint (arXiv:1509.00482)
Schaye J. et al., 2015, MNRAS, 446, 521
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Shan H. et al., 2015, preprint, (arXiv:1502.00313)
Shankar F. et al., 2014, ApJ, 797, L27
Shu Y., Bolton A. S., Schlegel D. J., Dawson K. S., Wake D. A., Brownstein J. R., Brinkmann J., Weaver B. A., 2012, AJ, 143, 90
Smee S. A. et al., 2013, AJ, 146, 32
Springel V., 2005, MNRAS, 364, 1105
Stoughton C. et al., 2002, in Tyson J. A., Wolff S., eds, Proc. SPIE Conf. Ser. Vol. 4836, Survey and Other Telescope Technologies and Discoveries. SPIE, Bellingham, p. 339
Szapudi I., Szalay A. S., 1998, ApJ, 494, L41
Trujillo-Gomez S., Klypin A., Primack J., Romanowsky A. J., 2011, ApJ, 742, 16
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
White M. et al., 2011, ApJ, 728, 126
York D. G. et al., 2000, AJ, 120, 1579
Zheng Z. et al., 2005, ApJ, 633, 791
Zu Y., Mandelbaum R., 2015, MNRAS, 454, 1161

[1]*Instituto de Física Teórica, (UAM/CSIC), Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*
[2]*Campus of International Excellence UAM+CSIC, Cantoblanco, E-28049 Madrid, Spain*
[3]*Departamento de Física Teórica M8, Universidad Autonoma de Madrid (UAM), Cantoblanco, E-28049 Madrid, Spain*
[4]*Leibniz-Institut für Astrophysik Potsdam (AIP), D-14482 Potsdam, Germany*
[5]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[6]*Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, E-18080 Granada, Spain*
[7]*Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 20030, China*
[8]*Department of Physics and Astronomy, University of Utah, 115 South 1400 East, Salt Lake City, UT 84112, USA*
[9]*Astronomy Department, New Mexico State University, Las Cruces, NM 88003, USA*
[10]*Severo Ochoa Associate Researcher at the Instituto de Física Teórica (UAM/CSIC), E-28049 Madrid, Spain*
[11]*Space Telescope Science Institute, Baltimore, MD 21218, USA*
[12]*Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003, USA*
[13]*Institute of Cosmology & Gravitation, University of Portsmouth, Dennis Sciama Building, Portsmouth PO1 3FX, UK*
[14]*Center for Astrophysics, Harvard University, 60 Garden Street, Cambridge, MA 02138, USA*
[15]*Department of Physics, Kansas State University, 116 Cardwell Hall, Manhattan, KS 66506, USA*
[16]*National Abastumani Astrophysical Observatory, Ilia State University, 2A Kazbegi Ave., GE-1060 Tbilisi, Georgia*
[17]*Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA*
[18]*Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA*

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

CHAPTER 4

MultiDark-Patchy mocks for the BOSS-DR12

Publication: Monthly Notices of the Royal Astronomical Society, Volume 456, Issue 4, p.4156-4173

Motivation

Unlike other branches of Physics, Cosmology and Astrophysics cannot have in-situ different realisations of their experiments. We consider our Universe as a huge experiment of the physical processes that govern its mass-energy content. This raises a problem in the analysis of the data coming from observations, because there is no simple way to estimate uncertainties from the data, although there are some methods that use subvolumes of the observed data to have an estimation of errors. However, they do not take into account all the relevant errors in the data set. At this point, cosmological simulations take an important place in these measurements. Theoretical models can reproduce the growth of structures with good agreement, so we can assume these simulations are possible realisations of the Universe. So we combine different ingredients to produce the most realistic mock catalogues thus improving the covariance matrices estimation.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **456,** 4156–4173 (2016) doi:10.1093/mnras/stv2826

The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: mock galaxy catalogues for the BOSS Final Data Release

Francisco-Shu Kitaura,[1⋆†] Sergio Rodríguez-Torres,[2,3,4‡] Chia-Hsun Chuang,[2§] Cheng Zhao,[5] Francisco Prada,[2,3,6] Héctor Gil-Marín,[7] Hong Guo,[8,9] Gustavo Yepes,[4] Anatoly Klypin,[10,11] Claudia G. Scóccola,[2,12,13] Jeremy Tinker,[14] Cameron McBride,[15] Beth Reid,[16,17] Ariel G. Sánchez,[18] Salvador Salazar-Albornoz,[18,19] Jan Niklas Grieb,[18,19] Mariana Vargas-Magana,[20] Antonio J. Cuesta,[21] Mark Neyrinck,[22] Florian Beutler,[16] Johan Comparat,[2] Will J. Percival[7] and Ashley Ross[7,23]

Affiliations are listed at the end of the paper

Accepted 2015 November 27. Received 2015 November 20; in original form 2015 September 19

ABSTRACT

We reproduce the galaxy clustering catalogue from the SDSS-III Baryon Oscillation Spectroscopic Survey Final Data Release (BOSS DR11&DR12) with high fidelity on all relevant scales in order to allow a robust analysis of baryon acoustic oscillations and redshift space distortions. We have generated (6000) 12 288 MultiDark PATCHY BOSS (DR11) DR12 light cones corresponding to an effective volume of $\sim$192 000 $[h^{-1}\,\mathrm{Gpc}]^3$ (the largest ever simulated volume), including cosmic evolution in the redshift range from 0.15 to 0.75. The mocks have been calibrated using a reference galaxy catalogue based on the halo abundance matching modelling of the BOSS DR11&DR12 galaxy clustering data and on the data themselves. The production follows three steps. First, we apply the PATCHY code to generate a dark matter field and an object distribution including non-linear stochastic galaxy bias. Secondly, we run the halo/stellar distribution reconstruction HADRON code to assign masses to the various objects. This step uses the mass distribution as a function of local density and non-local indicators (i.e. tidal field tensor eigenvalues and relative halo exclusion separation for massive objects) from the reference simulation applied to the corresponding patchy dark matter and galaxy distribution. Finally, we apply the SUGAR code to build the light cones. The resulting MultiDarkPATCHY mock light cones reproduce the number density, selection function, survey geometry, and in general within 1σ, for arbitrary stellar mass bins, the power spectrum up to $k = 0.3\,h\,\mathrm{Mpc}^{-1}$, the two-point correlation functions down to a few Mpc scales, and the three-point statistics of the BOSS DR11&DR12 galaxy samples.

Key words: methods: numerical – galaxies: haloes – galaxies: statistics – large-scale structure of Universe.

1 INTRODUCTION

The observable Universe represents a unique realization of an underlying physical cosmological process. Large galaxy redshift surveys like the Baryon Oscillation Spectroscopic Survey (BOSS; e.g. Bolton et al. 2012; Dawson et al. 2013; Alam et al. 2015), a branch of the ongoing Sloan Digital Sky Survey (SDSS-III; Eisenstein et al. 2011), scan the sky with unprecedented accuracy trying to unveil structure formation in an expanding Universe. One important question arises in the analysis of the data provided by such surveys: if the Universe is comparable to a huge unique experiment, how can we determine the uncertainties in the measurement of quantities derived from observing it? One strategy consists of dividing the observations into subvolumes, treating each of the subsamples

⋆E-mail: kitaura@aip.de
†Karl-Schwarzschild-fellow.
‡Campus de Excelencia Internacional UAM/CSIC Fellow.
§MultiDark Fellow.

as independent measurements, and computing the errors with jackknife or bootstrap estimates. While this approach continues being relevant as a way to obtain error estimates directly from the data (see e.g. Norberg et al. 2009), it also implies several disadvantages. First, it does not include systematic errors present in all subvolumes, secondly it does not lead to a physical understanding of the data by itself, and thirdly it introduces variance beyond the one already present in the data on scales larger than the subvolumes. The last point is especially critical when the signal sought has a large characteristic scale and its detection significance crucially depends on the volume, as is the case for baryon acoustic oscillations (BAOs; see e.g. Seo & Eisenstein 2005; White, Song & Percival 2009). During the past decades, there has been a huge effort to encode our physical knowledge of structure formation in computational algorithms, and compare the theoretical models to the actual observations. Pioneering works started with qualitative comparisons (see e.g. Klypin & Shandarin 1983; Blumenthal et al. 1984; Davis et al. 1985). Since then simulations have grown and such comparisons have turned increasingly more quantitative (see e.g. Klypin et al. 2003; Springel et al. 2005; Boylan-Kolchin et al. 2009; Klypin, Trujillo-Gomez & Primack 2011). These efforts are essential to understand structure formation and yet they suffer from a strong limitation: as simulations always push the computational limits, they are not suited for massive production. In fact, the number of current state-of-the-art large-volume *N*-body simulations is of order 10 (Kim et al. 2009; Alimi et al. 2012; Angulo et al. 2012; Prada et al. 2012; Fosalba et al. 2015; Ishiyama et al. 2015; Klypin et al. 2014; Skillman et al. 2014; Watson et al. 2014). However, an ideal approach to determine the uncertainties from current and upcoming surveys scanning large sky areas, and hence covering huge volumes, such as BOSS[1] (White et al. 2011), DESI[2]/BigBOSS (Schlegel et al. 2011), DES[3] (Frieman & Dark Energy Survey Collaboration 2013), LSST[4] (LSST Dark Energy Science Collaboration 2012), J-PAS[5] (Benitez et al. 2014), 4MOST[6] (de Jong et al. 2012), or *Euclid*[7] (Cimatti et al. 2009; Laureijs 2009), requires thousands of such simulations if the simplest error determination methods are used (Dodelson & Schneider 2013; Taylor, Joachimi & Kitching 2013; Percival et al. 2014). Alternative more efficient methods need to be considered to face this challenge. A few pioneering works explored a viable strategy more than a decade ago relying on simplified fast gravity solvers using perturbation theory (PT): PINOCCHIO (Monaco et al. 2002, 2013) and PTHALOS (Scoccimarro & Sheth 2002). Nevertheless, these methods are not trivial, need calibration with *N*-body simulations, and still demand high computational efforts. For this reason, some of the first analysis of large surveys (Percival et al. 2001; Cole et al. 2005) was done based on lognormal realizations (see also Percival, Verde & Peacock 2004; Beutler et al. 2011), which match the two-point statistics by construction (Coles & Jones 1991), although their three-point statistics is very different from the true one (see e.g. White, Tinker & McBride 2014; Chuang et al. 2015b). It is also not clear that their four-point statistics will be accurate (Cooray & Hu 2001; Takada & Hu 2013).

[1] http://www.sdss3.org/surveys/boss.php
[2] http://desi.lbl.gov/
[3] http://www.darkenergysurvey.org
[4] http://www.lsst.org/lsst/
[5] http://j-pas.org/
[6] http://www.aip.de/en/research/research-area-ea/research-groups-and-projects/4most
[7] http://www.euclid-ec.org

The analysis of past data releases of the BOSS collaboration utilized 1000 mocks, created based on an improved version of PTHALOS (Manera et al. 2013, 2015). The use of approximate gravity solvers in these methods came at the expense of only matching clustering statistics on a wide range of scales to ∼10 per cent precision (and strongly deviating towards small scales $\lesssim 20\,h^{-1}$ Mpc).

This sets the agenda for the current BOSS data release DR11&DR12 and the requirements for a new generation of mock galaxy catalogues. Ideally, one would like to base these on efficient solvers that are trained on exact solutions and deliver a comparable precision. A new generation of methods that can meet these high requirements have been developed during the past two years, in particular, PATCHY (Kitaura, Yepes & Prada 2014), QPM (White et al. 2014), and EZMOCKS (Chuang et al. 2015a). The key concept exploited by these methods is to rely only on the large-scale density field obtained from approximate gravity solvers and use biasing prescriptions to populate it with mock galaxies, in a similar way to the methods proposed to augment the resolution of *N*-body simulations (de la Torre & Peacock 2013; de la Torre et al. 2013; Angulo et al. 2014; Ahn et al. 2015). One should however be careful, as computing an accurate dark matter field is a necessary, but not sufficient condition to reproduce the correct halo/galaxy three-point statistics. The bias parameters are degenerate in the two-point statistics and need to be additionally constrained to reproduce higher order statistics (Kitaura et al. 2015). We will rely in this work on the PATCHY method due to its verified accuracy in the two- and three-point statistics for different populations of objects (see application of the HADRON code to PATCHY and EZMOCK; Zhao et al. 2015). An additional set of galaxy mocks fitting the BOSS DR11&DR12 (CMASS and LOWZ) data at two mean redshifts (respectively) based on QPM have been produced in an unprecedented effort. These are constructed with a different structure formation model based on low-resolution particle mesh solvers, and a different galaxy bias, based on a rank-ordering scheme assigning most massive objects to the highest density peaks (for a comparison of both sets of catalogues, see Section 3 and Gil-Marín et al. 2015a).

Another approach uses approximate PT-based solutions to speed up *N*-body solvers (see COLA method; Tassev, Zaldarriaga & Eisenstein 2013; Howlett, Manera & Percival 2015; Koda et al. 2015). This method is very promising to generate ensembles of reference mock catalogues; however, it has the drawback of requiring large computational memory for the force calculation and large number of particles to resolve the haloes (see Chuang et al. 2015b), and is therefore not suitable for the massive production aimed in this work. The speed of the method over *N*-body simulations comes at the expense of not resolving the substructures required to produce a realistic galaxy catalogue. This problem can be circumvented by, e.g., augmenting the missing objects with the halo occupation distribution (HOD) model, hereby losing some of the advantage of having a higher precise description of the non-linear clustering over the above-mentioned methods which rely only on the large-scale dark matter field, as shown in a comparison study (see Chuang et al. 2015b, and references therein). One may need an approach like COLA, to model the large-scale structure, combined with the galaxy bias presented in this work for future emission line galaxy-based surveys. We will, however, demonstrate here that this is not necessary to model the distribution of luminous red galaxies (LRGs) aimed in this work.

One could argue whether mock catalogues are required at all, as analytical models may deliver an almost direct computation of error bars and covariance matrices (Hartlap, Simon & Schneider 2007; Hamaus et al. 2010; Dodelson & Schneider 2013; Taylor et al. 2013;

Kalus, Percival & Samushia 2016). It still remains to be shown that these methods making simple assumptions, such as that the density field is Gaussian distributed, yield the same accuracy as covariance matrices based on large sets of mock catalogues.

Nevertheless, the purpose of mock catalogues is manifold, as they not only serve to provide error estimates, but also to provide an understanding of the systematics of the survey and of the methodology. Any analytical prediction or data analysis method should be cross-checked with large ensembles of mock galaxy catalogues for which the products of this work could be useful. One clear example is the case of BAO reconstruction techniques (see e.g. Eisenstein et al. 2007; Padmanabhan et al. 2012; Anderson et al. 2014; Ross et al. 2015).

We exploit the efficiency and accuracy of the PATCHY code to produce 12 288 galaxy mock catalogues[8] including the light-cone evolution of galaxy bias based on the halo abundance matching (HAM) technique applied to the reference BigMultiDark *N*-body simulation (see Rodríguez-Torres et al. 2015, companion paper), and to the peculiar motions based on the observational data, matching the two-, three-point statistics, in real and redshift space of the BOSS DR11&DR12 galaxy clustering data at different redshifts and for arbitrary stellar mass bins. Special care has been taken to include all relevant observational effects including selection functions and masking. The MultiDark PATCHY BOSS DR11 mock catalogues presented in this work are publicly available.[9]

This paper is structured as follows: in Section 2 we describe the methodology. This section starts with the generation of the reference catalogue using *N*-body simulations and the HAM technique. Subsequently, the scheme to massively generate mock catalogues is described. Then we show in Section 3 the statistical comparison between the mock catalogues and the BOSS DR12 data. Subsequently, we discuss future work (Section 4). Finally, in Section 5 we present the conclusions. The reader interested only in the results may skip Section 2 and directly go to Section 3.

2 METHODOLOGY

To construct high-fidelity mock light cones for interpreting the BOSS DR11&DR12 galaxy clustering, we adopt an iterative training procedure in which a reference catalogue is statistically reproduced with approximate gravity solvers and analytical–statistical biasing models. The whole algorithm involves several steps and is summarized in the flow chart in Fig. 1.

(i) The first step consists of the generation of an accurate reference catalogue. Here we rely on a large *N*-body simulation capable of resolving distinct haloes and the corresponding substructures. This permits us to apply the HAM technique to reproduce the clustering of the observations with only one parameter: the scatter in the stellar mass-to-halo mass relation (see Rodríguez-Torres et al. 2015, companion paper; and Section 2.1). This technique is applied at different redshift bins to obtain a detailed galaxy bias evolution spanning the redshift range covered by BOSS DR11&DR12 galaxies. In this way, we obtain mock galaxy catalogues in full cubical volumes of 2.5 h^{-1} Gpc side at different redshifts.

(ii) In the second step, we train the PATCHY code (Kitaura et al. 2014, 2015) to match the two- and three-point clustering of the full mock galaxy catalogues for each redshift bin. Here we consider all the mock galaxies together in a single bin irrespectively of their stellar mass.

(iii) In the third step, we apply the HADRON code (Zhao et al. 2015) to assign stellar masses to the individual objects.

(iv) In the fourth step, we apply the SUGAR code (see Rodríguez-Torres et al. 2015, companion paper) which includes selection effects, masking, and combines different boxes at different redshifts into a light cone.

(v) In the fifth step, the resulting MultiDark PATCHY mock catalogues are compared to the observations. The process is iterated until the desired accuracy for different statistical measures is reached.

In the next sections, we will describe in detail these steps described above for the massive generation of accurate mock galaxy catalogues. The reader interested only in the results may directly go to Section 3.

2.1 Reference mock catalogues

The reference catalogues are extracted from one of the BigMultiDark simulations[10] (Klypin et al. 2014), which was performed using GADGET-2 (Springel et al. 2005) with 3840^3 particles on a volume of $(2.5\,h^{-1}\,\mathrm{Mpc}\,)^3$ assuming Λ cold dark matter Planck cosmology with $\{\,\Omega_\mathrm{M} = 0.307\,115, \Omega_\mathrm{b} = 0.048\,206, \sigma_8 = 0.8288, n_\mathrm{s} = 0.9611\}$, and a Hubble constant ($H_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$) given by $h = 0.6777$. Haloes were defined based on the Bound Density Maximum halo finder (Klypin & Holtzman 1997).

We rely here on the HAM technique to connect haloes to galaxies (Kravtsov et al. 2004; Neyrinck, Hamilton & Gnedin 2004; Tasitsiomi et al. 2004; Vale & Ostriker 2004; Conroy, Wechsler & Kravtsov 2006; Kim, Park & Choi 2008; Guo et al. 2010; Wetzel & White 2010; Trujillo-Gomez et al. 2011; Nuza et al. 2013).

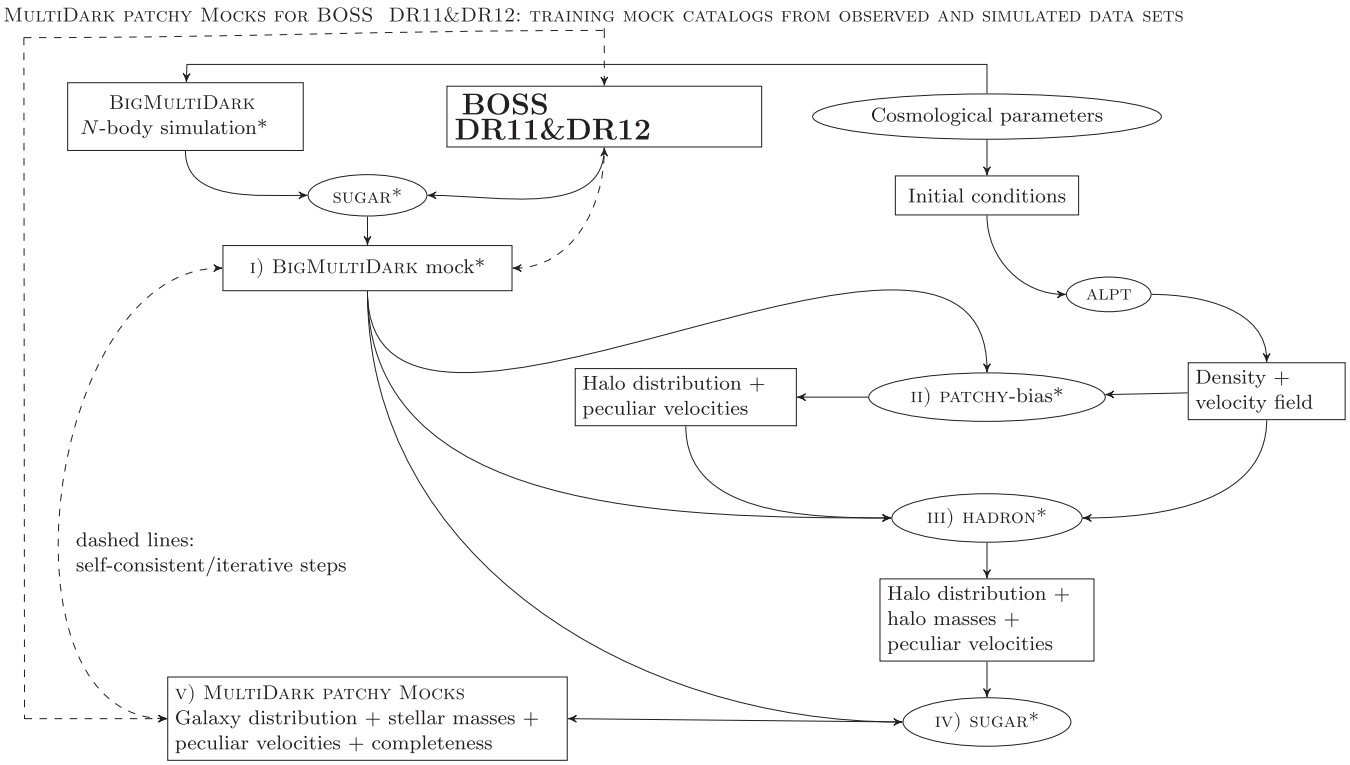
We note that there are alternative methods connecting haloes to galaxies like the HOD model, which we are not going to consider here (e.g. Berlind & Weinberg 2002; Kravtsov et al. 2004; Zehavi et al. 2005; Zentner et al. 2005; Zheng, Coil & Zehavi 2007; Ross & Brunner 2009; Skibba & Sheth 2009; Zheng et al. 2009; White et al. 2011). These methods are based on a statistical relation describing the probability that a halo of virial mass M hosts N galaxies with some specified properties. In general, theoretical HODs require the fitting of a function with several parameters, which we want to avoid here.

At first order HAM assumes a one-to-one correspondence between the luminosity and stellar or dynamical masses: galaxies with more stars are assigned to more massive haloes or subhaloes. The luminosity in a red band is sometimes used instead of stellar mass. There should be some degree of stochasticity in the relation between stellar and dynamical masses due to deviations in the merger history, angular momentum, halo concentration, and even observational errors (Tasitsiomi et al. 2004; Behroozi, Conroy & Wechsler 2010; Leauthaud et al. 2011; Trujillo-Gomez et al. 2011). Therefore, we include a scatter in that relation necessary to accurately fit the clustering of the BOSS data (see Rodríguez-Torres et al. 2015, companion paper). To do this, we modify the maximum

[8] This corresponds to an effective volume of $\sim$192 000 $[h^{-1}\,\mathrm{Gpc}]^3$, a factor of $\sim$20 times larger than the volume of the DEUS FUR simulation (Alimi et al. 2012), and a factor of $\sim$375 times larger than the DarkSky ds14 simulation (Skillman et al. 2014).

[9] http://data.sdss3.org/sas/dr11/boss/lss/dr11_patchy_mocks/ The BOSS DR12 mock catalogues will be made publicly available together with the data catalogue: http://data.sdss3.org/sas/dr12/boss/lss/dr12_patchy_mocks/.

[10] http://www.multidark.org/MultiDark/



*: the asterisc indicates the steps in which calibration with observations and simulations is required

ALPT: Augmented Lagrangian Perturbation Theory: Kitaura & Heß (2013)

BigMultiDark: N-body Planck simulation $(2.5\ h^{-1}\ \mathrm{Gpc})^3$ with $3{,}840^3$ particles: Klypin et al. (2014)

HADRON: Halo mAss Distribution ReconstructiON: Zhao et al. (2015)

HAM: Halo Abundance Matching, Rodríguez-Torres et al. (2015, companion paper)

MockFactory: White et al. (2014)

PATCHY: PerturbAtion Theory Catalog generator of Halo and galaxY distributions: Kitaura et al. (2014, 2015)

SUGAR: SUrvey GenerAtoR, Rodríguez-Torres et al. (see 2015, companion paper)
this code contains HAM and MockFactory

Figure 1. Flowchart of the methodology applied in this work for the generation of high-fidelity BOSS DR11&DR12 mock galaxy catalogues: i) starting from a reference mock catalogue calibrated with the observations, ii) followed by the reproduction of the whole catalogue, iii) with the subsequent mass assignment, iv) and survey generation. v) The final catalogues are compared with the observations and the simulation, and the previous steps are repeated until the mock catalogues are compatible with the observations within 1σ for the monopole and quadrupole up to $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$.

circular velocity ($V_{\rm max}$) of each object adding a Gaussian noise: $V_{\rm max}^{\rm scat} = V_{\rm max}(1 + \mathcal{N}(0, \sigma))$, where $\mathcal{N}(0, \sigma)$ is a Gaussian random number with mean 0 and standard deviation σ. Then, we sort all objects by $V_{\rm max}^{\rm scat}$, and then we selected objects starting from the one with larger $V_{\rm max}^{\rm scat}$ and we continue until we get the proper number density at different redshifts bins.

By construction, the method reproduces the observed luminosity function (or stellar mass function). It also reproduces the scale dependence of galaxy clustering over a large range of epochs (Conroy et al. 2006; Guo et al. 2010). When abundance matching is used for the observed stellar mass function (Li & White 2009), it gives also a reasonable fit to lensing results (Mandelbaum et al. 2006) and to the relation between stellar and virial mass (Guo et al. 2010).

2.2 Generation of mock galaxy catalogues

All covariance matrix estimates based on a finite number of mock catalogues, $N_{\rm s}$, are affected by noise, which must be propagated into the final constraints. The impact of the uncertainties in the covariance matrix on the derived cosmological constraints has been subject of several recent analyses (Dodelson & Schneider 2013; Taylor et al. 2013; Percival et al. 2014). In particular, Dodelson & Schneider (2013) showed that this additional uncertainty can be described by a rescaling of the parameter covariances derived from the distribution of measurements from a set of mocks with a factor given by

$$m = 1 + \frac{(N_{\rm s} - N_{\rm b} - 2)\left(N_{\rm b} - N_{\rm p}\right)}{(N_{\rm s} - N_{\rm b} - 1)(N_{\rm s} - N_{\rm b} - 4)}, \quad (1)$$

where $N_{\rm b}$ is the number of bins in the corresponding clustering measurements and $N_{\rm p}$ is the number of parameters measured. This implies that a large number of mock catalogues are necessary for a robust analysis of the galaxy clustering data.

For the anisotropic BAO measurements of Cuesta et al. (2015), the estimation of the full covariance matrix of the monopole and quadrupole of the two-dimensional correlation function from the ensemble of 1000 QPM corresponds to an additional uncertainty of 2 per cent on the constraints on $H(z)r_{\rm d}$ and $D_{\rm A}(z)/r_{\rm d}$. Using the 2048 MultiDark PATCHY mock catalogues, the effect is reduced to the order

of 1 per cent. Large sets of catalogues are even more important for full-shape fits of anisotropic clustering measurements, where the inclusion of information from smaller scales can significantly improve the constraints based on redshift space distortion (RSD; requiring a larger number of bins). For example, in the analysis of Sánchez et al. (in preparation), based on measurements of the clustering wedges statistic (Kazin, Sánchez & Blanton 2012), the use of mock catalogues corresponds to a rescaling of the parameter covariances by $m = 1.04$ and 1.085 when using 1000 or 2048 catalogues, respectively. This additional uncertainty corresponds to a degradation of the true constraining power of the clustering measurements, which should be minimized by using a larger number of mock catalogues. For this reason, we have made the effort in the BOSS collaboration of producing at least 1000 mocks for each BOSS DR11&DR12 subsample.

The strategy for the massive production of mock galaxy catalogues relies on generating dark matter fields with approximate gravity solvers on a mesh. We use grids of 960^3 cells with volumes of $(2.5\ h^{-1}\ \mathrm{Gpc})^3$ and resolutions of $2.6\,h^{-1}$ Mpc for which the structure formation model can be considered to be accurate (see Section 2.2.1). Then the galaxies are populated on the mesh according to a combined non-linear deterministic (see Section 2.2.2) and stochastic bias model (see Section 2.2.3). In a post-processing step, we assign halo/stellar masses to each object (see Section 2.2.5). Finally, we apply the survey geometry and selection functions (see Section 2.2.6).

Let us start describing the PATCHY code (PerturbAtion Theory Catalog generator of Halo and galaxY distributions).

2.2.1 Approximate fast structure formation model

We rely on augmented Lagrangian perturbation theory (ALPT) to simulate structure formation. Let us recap the basics of this method and refer for details to Kitaura & Heß (2013). In this approximation, the displacement field $\boldsymbol{\Psi}(\boldsymbol{q},z)$, mapping a distribution of dark matter particles at initial Lagrangian positions $\boldsymbol{q}$ to the final Eulerian positions $\boldsymbol{x}(z)$ at redshift z ($\boldsymbol{x}(z)=\boldsymbol{q}+\boldsymbol{\Psi}(\boldsymbol{q},z)$), is split into a long-range $\boldsymbol{\Psi}_{\rm L}(\boldsymbol{q},z)$ and a short-range component $\boldsymbol{\Psi}_{\rm S}(\boldsymbol{q},z)$, i.e. $\boldsymbol{\Psi}(\boldsymbol{q},z)=\boldsymbol{\Psi}_{\rm L}(\boldsymbol{q},z)+\boldsymbol{\Psi}_{\rm S}(\boldsymbol{q},z)$.

We rely on second order LPT (2LPT) for the long-range component $\boldsymbol{\Psi}_{\rm 2LPT}$ (for details on 2LPT, see Buchert 1994; Bouchet et al. 1995; Catelan 1995).

The resulting displacement field is filtered with a kernel $\mathcal{K}$: $\boldsymbol{\Psi}_{\rm L}(\boldsymbol{q},z)=\mathcal{K}(\boldsymbol{q},r_{\rm S})\circ\boldsymbol{\Psi}_{\rm 2LPT}(\boldsymbol{q},z)$. We apply a Gaussian filter $\mathcal{K}(\boldsymbol{q},r_{\rm S})=\exp\left(-|\boldsymbol{q}|^2/(2r_{\rm S}^2)\right)$, with $r_{\rm S}$ being the smoothing radius. We use the spherical collapse approximation to model the short-range component $\boldsymbol{\Psi}_{\rm SC}(\boldsymbol{q},z)$ (see Bernardeau 1994; Mohayaee et al. 2006; Neyrinck 2013). The combined ALPT displacement field

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

is used to move a set of homogeneously distributed particles from Lagrangian initial conditions to the Eulerian final ones. We then grid the particles following a clouds-in-cell scheme to produce a smooth density field $\delta^{\rm ALPT}$. One may get some improvements preventing voids within larger collapsing regions, which essentially extends the collapsing region towards moderate underdensities (see MUSCLE method in Neyrinck 2016). This approach requires about eight additional convolutions being about twice as expensive, as the approached used here. Moreover, we have checked that the improvement provided by including MUSCLE is not perceptible when using grids with cell sizes of $2.6\ h^{-1}$ Mpc.

2.2.2 Deterministic bias relations

In this section, we describe the deterministic part of our bias model. This is combined with a stochastic element, described in Section 2.2.3, and a non-local element, described in Section 2.2.5, to produced the full model. The deterministic bias relates the expected number counts of haloes or galaxies $\rho_{\rm g}\equiv\langle N_{\rm g}\rangle_{\partial V}$ at a given finite volume to the underlying dark matter field $\rho_{\rm M}$, with $\langle[\cdots]\rangle_{\partial V}$ being the ensemble average over the differential volume element ∂V (in our case the cell of a regular mesh). This relation is known to be non-linear, non-local, and stochastic (Press & Schechter 1974; Peacock & Heavens 1985; Bardeen et al. 1986; Fry & Gaztanaga 1993; Mo & White 1996, 2002; Dekel & Lahav 1999; Sheth & Lemson 1999; Seljak 2000; Berlind & Weinberg 2002; Smith, Scoccimarro & Sheth 2007; Desjacques et al. 2010; Beltrán Jiménez & Durrer 2011; Valageas & Nishimichi 2011; Baldauf et al. 2012, 2013; Chan, Scoccimarro & Sheth 2012; Elia, Ludlow & Porciani 2012; Ahn et al. 2015). In general, this bias relation will be arbitrarily complex:

$$\rho_{\rm g}=f_{\rm g}\,B(\rho_{\rm M}), \quad (3)$$

with $B(\rho_{\rm M})$ being a general bias function, $f_{\rm g}=\frac{\langle\rho_{\rm g}\rangle_V}{\langle B(\rho_{\rm M})\rangle_V}$, $\langle\rho_{\rm g}\rangle_V$ being the number density $\bar{N}_{\rm g}$, and $\langle[\cdots]\rangle_V$ being the ensemble average over the whole considered volume V (in our case the volume of the considered mesh).

The deterministic bias model we consider in this work has the following form:

$$\rho_{\rm g}=f_{\rm g}\,\theta(\rho_{\rm M}-\rho_{\rm th})\,\exp\left[-\left(\frac{\rho_{\rm M}}{\rho_\epsilon}\right)^\epsilon\right]\,\rho_{\rm M}^\alpha\,(\rho_{\rm M}-\rho_{\rm th})^\tau, \quad (4)$$

with

$$f_{\rm g}=\bar{N}_{\rm g}/\langle\theta(\rho_{\rm M}-\rho_{\rm th})\,\exp\left[-\left(\frac{\rho_{\rm M}}{\rho_\epsilon}\right)^\epsilon\right]\,\rho_{\rm M}^\alpha\,(\rho_{\rm M}-\rho_{\rm th})^\tau\rangle_V, \quad (5)$$

and $\{\ \rho_{\rm th},\ \alpha,\ \epsilon,\ \rho_\epsilon,\ \tau\}$ the parameters of the model. We have modelled threshold bias (Kaiser 1984; Bardeen et al. 1986; Cole & Kaiser 1989; Sheth, Mo & Tormen 2001; Mo & White 2002) as a combination of a step function $\theta(\rho_{\rm M}-\rho_{\rm th})$ (Kitaura et al. 2014) and an exponential cut-off $\exp\left[-\left(\frac{\rho_{\rm M}}{\rho_\epsilon}\right)^\epsilon\right]$ (Neyrinck et al. 2014). The local bias expansion (Cen & Ostriker 1993; Fry & Gaztanaga 1993) is summarized by a power law (de la Torre & Peacock 2013; Kitaura et al. 2014). In addition, we consider a bias $(\rho_{\rm M}-\rho_{\rm th})^\tau$ which compensates for the missing power of PT-based methods.

Non-local bias has been recently found to be relevant (McDonald & Roy 2009; Baldauf et al. 2012; Chan, Scoccimarro & Sheth 2012; Sheth, Chan & Scoccimarro 2013; Saito et al. 2014). A non-local bias introduces a scatter in the local deterministic bias relations described above. In this work, the scatter is first described by a stochastic bias relation (see Section 2.2.3). We have investigated second-order non-local bias with PATCHY without finding that this can have a relevant effect on the mock catalogues considering stochastic bias and the full (one single mass bin) catalogue (see Autefage et al., in preparation). In fact, once one considers different populations of halo or stellar mass objects, then non-local bias plays an important role. We solve this in a post-processing step when assigning the masses to each galaxy (see Section 2.2.5 and Zhao et al. 2015).

2.2.3 *Stochastic biasing*

The halo distribution is a discrete sample $N_{g,i}$ of the continuous underlying dark matter distribution $\rho_{g,i}$:

$$N_{g,i} \curvearrowleft P(N_{g,i} \mid \rho_{g,i}, \{p_{SB}\}), \tag{6}$$

for each cell i and $\{p_{SB}\}$ being the set of stochastic bias parameters. To account for the shot noise, one could do Poissonian realizations of the halo density field as given by the deterministic bias and the dark matter field (see e.g. de la Torre & Peacock 2013). However, it is known that the excess probability of finding haloes in high-density regions generates overdispersion (Somerville et al. 2001; Casas-Miranda et al. 2002).

The strategy up to now has been to generate a mock catalogue which reproduces the clustering of the whole population of galaxies for a given redshift. This has the advantage that by mixing massive and low-mass galaxies we will always be dominated by overdispersion, which is much easier to model than underdispersion. In particular, we consider the negative binomial probability distribution function (for non-Poissonian distributions, see Saslaw & Hamilton 1984; Sheth 1995) including an additional parameter β to model overdispersion (tends towards the Poisson probability distribution function for $\beta \to \infty$ and for low λ values).

We note that a proper treatment of the deviation from Poissonity is also crucial to get accurate density reconstructions (see Ata, Kitaura & Müller 2015 and Ata et al., in preparation).

We will need, however, to take care of the different statistical nature of each population of galaxies when we assign masses to each object (see Section 2.2.5).

2.2.4 *Redshift space distortions*

Let us recap here the way in which RSDs are treated in the PATCHY code (see Kitaura et al. 2014).

The mapping between Eulerian real space $\boldsymbol{x}(z)$ and redshift space $\boldsymbol{s}(z)$ is given by $\boldsymbol{s}(z) = \boldsymbol{x}(z) + \boldsymbol{v}_r(z)$, with $\boldsymbol{v}_r \equiv (\boldsymbol{v}\cdot\hat{\boldsymbol{r}})\hat{\boldsymbol{r}}/(Ha)$, where $\hat{\boldsymbol{r}}$ is the unit sightline vector, H the Hubble constant, a the scale factor, and $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x})$ the 3D velocity field interpolated at the position of each halo in Eulerian space $\boldsymbol{x}$ using the displacement field $\boldsymbol{\Psi}_{ALPT}(\boldsymbol{q}, z)$. We split the peculiar velocity field into a coherent $\boldsymbol{v}^{coh}$ and a (quasi-) virialized component $\boldsymbol{v}_\sigma$: $\boldsymbol{v} = \boldsymbol{v}^{coh} + \boldsymbol{v}^{\sigma}$. The coherent peculiar velocity field is computed in Lagrangian space from the linear Gaussian field $\delta^{(1)}(\boldsymbol{q})$ using the ALPT formulation consistently with the displacement field (see equation 2):

$$\begin{aligned}\boldsymbol{v}^{coh}_{ALPT}(\boldsymbol{q}, z) = {} & \mathcal{K}(\boldsymbol{q}, r_S) \circ \boldsymbol{v}_{2LPT}(\boldsymbol{q}, z) \\ & + (1 - \mathcal{K}(\boldsymbol{q}, r_S)) \circ \boldsymbol{v}_{SC}(\boldsymbol{q}, z),\end{aligned} \tag{7}$$

with $\boldsymbol{v}_{2LPT}(\boldsymbol{q}, z)$ being the second-order and $\boldsymbol{v}_{SC}(\boldsymbol{q}, z)$ being the spherical collapse component (for details see Kitaura et al. 2014).

We use the high correlation between the local density field and the velocity dispersion to model the displacement due to (quasi-) virialized motions. Effectively, we sample a Gaussian distribution function ($\mathcal{G}$) with a dispersion given by $\sigma_v \propto \left(1 + b^{ALPT}\delta^{ALPT}(\boldsymbol{x})\right)^{\gamma}$. Consequently,

$$\boldsymbol{v}^{\sigma}_r \equiv (\boldsymbol{v}^{\sigma}\cdot\hat{\boldsymbol{r}})\hat{\boldsymbol{r}}/(Ha) = \mathcal{G}\left(g \times \left(1 + \delta^{ALPT}(\boldsymbol{x})\right)^{\gamma}\right)\hat{\boldsymbol{r}}. \tag{8}$$

For the Gaussian streaming model see Reid & White (2011), and for non-Gaussian models see e.g. Tinker (2007). In closely virialized systems, the kinetic energy approximately equals the gravitational energy and a Keplerian law predicts γ close to 0.5, leaving only the proportionality constant g as a free parameter in the model (see also Heß, Kitaura & Gottlöber 2013). We assign larger dispersion velocities to low-mass objects considered to be satellites. The parameters g and γ have been adjusted to fit the damping effect in the monopole and quadrupole as found in the BigMultiDark N-body simulation first and later further constrained with the BOSS DR12 data for different redshift bins (see discussion in Section 3).

2.2.5 *Halo/stellar mass distribution reconstruction*

Once we have a spatial distribution of objects $\{\boldsymbol{r}_g\}$ which accurately reproduce the clustering of the whole galaxy sample at a given redshift, we assign the halo/stellar masses M^l_g to each object l according to the statistical information extracted from the BigMultiDark simulation using the Halo mAss Distribution ReconstructiON (HADRON) code (for technical details see Zhao et al. 2015). In particular, we sample the following conditional probability distribution function

$$M^l_g \curvearrowleft P(M^l_g|\{\boldsymbol{r}_g\}, \rho_M, T, \Delta r^M_{min}, \{p_c\}, z), \tag{9}$$

where ρ_M is the local density, T the tidal field tensor (in particular the eigenvalues), Δr^M_{min} a minimum separation between massive objects due to exclusion effects, $\{p_c\}$ a set of cosmological parameters, and z the redshift at which we want to apply the mass reconstruction. We note that at this stage we consider non-local biasing through the tidal field tensor and the minimum separation of objects. Using all this information, it has been proven that one can recover compatible clustering for arbitrary halo mass cuts with the N-body simulation up to scales of about $k = 0.3\,h^{-1}$ Mpc (Zhao et al. 2015). We extend the algorithm to stellar masses including the rank-ordering relation and scatter described in Section 2.1.

2.2.6 *Survey generator*

The SUrvey GenerAtoR (SUGAR) code is an openMP code which constructs light cones from mock galaxy catalogues (see Rodríguez-Torres et al. 2015, companion paper). This code applies geometrical features of the survey, including the geometry (using the publicly available MANGLE mask; Swanson et al. 2008), sector completeness, veto masks, and radial selection functions.

The SUGAR code can construct light cones using a single box or multiples boxes at different redshifts, in order to include the redshift evolution in the final catalogue. The first step in the construction of the light cone is to locate the observer ($z = 0$) and to transform from comoving Cartesian coordinates to equatorial coordinates (RA,Dec.) and redshift. To compute the observed redshift (redshift space) of an object, first we compute the comoving distance from the observer to the object, and then we transform it to redshift space following $s = r_c + (\boldsymbol{v}\cdot\hat{\boldsymbol{r}})/aH(z_{real})$ (see Section 2.2.4), where $r_c(z)$ is computed from $r_c(z) = \int_0^{z_{real}} \frac{c\,dz'}{H_0\sqrt{\Omega_M(1+z')^3+\Omega_\Lambda}}$.

Once we compute the redshift of each galaxy, we consider two options to select objects in the radial direction:

(i) downsampling: this option preserves the clustering of the input box selecting objects randomly to have the desired number density.

(ii) selecting by halo property: this consists of rank ordering objects by a halo property and selecting them sequentially until the correct number density is obtained.

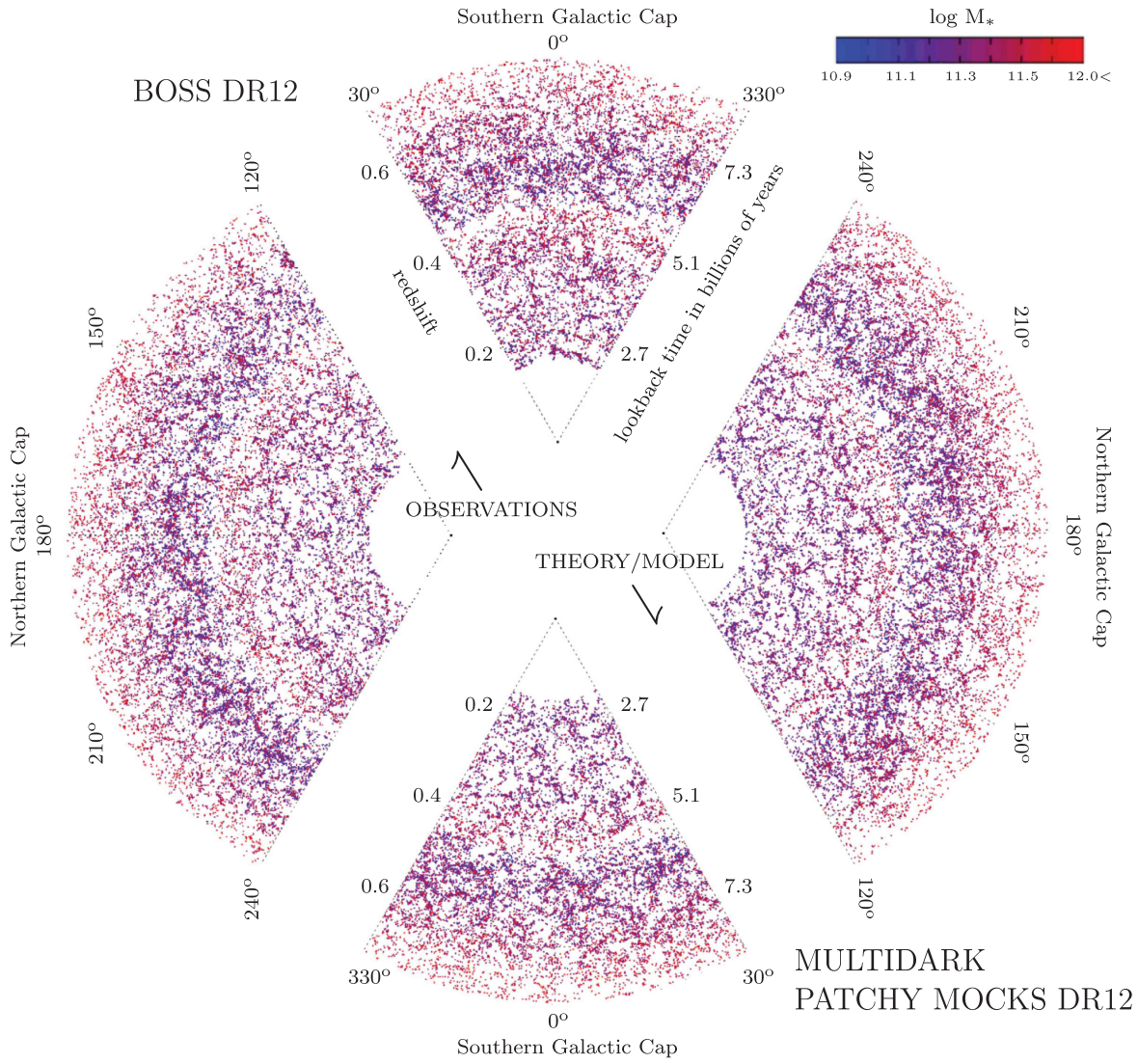


Figure 2. Pie plot of the BOSS DR12 observations (upper-left region) and one MultiDark PATCHY mock realization (lower-right region).

3 RESULTS: STATISTICAL COMPARISON BETWEEN THE MULTIDARK PATCHY MOCKS AND THE BOSS DR12 DATA

Following the method described in Section 2, we generate 12 288 mock light-cone galaxy catalogues for BOSS DR12[11] (2048 for each LOWZ, CMASS, combined, southern, and northern galactic cap). We call these catalogues MultiDark PATCHY mocks, MD PATCHY mocks in short. The corresponding computations required about 500 000 CPU hours (30–50 min for each box on 16 cores and a total of 40 960 boxes). Since each PATCHY+HADRON run requires less than 24 Gb shared memory for a grid with 960^3 cells, we were able to make use of 128 nodes with 32 Gb each in parallel from the BSC Marenostrum facilities, taking about one week wall clock time for all 40 960 catalogues. The light-cone generation with SUGAR required an additional ∼1000 CPU hours. The equivalent computations based on *N*-body simulations would have required about 9000 million CPU hours (∼2.3 million CPU hours for each light cone). The effective number of particles is $\sim(61\,440)^3$ (given that the reference catalogue required 3840^3 particles to resolve the objects we reproduce in the MD PATCHY catalogues).

We used 10 redshift bins to construct the light cones. This permits us to obtain the galaxy bias, the growth, and the peculiar motion evolution as a function of redshift. A visualization of the BOSS DR12 and one MD PATCHY mock realization is shown in Fig. 2. We can clearly see from this plot that both the data and the mocks follow the same selection criteria including the survey mask (the colour code stands for the stellar mass), and there are no obvious visual differences beyond cosmic variance. The empty regions seem to be similarly distributed for both cases, indicating that the three-point statistics should be close, and the statistical comparison between the MD PATCHY mock galaxy catalogues and the observations of BOSS DR12 yields good agreement. The number densities for LOWZ and CMASS galaxy samples are recovered by construction (see Fig. 3). We investigate the performance of the mock galaxy catalogues in detail in the following subsections.

[11] We have produced half the amount of mock catalogues for DR11, i.e. 1024 for each LOWZ, CMASS, combined, southern, and northern galactic cap.

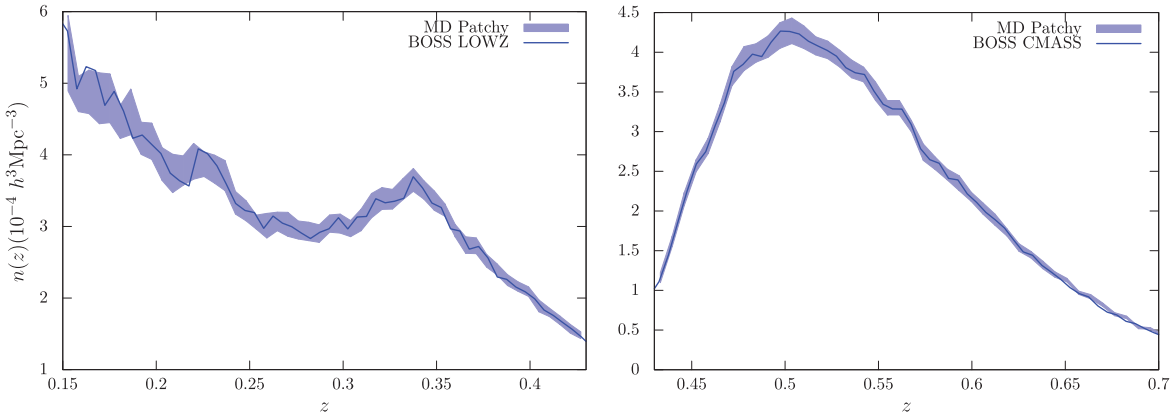


Figure 3. Number density for the LOWZ (left) and CMASS (right) samples. The observations are given by the blue solid lines. The shaded contours represent the 1σ regions according to the MD PATCHY mocks.

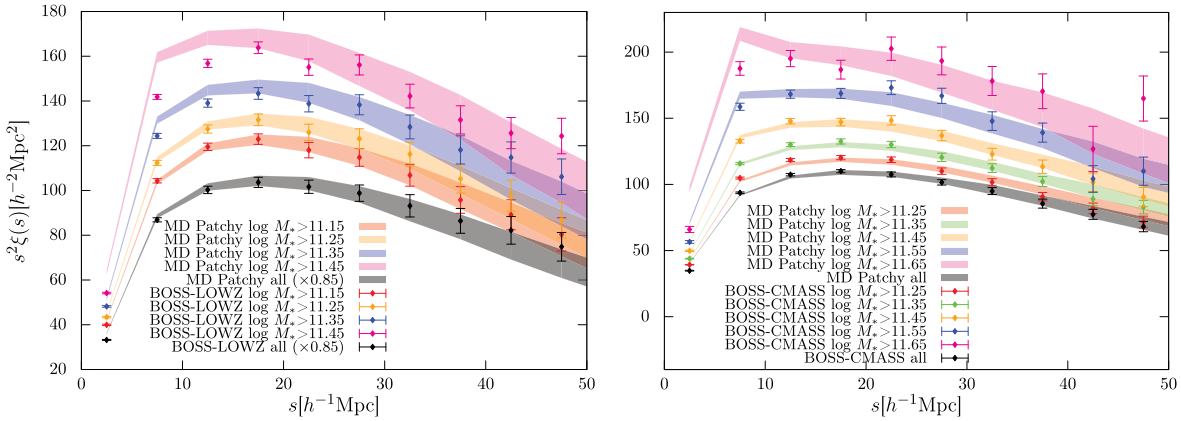


Figure 4. Monopole for different stellar mass bins as indicated in the legend with the corresponding colour code. The error bars represent the BOSS DR12 data. The shaded contours represent the 1σ regions according to the MD PATCHY mocks.

To avoid redundancy, we show only the results for BOSS DR12, as the only difference with respect to the BOSS DR11 mocks is the applied mask and selection function.

3.1 Two-point and three-point correlation functions

We perform first an analysis in configuration space computing the two- and three-point correlation functions. To compute the clustering signal in the correlation function for the MD PATCHY mock light cones and the observed data, we rely on the Landy & Szalay (1993) estimator. We will follow their notation referring to the data sample (either simulation or observed data) as D and to the random catalogue as R.

The correlation function is then constructed in the following way:

$$\xi(s) = \frac{DD - 2DR + RR}{RR}, \tag{10}$$

as a function of separation between pairs of galaxies in redshift space s.

The three-point correlation function gives a description of the probability of finding three objects in three different volumes, and can be computed following Szapudi & Szalay (1998),

$$\zeta(s_{12}, s_{23}, s_{13}) = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}, \tag{11}$$

as a function of separation between the vertices of triangles spanned by triplets of galaxies in redshift space s_{12}, s_{23}, s_{13}.

Fig. 4 shows that we accurately recover the clustering (monopole) for arbitrary stellar mass bins showing almost perfect agreement with observations. Only for the two largest stellar mass bin, we find deviations larger than 1σ. This is mainly due to the 'halo exclusion effect', which is only approximately modelled, assuming a minimum separation for massive galaxies, and not the full separation distribution function (Zhao et al. 2015). We find, however, that these differences are not critical, as they are restricted to small scales ($\lesssim 20\,h^{-1}$ Mpc) and only a low number of objects are affected. We further compute the monopole and quadrupole for LOWZ and CMASS (see Fig. 5 and Section 3.3). The monopole agrees towards small scales down to a few Mpc within 1σ.

There is a deviation of the monopole around the BAO peak and towards larger scales. While the galaxy mock catalogues cross zero right after the BAO peak, the observations do not. In this study, we have applied all of the systematic weights, such as the stellar density contamination, detailed in Reid et al. (2016) and Ross et al. (in preparation). The correlation function measurements are quite covariant between s bins at these scales, making the deviations less significant than one would expect by the visual impression. The significance and potential causes of the large-scale excess are studied in Ross et al. (in preparation), where it is also shown that it has no significant impact on BAO measurements. This is even more so, as the overall shape of $\xi(s)$ in BAO measurements is

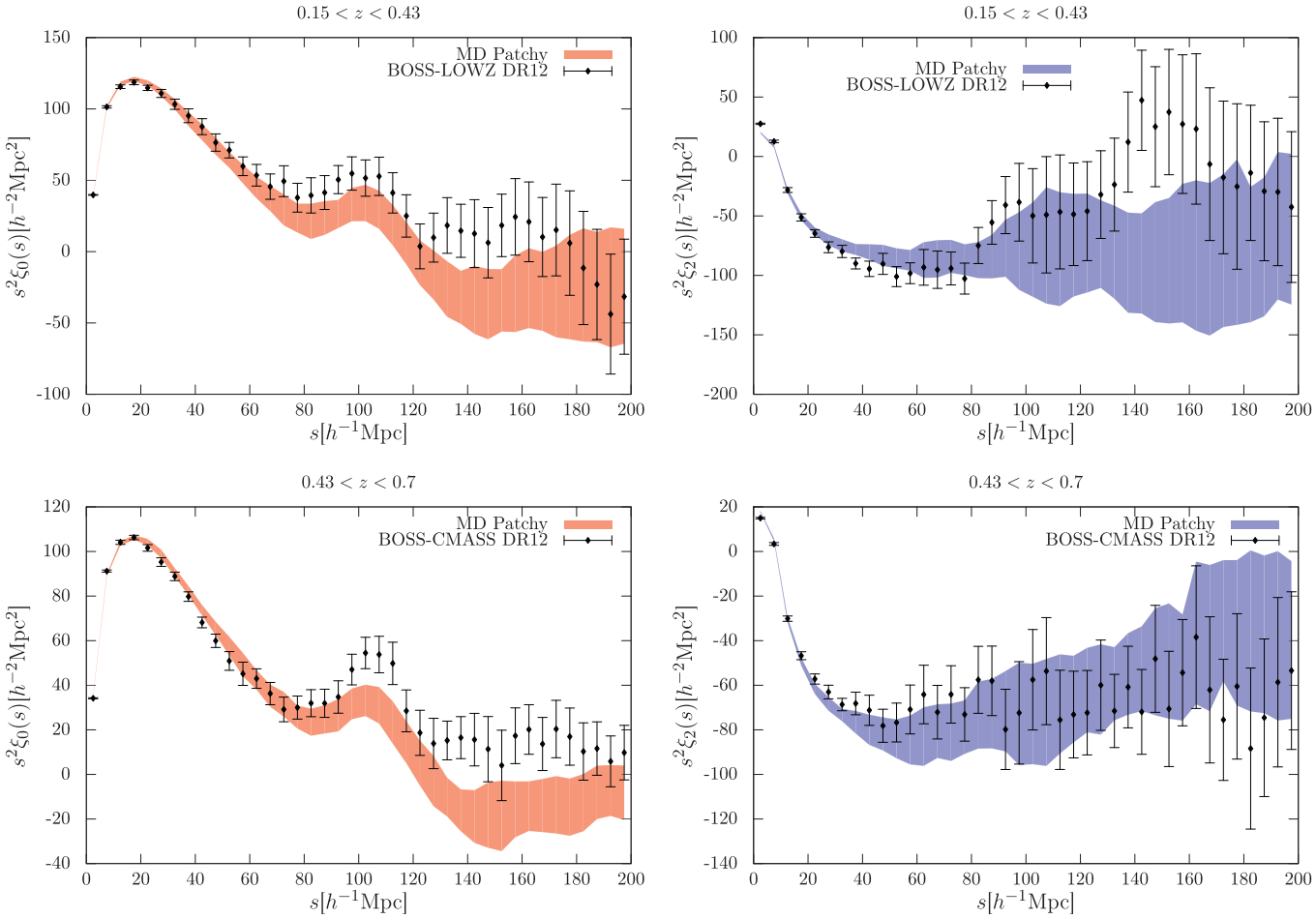


Figure 5. Monopole (on the left) and quadrupole (on the right) for LOWZ and CMASS in the first and second rows, respectively. The shaded contours represent the 1σ regions according to the MD PATCHY mocks, correlation function in red, quadrupole in blue.

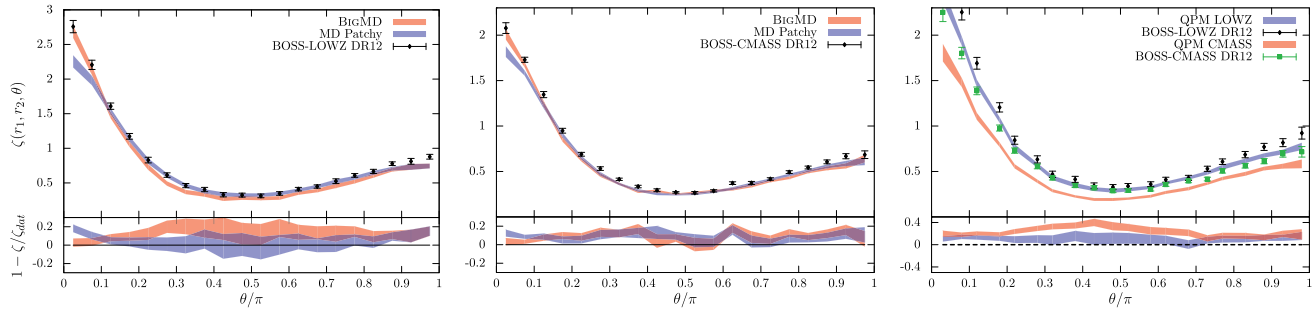


Figure 6. Left-hand and central panels: three-point statistics comparing the MD PATCHY mocks (blue shaded region) with the BigMultiDark mocks of the *N*-body simulation (red shaded region) and the observations (black error bars) for LOWZ (left) and CMASS galaxies (central). Right-hand panel: three-point statistics comparing the QPM mocks (LOWZ: blue shaded region, CMASS: red shaded region) to the observations (LOWZ: black error bars, CMASS: green error bars). Corresponding ratios are shown in the bottom panels. Shaded area shows 1σ uncertainties, $r_1 = 10$ and $r_2 = 20\,h^{-1}$ Mpc and θ is the angle between r_1 and r_2 h^{-1} Mpc.

marginalized over with a polynomial (see e.g. Anderson et al. 2014). See also Ross et al. (2012) and Chuang et al. (2013) for similar studies on an earlier BOSS data set and Huterer et al. (2013) for potential photometric calibration systematics, which have not been accounted for in this analysis.

In the case of RSD measurements, one has to make sure that the analysis is performed on scales which are not affected by systematics (Gil-Marín et al. 2015a, companion paper). The quadrupole is in very good agreement on all scales, further supporting that RSD analysis should be safe, even in case there are some remnant systematics in the data.

An investigation of the three-point function demonstrates that the MD PATCHY mocks have a quality very similar to those based on *N*-body simulations after calibration (see the left-hand and central panels in Fig. 6). We have constrained the galaxy bias parameters (see Sections 2.2.2, 2.2.3, and 2.2.5) based on the reference

catalogues from the BigMultiDark simulation on cubical full volumes at each of the 10 redshift bins, matching the two- and the three-point statistics. To fit the latter, we focused on matching the higher order correlation functions through the probability distribution function of galaxies in the reference catalogues following the approach presented in Kitaura et al. (2015). Using the observations to constrain the three-point statistics is not trivial, due to incompleteness effects. This explains why the MD PATCHY mock catalogues better fit the reference catalogue than the data, especially for the CMASS galaxies. The three-point statistics performs worse for the QPM mocks, possibly because they do not include an iterative validation step fitting higher order statistics (beyond the two-point correlation function). The non-linear RSD parameter (see Section 2.2.4) was iteratively constrained based on the observations, as we explain in the next section.

3.2 Monopole and quadrupole in Fourier space

The galaxy power spectrum P and the galaxy bispectrum B are the two- and three-point correlation functions in Fourier space. Given the Fourier transform of the galaxy overdensity, $\delta_{\rm g}(\boldsymbol{x}) \equiv \rho_{\rm g}(\boldsymbol{x})/\bar{\rho}_{\rm g} - 1$,

$$\delta_{\rm g}(\boldsymbol{k}) = \int {\rm d}^3\boldsymbol{x}\, \delta_{\rm g}(\boldsymbol{x}) \exp(-{\rm i}\boldsymbol{k}\cdot\boldsymbol{x}), \tag{12}$$

where $\rho_{\rm g}(\boldsymbol{x})$ is the number density of objects and $\bar{\rho}_{\rm g}$ its mean value, and the galaxy power spectrum and galaxy bispectrum are defined as

$$\langle \delta_{\rm g}(\boldsymbol{k})\delta_{\rm g}(\boldsymbol{k}')\rangle \equiv (2\pi)^3 P(k)\delta^D(\boldsymbol{k}+\boldsymbol{k}'), \tag{13}$$

$$\langle \delta_{\rm g}(\boldsymbol{k}_1)\delta_{\rm g}(\boldsymbol{k}_2)\delta_{\rm g}(\boldsymbol{k}_3)\rangle \equiv (2\pi)^3 B(\boldsymbol{k}_1,\boldsymbol{k}_2)\delta^D(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3), \tag{14}$$

with δ^D being the Dirac delta function. Note that the bispectrum is only well defined when the set of k-vectors, k_1, k_2, and k_3, close to form a triangle, $\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3 = \boldsymbol{0}$. It is common to define the reduced bispectrum Q as

$$Q(\alpha_{12}|\boldsymbol{k}_1,\boldsymbol{k}_2) \equiv \frac{B(\boldsymbol{k}_1,\boldsymbol{k}_2)}{P(k_1)P(k_2)+P(k_2)P(k_3)+P(k_1)P(k_3)}, \tag{15}$$

where α_{12} is the angle between $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$. This quantity is independent of the overall scale k and redshift at large scales and for a power spectrum that follows a power law. Moreover, it presents a characteristic 'U-shape' predicted by gravitational instability. Mode coupling and power-law deviations in the actual power spectrum induce a slight scale and time dependence in this quantity. However, in practice it has been observed that at scales of the order of $k \sim 0.1\,h\,{\rm Mpc}^{-1}$ the reduced bispectrum does not present a high variation in its amplitude.

The measurement of the bispectrum is performed in the same way as the approach described in Gil-Marín et al. (2015c). This method consists of generating k-triangles and randomly orientating them in k-space. When the number of random triangles is sufficiently large, the mean value of their bispectra tends to the fiducial bispectrum (for details see Gil-Marín et al. 2015c).

Discreteness adds a shot noise contribution to the measured power spectrum and bispectrum. In this paper, we assume that these contributions are of Poisson type and therefore are given by

$$P_{\rm sn}(k) = \frac{1}{\bar{n}} \tag{16}$$

$$B_{\rm sn}(\boldsymbol{k}_1,\boldsymbol{k}_2) = \frac{1}{\bar{n}}\left[P(k_1)+P(k_2)+P(k_3)\right] + \frac{1}{\bar{n}^2}, \tag{17}$$

where $k_3 = |\boldsymbol{k}_1+\boldsymbol{k}_2|$ and $\bar{n}$ is the number density of haloes.

For both power spectrum and bispectrum, we present the BOSS DR12 data error bars computed from the dispersion among 2048 and 100 realizations of MD PATCHY mock catalogues, respectively.

The Fourier space analysis has been used to improve the modelling of the RSDs in the galaxy mock catalogues. We have assigned higher peculiar random motions to about 10 per cent of the galaxies to fit the quadrupole of the data with a specific value for each of the 10 redshift bins. The resulting monopoles and the quadrupoles show good agreement with the observations over the range relevant to BAOs and RSDs up to at least $k \simeq 0.3\,h^{-1}$ Mpc for both LOWZ and CMASS (see Fig. 7). This agreement is further supported after BAO reconstruction, as can be seen in Fig. 8. Only towards the very large scales ($k \lesssim 0.02\,h^{-1}$ Mpc), we can find that the observed monopole tends to be larger than the mock catalogues (both MD PATCHY and QPM). This hints towards the discrepancy in the monopole found in configuration space (see the previous section). Although the PATCHY method can potentially yield accurate two-point statistics up to $k \sim 1\,h^{-1}$ Mpc (see Kitaura et al. 2014; Chuang et al. 2015b), we have restricted the study to lower ks, as the analysis of BAOs and

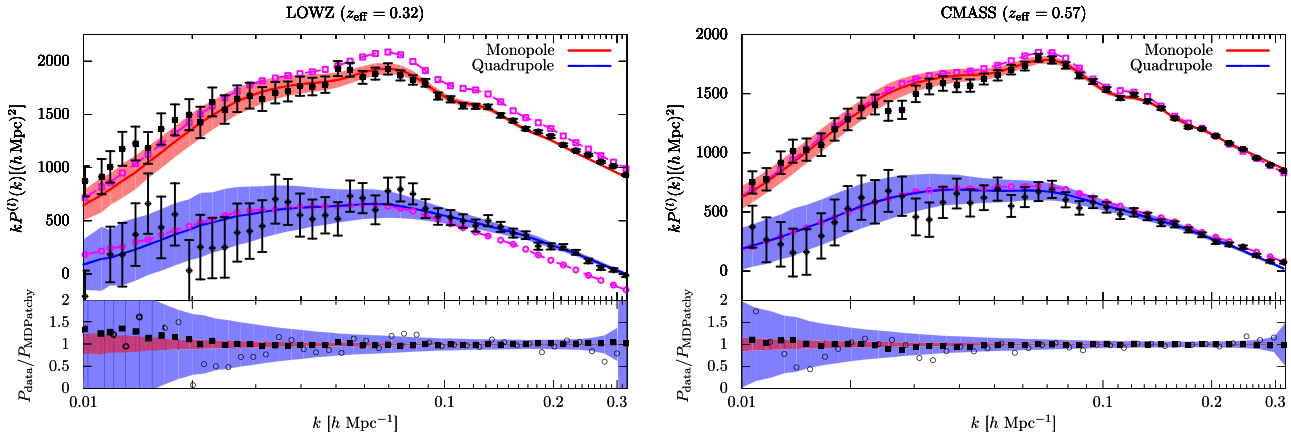


Figure 7. Monopole (red) and quadrupole (blue) in Fourier space for the LOWZ (left) and CMASS galaxies (right) for the mean over 2048 MD PATCHY mocks for both southern and northern galactic caps, the average and 1σ uncertainties are shown. The results for QPM (1000 mocks for each LOWZ/CMASS, and north/south) are shown with dashed magenta lines. The error bars assigned to the data points have been computed based on 2048 MD PATCHY mocks. The ratio plots in the bottom panels have been only done for the MD PATCHY mocks.

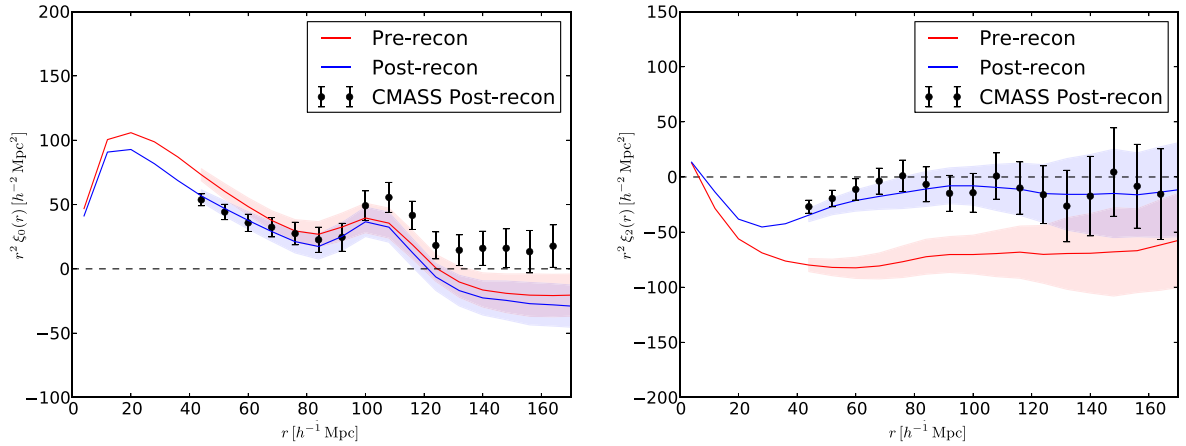


Figure 8. Monopole (on the left) and quadrupole (on the right) before and after BAO reconstruction (see Vargas-Magana et al., in preparation). The error bars represent the BOSS DR12 data. The solid lines correspond to the mean, and the shaded contours represent the 1σ regions, according to the MD PATCHY mocks (red pre-, and blue post-reconstruction).

RSDs will not be done beyond $k = 0.3\,h^{-1}$ Mpc, and the computation of power spectra for thousands of mocks with large grids becomes very expensive.

This fitting procedure had, however, as a consequence that the three-point correlation function is slightly less precise at angles close to $\theta \sim 0$ and $\theta \sim \pi$, as can be seen in Fig. 6, which prior to this operation was fully compatible with the reference catalogue. In fact, the reference BigMultiDark catalogue used in this study showed a highly discrepant quadrupole, as compared to the observations. This has been deeply analysed and better agreement has been found based on an improved HAM procedure applied to the BigMultiDark simulation (see Rodríguez-Torres et al. 2015, companion paper), which however was not available at the moment of the generation of the MD PATCHY mocks. The HOD model adopted in the QPM mock catalogues assumed about 10 per cent satellite galaxies. This yields a compatible quadrupole for the CMASS galaxies. However, as these catalogues were not iteratively calibrated for different redshift slices, their agreement with the LOWZ galaxies is less accurate.

A detailed analysis of the bispectra is presented in Figs 9 and 10 demonstrating reasonable agreement between the mocks and the observations for different configurations of triangles across a wide range of scales, given the high uncertainties introduced by the mask, selection function, and cosmic variance.

3.3 Cosmic evolution

The cosmic evolution modelled in the MD PATCHY mocks was achieved by fitting the clustering of 10 redshift bins for the full redshift range spanning about 5 Gyr. This implied running structure formation with ALPT for each redshift, i.e. modelling the growth of structures and the growth rate, and additionally fitting the galaxy bias evolution and the non-linear RSDs. The evolution of clustering for both sets of mocks in the full redshift range is shown in Fig. 11. While the correlation function for CMASS galaxies does not show strong differences along the CMASS redshift range, this evolution is very apparent for the LOWZ sample. Fig. 12 shows the comparison between the mocks and the observations for different LOWZ in more detail. The QPM mocks do not include a detailed cosmic evolution within LOWZ or CMASS being based on mean redshifts for each case. This explains why these mocks lose accuracy in the two-point statistics towards low redshifts.

We investigate now the cosmic evolution of the covariance matrices derived from the MD PATCHY mocks[12] computed as in Anderson et al. (2014):

$$\mathrm{cov}[i,\,j] = \frac{\sum_l (\xi_i^l - \langle\xi_i^l\rangle)(\xi_j^l - \langle\xi_j^l\rangle)}{N_\mathrm{s} - 1}\,, \tag{18}$$

with bins i and j, mock sample l, and N_s being the number of simulations.

The correlation matrices for different redshift bins shown in Fig. 13 were constructed upon the covariance matrices following

$$\mathcal{C}[i,\,j] = \frac{\mathrm{cov}[i,\,j]}{\sqrt{\mathrm{cov}[i,\,i]}\sqrt{\mathrm{cov}[j,\,j]}}\,. \tag{19}$$

We find that the correlation matrices vary in subsequent redshift bins. First, the correlation matrices are increasingly correlated close to the diagonal for both the monopole and the quadrupole towards lower redshifts, as expected from gravitational evolution coupling different scales. This is seen in Fig. 13 as the diagonal red band becomes broader especially comparing the highest redshift bin with the lower ones. Secondly, we find that moderate off-diagonal correlations present at higher redshifts disappear towards lower redshifts. And thirdly, we can see that the correlation between the monopole and the quadrupole at large scales becomes maximal in the redshift bin $0.43 < z < 0.55$, as can be seen in the white region in the lower-right and upper-left blocks. This 'triangular' correlation is expected from linear theory (see equations 7 and 9 in Chuang & Wang 2013).

Further calculations of the correlation functions including QPM mocks are shown in companion publications (Gil-Marín et al. 2015a,b, companion papers).

Additionally, we show in Fig. 14 the angular correlation function and in Fig. 15 the multipole moments (including the hexadecapole) for different redshift bins based on the combined sample showing good agreement between the MD PATCHY mocks and the data.

4 FUTURE WORK

We have taken advantage in this survey of the characteristic bias of LRGs, being massive objects residing in high-density regions.

[12] Covariance matrices for the different catalogues (LOWZ, CMASS, and combined sample) will be made publicly available with the publication of the galaxy catalogue.

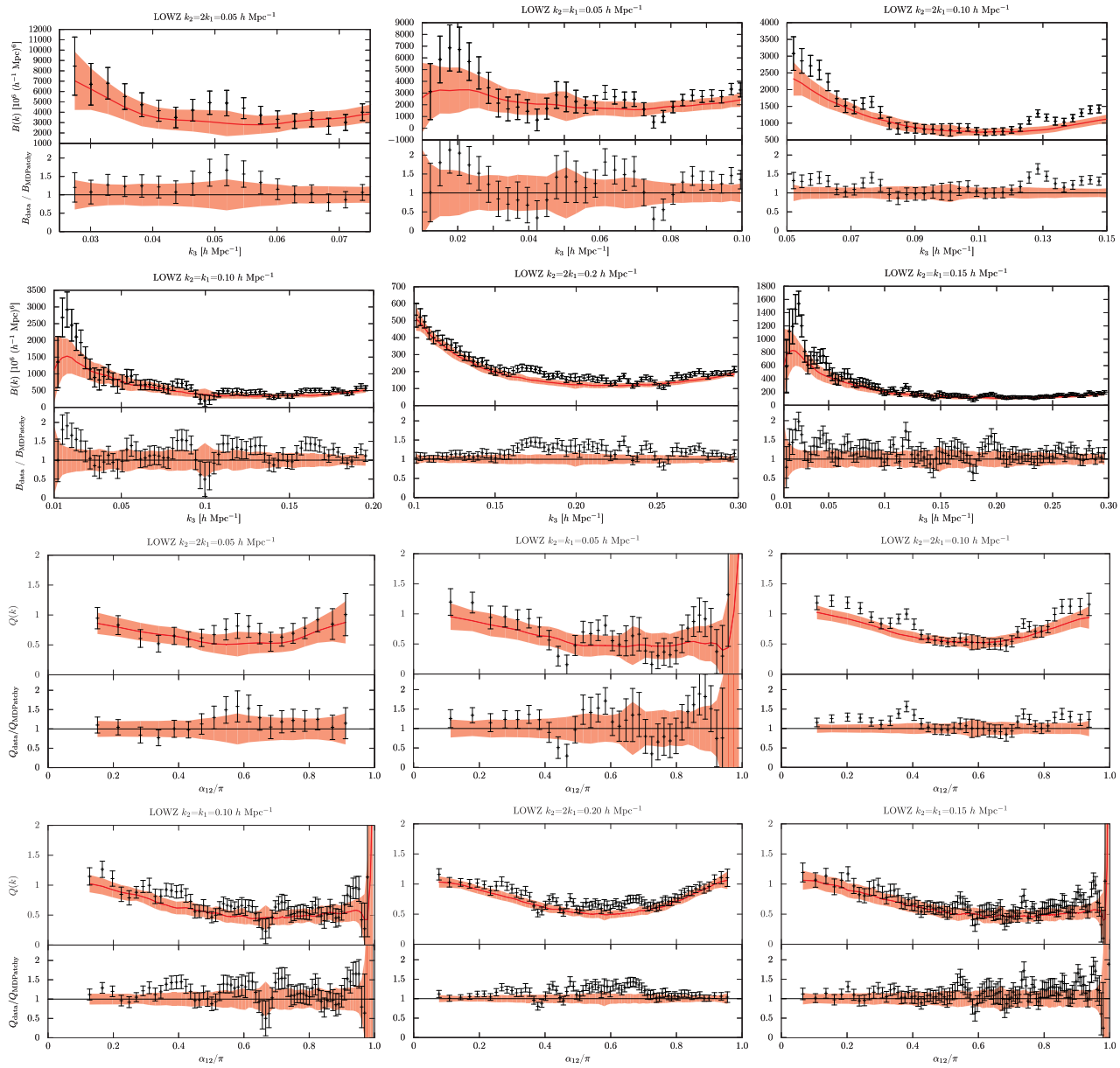


Figure 9. Bispectra and reduced bispectra for LOWZ mocks and observed galaxies for different configurations indicated above each panel. The red solid line corresponds to the mean and the red shaded region to the 1σ contour of 100 MD PATCHY mocks. The black dots correspond to the BOSS DR12 data with the error bars taken from the MD PATCHY mocks.

This work confirms that threshold bias is an essential ingredient to explain the clustering of LRGs. This facilitates our analysis, since the low-density filamentary network did not need to be accurately described, and it has permitted us to rely on low-resolution (augmented Lagrangian) PT-based methods. This will no longer apply for upcoming surveys based on emission line galaxies residing in the whole cosmic web. One could improve the methodology presented in this work by substituting the structure formation model based on PT with a more accurate one (e.g. COLA). Whether this is necessary, or whether more efficient alternative approaches are sufficient (e.g. ALPT with MUSCLE corrections), will be investigated in future works.

Non-local bias was only considered in the mass assignment step, but neglected in the generation of the full galaxy population. This may become important to model for emission line galaxies, and needs a deeper analysis.

The approximate 'halo exclusion' modelling is mainly responsible for the deviation in the clustering of the most massive objects, and could be improved by taking their full distribution of relative distances, instead of taking a sharp minimum separation for each mass bin, as is done here.

Another aspect which still needs to be improved in the catalogues is the clustering on sub-Mpc scales. We have randomly assigned positions of dark matter particles to the mock galaxies without considering that some of them are satellites of central galaxies. This implies that these mocks are not appropriate for fibre-collision analysis. For the time being, we will leave the mock catalogues as they are, since most of the studies are not affected by this. Nevertheless, we would like to stress that this aspect can easily be corrected by assigning to a fraction of the mock galaxies close positions to the major most massive ones in the neighbourhood, without the need of redoing the catalogues. The QPM mocks better model fibre

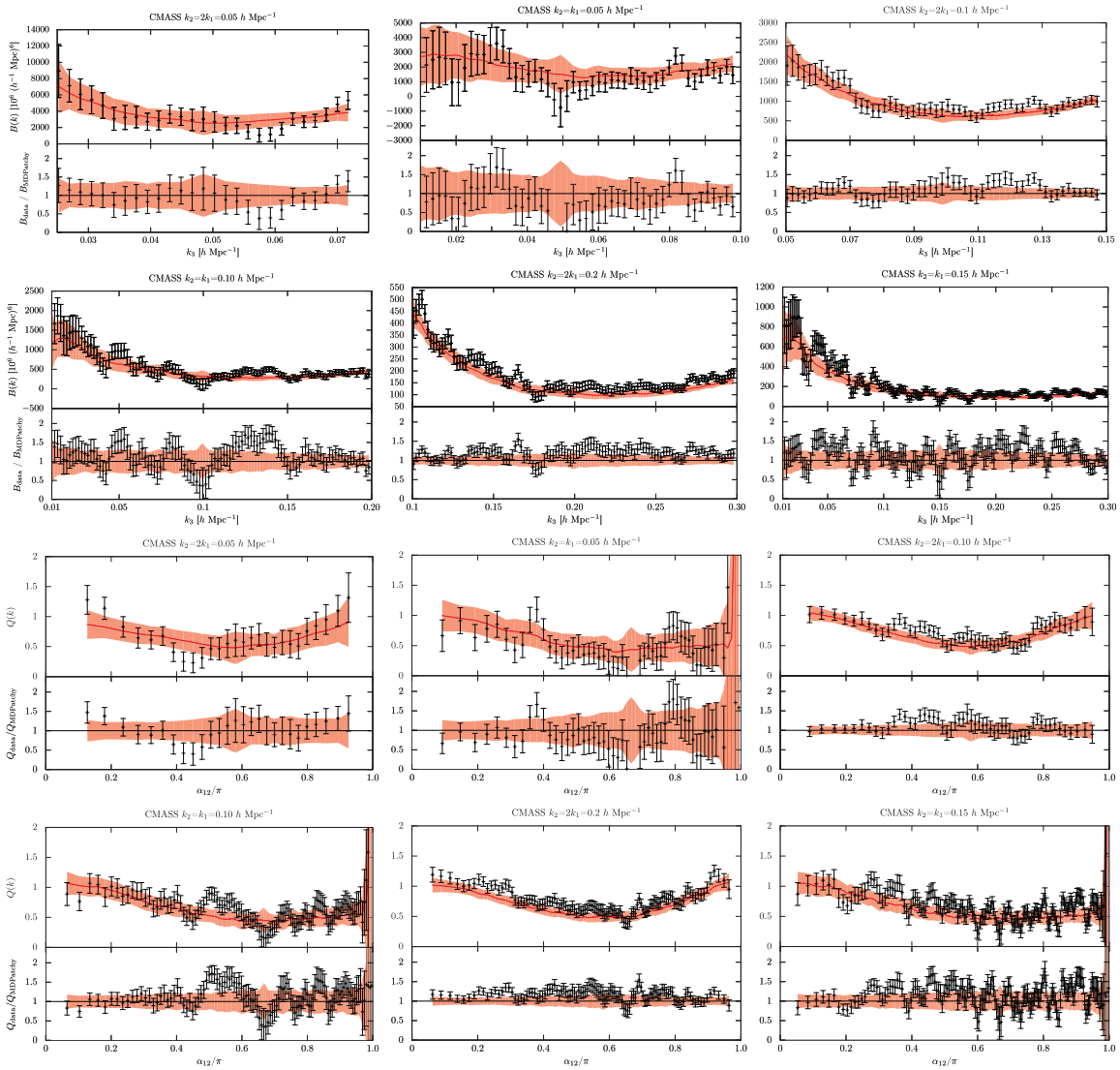


Figure 10. Bispectra and reduced bispectra for CMASS mocks and observed galaxies for different configurations. The red solid line corresponds to the mean and the red shaded region to the 1σ contour of 100 MD PATCHY mocks. The black dots correspond to the BOSS DR12 data with the error bars taken from the MD PATCHY mocks.

collisions, as the HOD adopted in this work successfully reproduced the fraction of close satellites and central galaxies (Gil-Marín et al. 2015a, companion paper).

Also the photometric calibration systematics, presumably responsible for the excess of power in the data towards large scales, require further investigation.

We have considered one fiducial cosmology. It would be, however, interesting to provide sets of mock catalogues running over different combinations of cosmological parameters.

Let us finally mention that we have ignored in this study super-survey modes, which may be especially relevant for the analysis of the power spectrum at very large scales (Takada & Hu 2013; Li, Hu & Takada 2014a,b; Carron & Szapudi 2015).

We aim at addressing all these issues in future works.

5 SUMMARY AND CONCLUSIONS

We have presented 12 288 mock galaxy catalogues for the BOSS DR12, including all relevant physical and observational effects, to enable a robust analysis of BAOs and RSDs.

The main features of these mock catalogues are as follows:

(i) large number of catalogues: 2048 for each LOWZ, CMASS, and combined LOWZ+CMASS and northern and southern galactic cap,

(ii) accurate structure formation model on scales of a few Mpc,

(iii) accurate galaxy bias model including non-linear, stochastic, threshold bias, and a non-local bias dependence on the tidal field tensor and the exclusion effect separation of massive objects,

(iv) modelling redshift evolution of galaxy bias, growth of structures, growth rate, and non-linear RSDs,

(v) and additional survey features, such as geometry, sector completeness, veto masks, and radial selection functions.

The same degree of accuracy is achieved for the BOSS DR11 MD PATCHY mocks, for which only 6000 light-cone mock catalogues were produced (1000 for each LOWZ, CMASS, and combined LOWZ+CMASS and northern and southern galactic cap).

The MD PATCHY mocks have shown a better match to the data than the QPM mocks in terms of two- and three-point statistics. Investigating the origin for these differences can be interesting as

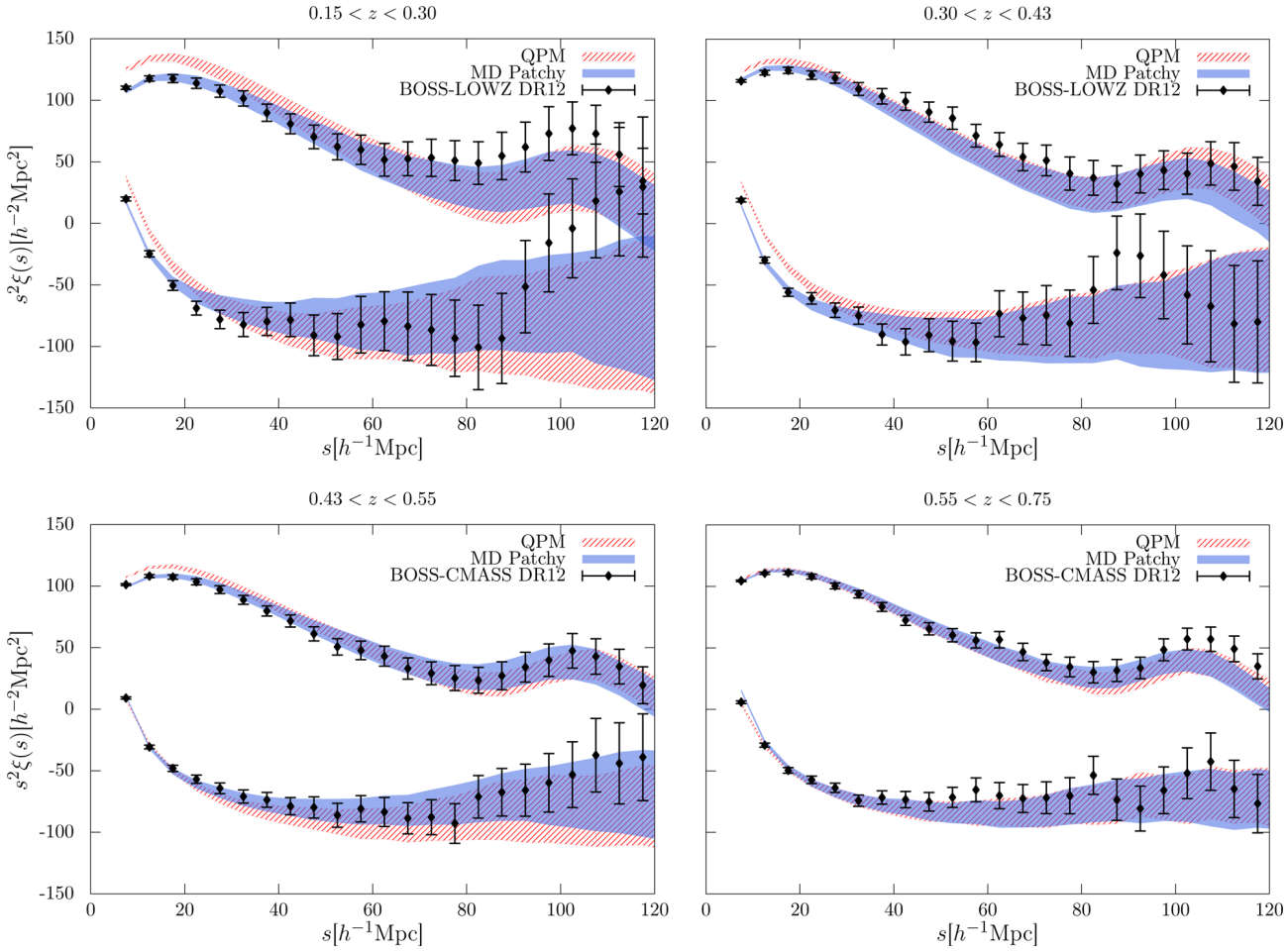


Figure 11. Monopole and the quadrupole for different redshift bins over the redshift range $0.15 < z < 0.7$. The black error bars stand for the BOSS DR12 data. The shaded contours represent the 1σ regions according to the MD PATCHY mocks in blue and according to the QPM mocks in red. These measurements are used in the BAO and RSD analysis in Chuang et al. (in preparation).

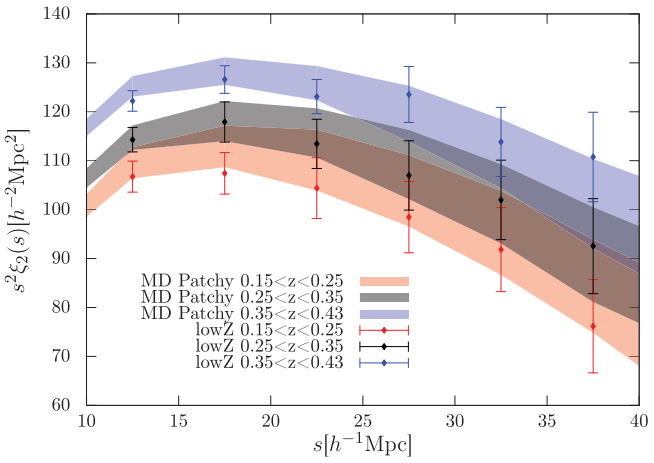


Figure 12. Monopole showing the evolution for LOWZ. The corresponding redshift bins for the PATCHY mocks are represented by shaded regions, and the observations by the error bars.

the physical models, and in particular the galaxy bias, adopted in each method are quite different.

We note that neglecting the stochastic bias considered in the MD PATCHY mocks, modelling the deviation from Poisson shot noise (predominantly overdispersion), could underestimate the clustering uncertainties.

The mock catalogues have enabled a robust analysis of the BOSS data yielding the necessary error estimates and the validation of the analysis methods. In particular, the studies include the following:

(i) a full clustering analysis (Grieb et al., in preparation; Sánchez et al., in preparation: see Fig. 15),
(ii) a tomographic analysis of the large-scale angular galaxy clustering, where full light-cone effects (e.g. growth, bias, and velocity field evolution) are essential (Salazar-Albornoz et al., in preparation: see Fig. 14),
(iii) a study of the BAO reconstructions (see Vargas-Magana et al., in preparation, and Fig. 8 showing the performance on the MD PATCHY mocks),
(iv) and an RSD analysis (Gil-Marín et al. 2015a, companion paper; Beutler et al., in preparation).

We have demonstrated that the MD PATCHY BOSS DR12 mock galaxies match, in general within 1σ, the clustering properties of the BOSS LRGs for the monopole, quadrupole, and hexadecapole of the two-point correlation function both in configuration and Fourier space. In particular, we achieve a high accuracy in the modelling of the monopole up to $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$. We have furthermore shown that we also obtain three-point statistics with the same level of

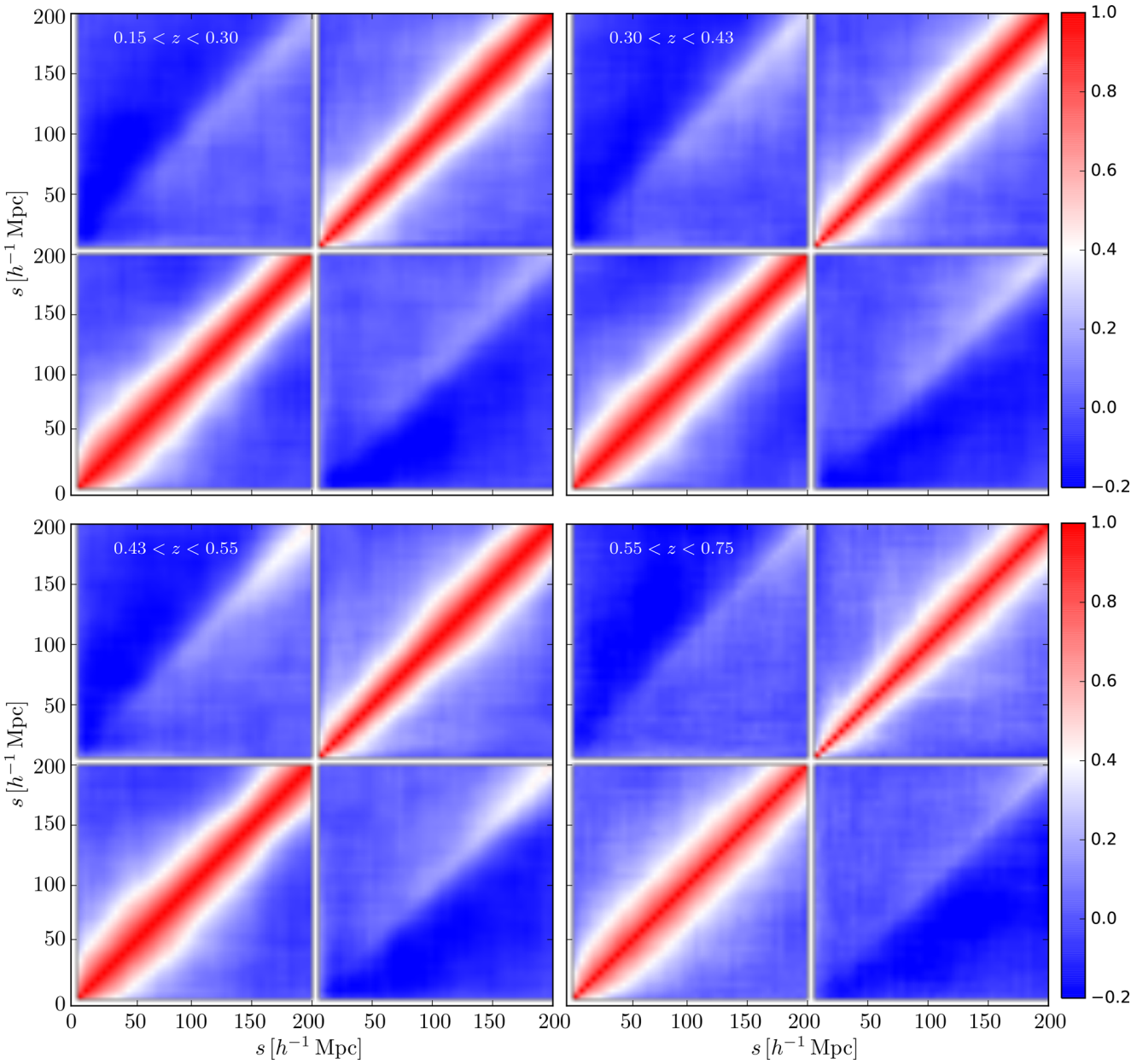


Figure 13. Cosmic evolution of the correlation matrices for different redshift bins indicated in the legend in bins of 5 h^{-1} Mpc. Lower-left block for the monopole, upper-right block for the quadrupole, and upper-left and lower-right blocks for the correlations between the monopole and the quadrupole. See Section 3.3 for details of the calculation. These correlation matrices are used in the BAO and RSD analysis in Chuang et al. (in preparation).

accuracy as *N*-body-based catalogues at scales larger than a few Mpc, which are close to the observations.

The good agreement between the models and the observations demonstrates the level of accuracy reached in cosmology, our understanding of structure formation, galaxy bias, and observational systematics.

All the mock galaxy catalogues and the corresponding covariance matrices will be made publicly available together with the release of the BOSS DR12 galaxy catalogue.

ACKNOWLEDGEMENTS

FSK, SRT, CC, CZ, FP, AK, and CGS acknowledge support from the Spanish MICINNs Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064, MINECO Centro de Excelencia Severo Ochoa Programme under grant SEV- 2012-0249, and grant AYA2014-60641-C2-1-P. FSK and CZ also want to thank the Instituto de Física Teórica UAM/CSIC for the hospitality and support during several visits, where part of this work was completed. FP wishes to thank the Lawrence Berkeley National Laboratory for the hospitality during the development of this work; he also acknowledges the Spanish MEC 'Salvador de Madariaga' programme, Ref. PRX14/00444. HGM is grateful for support from the UK Science and Technology Facilities Council through the grant ST/I001204/1. HG acknowledges the support of the 100 Talents Program of the Chinese Academy of Sciences. GY wishes to thank MINECO (Spain) for financial support under project grants AYA2012-31101 and FPA2012-34694. He also thanks the Red Española de Supercomputación for granting computing time in the Marenostrum supercomputer, in which part of this work has been done. AGS, SSA,

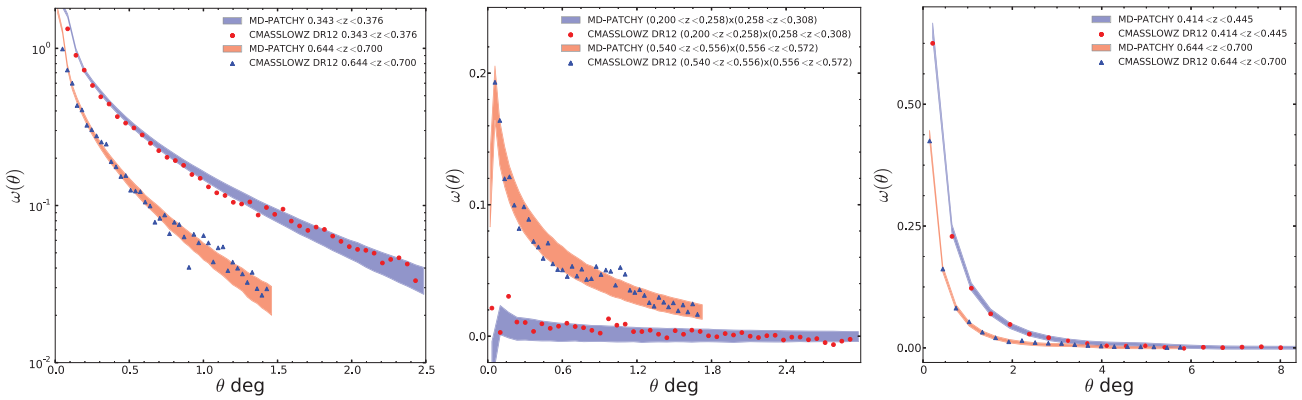


Figure 14. Angular correlation functions based on the combined sample. Left-hand panel: angular auto-correlation function on small scales for two different tomographic bins (see key for redshift ranges), where colour bands are the mean and 1σ region of MD PATCHY and symbols correspond to the measurements on the DR12 combined sample. Central panel: angular cross-correlation function on small scales between different tomographic bins, following the same key as the left-hand panel. Right-hand panel: large-scale angular auto-correlation function for two different redshift bins. These measurements are used in the tomographic analysis of galaxy clustering in Salazar-Albornoz et al. (in preparation).

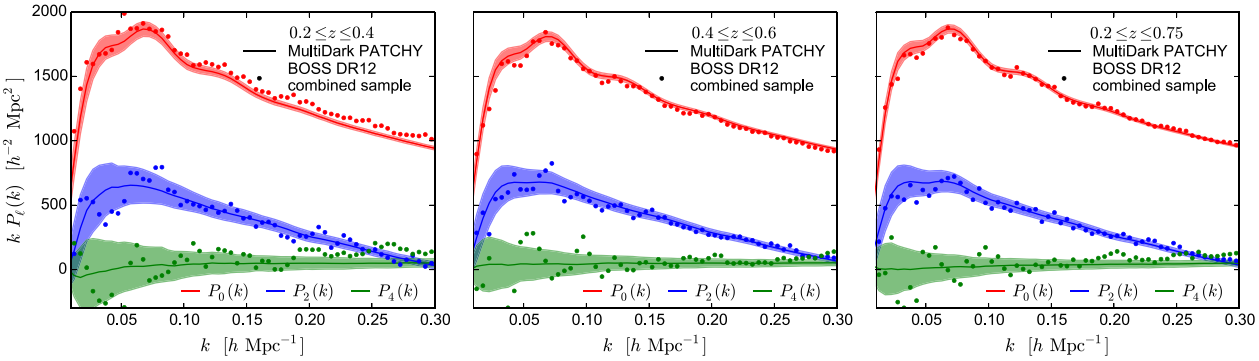


Figure 15. Multipole moments based on the combined sample: monopole P_0, quadrupole P_2, and hexadecapole P_4 for different redshift bins (see key for redshift ranges), where colour bands are the mean of MD PATCHY and symbols correspond to the measurements on the DR12 combined sample. These measurements are used in the wedges analysis of galaxy clustering in Grieb et al. (in preparation).

and JNG acknowledge support from the Transregional Collaborative Research Centre TR33 'The Dark Universe' of the German Research Foundation (DFG). AJC is supported by supported by the European Research Council under the European Community's Seventh Framework Programme FP7-IDEAS-Phys.LSS 240117. Funding for this work was partially provided by the Spanish MINECO under project MDM-2014-0369 of ICCUB (Unidad de Excelencia 'María de Maeztu').

The massive production of all MultiDark PATCHY BOSS DR12 mocks has been performed at the BSC Marenostrum supercomputer, the Hydra cluster at the Instituto de Física Teórica UAM/CSIC, and NERSC at the Lawrence Berkeley National Laboratory.

The BigMultiDark simulations have been performed on the SuperMUC supercomputer at the Leibniz-Rechenzentrum (LRZ) in Munich, using the computing resources awarded to the PRACE project number 2012060963. We want to thank V. Springel for providing us with the optimized version of GADGET-2.

Numerical computations for the power spectrum multipoles and bispectrum were performed on the Sciama High Performance Compute (HPC) cluster which is supported by the ICG, SEPNet, and the University of Portsmouth.

This research also used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the US Department of Energy Office of Science. The SDSS-III website is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, University of Cambridge, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

REFERENCES

Ahn K., Iliev I. T., Shapiro P. R., Srisawat C., 2015, MNRAS, 450, 1486
Alam S. et al., 2015, ApJS, 219, 12
Alimi J.-M. et al., 2012, preprint (arXiv:1206.2838)
Anderson L. et al., 2014, MNRAS, 441, 24
Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, MNRAS, 426, 2046
Angulo R. E., Baugh C. M., Frenk C. S., Lacey C. G., 2014, MNRAS, 442, 3256
Ata M., Kitaura F.-S., Müller V., 2015, MNRAS, 446, 4250
Baldauf T., Seljak U., Desjacques V., McDonald P., 2012, Phys. Rev. D, 86, 083540
Baldauf T., Seljak U., Smith R. E., Hamaus N., Desjacques V., 2013, Phys. Rev. D, 88, 083507
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Beltrán Jiménez J., Durrer R., 2011, Phys. Rev. D, 83, 103509
Benitez N. et al., 2014, preprint (arXiv:1403.5237)
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Bernardeau F., 1994, ApJ, 427, 51
Beutler F. et al., 2011, MNRAS, 416, 3017
Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517
Bolton A. S. et al., 2012, AJ, 144, 144
Bouchet F. R., Colombi S., Hivon E., Juszkiewicz R., 1995, A&A, 296, 575
Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, MNRAS, 398, 1150
Buchert T., 1994, MNRAS, 267, 811
Carron J., Szapudi I., 2015, MNRAS, 447, 671
Casas-Miranda R., Mo H. J., Sheth R. K., Boerner G., 2002, MNRAS, 333, 730
Catelan P., 1995, MNRAS, 276, 115
Cen R., Ostriker J. P., 1993, ApJ, 417, 415
Chan K. C., Scoccimarro R., Sheth R. K., 2012, Phys. Rev. D, 85, 083509
Chuang C.-H., Wang Y., 2013, MNRAS, 431, 2634
Chuang C.-H. et al., 2013, preprint (arXiv:1312.4889)
Chuang C.-H., Kitaura F. S., Prada F., Zhao C., Yepes G., 2015a, MNRAS, 446, 2621
Chuang C.-H. et al., 2015b, MNRAS, 452, 686
Cimatti A. et al., 2009, Exp. Astron., 23, 39
Cole S., Kaiser N., 1989, MNRAS, 237, 1127
Cole S. et al., 2005, MNRAS, 362, 505
Coles P., Jones B., 1991, MNRAS, 248, 1
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Cooray A., Hu W., 2001, ApJ, 554, 56
Cuesta A. J., Verde L., Riess A., Jimenez R., 2015, MNRAS, 448, 3463
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Dawson K. S. et al., 2013, AJ, 145, 10
de Jong R. S. et al., 2012, Proc. SPIE, 8446, 84460T
de la Torre S., Peacock J. A., 2013, MNRAS, 435, 743
de la Torre S. et al., 2013, A&A, 557, A54
Dekel A., Lahav O., 1999, ApJ, 520, 24
Desjacques V., Crocce M., Scoccimarro R., Sheth R. K., 2010, Phys. Rev. D, 82, 103529
Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537
Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007, ApJ, 664, 675
Eisenstein D. J. et al., 2011, AJ, 142, 72
Elia A., Ludlow A. D., Porciani C., 2012, MNRAS, 421, 3472
Fosalba P., Crocce M., Gaztañaga E., Castander F. J., 2015, MNRAS, 448, 2987
Frieman J., Dark Energy Survey Collaboration, 2013, in American Astronomical Society Meeting Abstracts, 221, p. 335.01
Fry J. N., Gaztanaga E., 1993, ApJ, 413, 447
Gil-Marín H. et al., 2015a, preprint (arXiv:1509.06386)
Gil-Marín H. et al., 2015b, preprint (arXiv:1509.06373)
Gil-Marín H., Noreña J., Verde L., Percival W. J., Wagner C., Manera M., Schneider D. P., 2015c, MNRAS, 451, 539
Guo Q., White S., Li C., Boylan-Kolchin M., 2010, MNRAS, 404, 1111
Hamaus N., Seljak U., Desjacques V., Smith R. E., Baldauf T., 2010, Phys. Rev. D, 82, 043515
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Heß S., Kitaura F. S., Gottlöber S., 2013, MNRAS, 435, 2065
Howlett C., Manera M., Percival W. J., 2015, Astron. Comput., 12, 109
Huterer D., Cunha C. E., Fang W., 2013, MNRAS, 432, 2945
Ishiyama T., Enoki M., Kobayashi M. A. R., Makiya R., Nagashima M., Oogi T., 2015, PASJ, 67, 61
Kaiser N., 1984, ApJ, 284, L9
Kalus B., Percival W. J., Samushia L., 2016, MNRAS, 455, 2573
Kazin E. A., Sánchez A. G., Blanton M. R., 2012, MNRAS, 419, 3223
Kim J., Park C., Choi Y.-Y., 2008, ApJ, 683, 123
Kim J., Park C., Gott J. R., III, Dubinski J., 2009, ApJ, 701, 1547
Kitaura F. S., Heß S., 2013, MNRAS, 435, L78
Kitaura F. S., Yepes G., Prada F., 2014, MNRAS, 439, L21
Kitaura F. S., Gil-Marín H., Scóccola C. G., Chuang C.-H., Müller V., Yepes G., Prada F., 2015, MNRAS, 450, 1836
Klypin A., Holtzman J., 1997, preprint (arXiv:e-prints)
Klypin A. A., Shandarin S. F., 1983, MNRAS, 204, 891
Klypin A., Hoffman Y., Kravtsov A. V., Gottlöber S., 2003, ApJ, 596, 19
Klypin A. A., Trujillo-Gomez S., Primack J., 2011, ApJ, 740, 102
Klypin A., Yepes G., Gottlober S., Prada F., Hess S., 2014, preprint (arXiv:1411.4001)
Koda J., Blake C., Beutler F., Kazin E., Marin F., 2015, preprint (arXiv:1507.05329)
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Laureijs R., 2009, preprint (arXiv:0912.0914)
Leauthaud A., Tinker J., Behroozi P. S., Busha M. T., Wechsler R. H., 2011, ApJ, 738, 45
Li C., White S. D. M., 2009, MNRAS, 398, 2177
Li Y., Hu W., Takada M., 2014a, Phys. Rev. D, 89, 083519
Li Y., Hu W., Takada M., 2014b, Phys. Rev. D, 90, 103530
LSST Dark Energy Science Collaboration 2012, preprint (arXiv:1211.0310)
McDonald P., Roy A., 2009, J. Cosmol. Astropart. Phys., 8, 20
Mandelbaum R., Seljak U., Kauffmann G., Hirata C. M., Brinkmann J., 2006, MNRAS, 368, 715
Manera M. et al., 2013, MNRAS, 428, 1036
Manera M. et al., 2015, MNRAS, 447, 437
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Mo H. J., White S. D. M., 2002, MNRAS, 336, 112
Mohayaee R., Mathis H., Colombi S., Silk J., 2006, MNRAS, 365, 939
Monaco P., Theuns T., Taffoni G., Governato F., Quinn T., Stadel J., 2002, ApJ, 564, 8
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Neyrinck M. C., 2013, MNRAS, 428, 141
Neyrinck M. C., 2016, MNRAS, 455, L11
Neyrinck M. C., Hamilton A. J. S., Gnedin N. Y., 2004, MNRAS, 348, 1
Neyrinck M. C., Aragón-Calvo M. A., Jeong D., Wang X., 2014, MNRAS, 441, 646
Norberg P., Baugh C. M., Gaztañaga E., Croton D. J., 2009, MNRAS, 396, 19
Nuza S. E. et al., 2013, MNRAS, 432, 743
Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A. J., Mehta K. T., Kazin E., 2012, MNRAS, 427, 2132
Peacock J. A., Heavens A. F., 1985, MNRAS, 217, 805
Percival W. J. et al., 2001, MNRAS, 327, 1297
Percival W. J., Verde L., Peacock J. A., 2004, MNRAS, 347, 645
Percival W. J. et al., 2014, MNRAS, 439, 2531
Prada F., Klypin A. A., Cuesta A. J., Betancort-Rijo J. E., Primack J., 2012, MNRAS, 423, 3018
Press W. H., Schechter P., 1974, ApJ, 187, 425
Reid B. A., White M., 2011, MNRAS, 417, 1913
Reid B. et al., 2016, MNRAS, 455, 1553

Rodríguez-Torres S. A. et al., 2015, preprint (arXiv:1509.06404)
Ross A. J., Brunner R. J., 2009, MNRAS, 399, 878
Ross A. J. et al., 2012, MNRAS, 424, 564
Ross A. J., Samushia L., Howlett C., Percival W. J., Burden A., Manera M., 2015, MNRAS, 449, 835
Saito S., Baldauf T., Vlah Z., Seljak U., Okumura T., McDonald P., 2014, Phys. Rev. D, 90, 123522
Saslaw W. C., Hamilton A. J. S., 1984, ApJ, 276, 13
Schlegel D., Abdalla F., Abraham T., Ahn C., Allende Prieto C., Annis J., Aubourg E. et al., 2011, preprint (arXiv:1106.1706)
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Seljak U., 2000, MNRAS, 318, 203
Seo H.-J., Eisenstein D. J., 2005, ApJ, 633, 575
Sheth R. K., 1995, MNRAS, 274, 213
Sheth R. K., Lemson G., 1999, MNRAS, 304, 767
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Sheth R. K., Chan K. C., Scoccimarro R., 2013, Phys. Rev. D, 87, 083002
Skibba R. A., Sheth R. K., 2009, MNRAS, 392, 1080
Skillman S. W., Warren M. S., Turk M. J., Wechsler R. H., Holz D. E., Sutter P. M., 2014, preprint (arXiv:1407.2600)
Smith R. E., Scoccimarro R., Sheth R. K., 2007, Phys. Rev. D, 75, 063512
Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, MNRAS, 320, 289
Springel V. et al., 2005, Nature, 435, 629
Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, MNRAS, 387, 1391
Szapudi I., Szalay A. S., 1998, ApJ, 494, L41
Takada M., Hu W., 2013, Phys. Rev. D, 87, 123504
Tasitsiomi A., Kravtsov A. V., Gottlöber S., Klypin A. A., 2004, ApJ, 607, 125
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, J. Cosmol. Astropart. Phys., 6, 36
Taylor A., Joachimi B., Kitching T., 2013, MNRAS, 432, 1928
Tinker J. L., 2007, MNRAS, 374, 477
Trujillo-Gomez S., Klypin A., Primack J., Romanowsky A. J., 2011, ApJ, 742, 16
Valageas P., Nishimichi T., 2011, A&A, 527, A87
Vale A., Ostriker J. P., 2004, MNRAS, 353, 189
Watson W. A., Iliev I. T., Diego J. M., Gottlöber S., Knebe A., Martínez-González E., Yepes G., 2014, MNRAS, 437, 3776
Wetzel A. R., White M., 2010, MNRAS, 403, 1072
White M., Song Y.-S., Percival W. J., 2009, MNRAS, 397, 1348
White M. et al., 2011, ApJ, 728, 126
White M., Tinker J. L., McBride C. K., 2014, MNRAS, 437, 2594
Zehavi I. et al., 2005, ApJ, 630, 1
Zentner A. R., Berlind A. A., Bullock J. S., Kravtsov A. V., Wechsler R. H., 2005, ApJ, 624, 505
Zhao C., Kitaura F. S., Chuang C.-H., Prada F., Yepes G., Tao C., 2015, MNRAS, 451, 4266
Zheng Z., Coil A. L., Zehavi I., 2007, ApJ, 667, 760
Zheng Z., Zehavi I., Eisenstein D. J., Weinberg D. H., Jing Y. P., 2009, ApJ, 707, 554

[1]*Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany*
[2]*Instituto de Física Teórica, (UAM/CSIC), Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*
[3]*Campus of International Excellence UAM+CSIC, Cantoblanco, E-28049 Madrid, Spain*
[4]*Departamento de Física Teórica, Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*
[5]*Tsinghua Center for Astrophysics, Department of Physics, Tsinghua University, Haidian District, Beijing 100084, China*
[6]*Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, E-18080 Granada, Spain*
[7]*Institute of Cosmology & Gravitation, University of Portsmouth, Dennis Sciama Building, Portsmouth PO1 3FX, UK*
[8]*Key Laboratory for Research in Galaxies and Cosmology of Chinese Academy of Sciences, Shanghai Astronomical Observatory, Shanghai 200030, China*
[9]*Department of Physics and Astronomy, University of Utah, UT 84112, USA*
[10]*Astronomy Department, New Mexico State University, Las Cruces, NM 88003, USA*
[11]*Severo Ochoa Associate Researcher at the Instituto de Fisica Teorica (UAM/CSIC), E-28049 Madrid, Spain*
[12]*Instituto de Astrofísica de Canarias (IAC), C/Vía Láctea, s/n, La Laguna, E-38200 Tenerife, Spain*
[13]*Dpto. Astrofísica, Universidad de La Laguna (ULL), E-38206 La Laguna, Tenerife, Spain*
[14]*Center for Cosmology and Particle Physics, New York University, 4 Washington Place, New York, NY 10003, USA*
[15]*Harvard–Smithsonian Center for Astrophysics, 60 Garden St., Cambridge, MA 02138, USA*
[16]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[17]*Departments of Physics and Astronomy, University of California, Berkeley, CA 94720, USA*
[18]*Max-Planck-Institut für extraterrestrische Physik, Postfach 1312, Giessenbachstr., D-85741 Garching, Germany*
[19]*Universitäts-Sternwarte München, Scheinerstrasse 1, D-81679 Munich, Germany*
[20]*Instituto de Física, Universidad Nacional Autónoma de México, Apdo. Postal 20-364, 01000 México, D.F., México*
[21]*Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, E-08028 Barcelona, Spain*
[22]*Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218, USA*
[23]*Center for Cosmology and AstroParticle Physics, The Ohio State University, Columbus, OH 43210, USA*

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

CHAPTER 5

Clustering of Quasars in the eBOSS-Y1Q

Publication: Monthly Notices of the Royal Astronomical Society, Volume 468, Issue 1, p.728-740

Motivation

Current and future surveys are mapping deep regions of the Universe. They are covering huge volumes which becomes a challenge for cosmological simulations. Quasars allow us to explore these regions. However, the connection between these objects and dark matter is not well understood. Most models analyse them at large scales. Here, we produce mock catalogues using a new method to describe the quasar-halo connection and learn more about the evolution of the structures shown by these quasars.

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **468,** 728–740 (2017)
Advance Access publication 2017 February 24

doi:10.1093/mnras/stx454

Clustering of quasars in the first year of the SDSS-IV eBOSS survey: interpretation and halo occupation distribution

Sergio A. Rodríguez-Torres,[1,2,3]★† Johan Comparat,[1,3]‡ Francisco Prada,[1,2,4] Gustavo Yepes,[3] Etienne Burtin,[5] Pauline Zarrouk,[5] Pierre Laurent,[5] ChangHoon Hahn,[6] Peter Behroozi,[7] Anatoly Klypin,[8,9] Ashley Ross,[10,11] Rita Tojeiro[12] and Gong-Bo Zhao[11,13]

[1]*Instituto de Física Teórica, (UAM/CSIC), Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*
[2]*Campus of International Excellence UAM+CSIC, Cantoblanco, E-28049 Madrid, Spain*
[3]*Departamento de Física Teórica M8, Universidad Autonóma de Madrid (UAM), Cantoblanco, E-28049, Madrid, Spain*
[4]*Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, E-18080 Granada, Spain*
[5]*CEA, Centre de Saclay, IRFU/SPP, F-91191 Gif-sur-Yvette, France*
[6]*Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003, USA*
[7]*Theoretical Astrophysics Center, Astronomy and Physics Departments, University of California, Berkeley, Berkeley CA 94720, USA*
[8]*Astronomy Department, New Mexico State University, Las Cruces, NM-88003, USA*
[9]*Severo Ochoa Associate Researcher at the Instituto de Física Teórica (UAM/CSIC), Cantoblanco, E-28049, Madrid, Spain*
[10]*Center for Cosmology and AstroParticle Physics, The Ohio State University, Columbus, OH 43210, USA*
[11]*Institute of Cosmology and Gravitation, Dennis Sciama Building, University of Portsmouth, Portsmouth PO1 3FX, UK*
[12]*School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews KY16 9SS, UK*
[13]*National Astronomy Observatories, Chinese Academy of Science, Beijing 100012, P.R. China*

Accepted 2017 February 20. Received 2017 February 17; in original form 2016 September 17

ABSTRACT

In current and future surveys, quasars play a key role. The new data will extend our knowledge of the Universe as it will be used to better constrain the cosmological model at redshift $z > 1$ via baryon acoustic oscillation and redshift space distortion measurements. Here, we present the first clustering study of quasars observed by the extended Baryon Oscillation Spectroscopic Survey. We measure the clustering of ∼70 000 quasars located in the redshift range $0.9 < z < 2.2$ that cover 1168 deg^2. We model the clustering and produce high-fidelity quasar mock catalogues based on the BigMultiDark Planck simulation. Thus, we use a modified (sub)halo abundance matching model to account for the specificities of the halo population hosting quasars. We find that quasars are hosted by haloes with masses $\sim 10^{12.7}$ M$_\odot$ and their bias evolves from 1.54 ($z = 1.06$) to 3.15 ($z = 1.98$). Using the current extended Baryon Oscillation Spectroscopic Survey data, we cannot distinguish between models with different fractions of satellites. The high-fidelity mock light-cones, including properties of haloes hosting quasars, are made publicly available.

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

1 INTRODUCTION

How quasars (QSO) populate the large-scale structure is a puzzle in modern cosmology. It is known that these objects trace the dark matter density field. Therefore, using measurements of the baryon acoustic oscillations (BAO) or redshift space distortions (RSD) from quasars, one can infer information of the cosmological model. However, for these studies or to increase the knowledge of the evolution of quasars, we require a good estimation of their distribution at all scales. Thus, spectroscopic surveys and high-fidelity galaxy mocks from simulations are a great help when solving many riddles concerning quasars.

Large galaxy spectroscopic surveys are an excellent tool to construct a precise 3D map of our Universe. They allow us to study the distribution of different populations in the Universe and constrain cosmological information via BAO scale or RSD measurements. The Sloan Digital Sky Survey (SDSS; York et al. 2000) and the two degree field galaxy redshift survey (Norberg et al. 2001) first measured the BAO scale in the local universe (Cole et al. 2005;

★ E-mail: sergio.rodriguez@uam.es
† Campus de Excelencia Internacional UAM/CSIC Scholar.
‡ Severo Ochoa Fellow.

Eisenstein et al. 2005). The Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013), included in the SDSS III program (Eisenstein et al. 2011), recently provided accurate redshifts for 1.5 million galaxies as faint as $i = 19.1$, that cover the redshift range $0.2 < z < 0.75$ on 10 000 deg^2. In combination with SDSS-I/II (York et al. 2000), it provided a subpercent level measurement of the position of the BAO peak at redshift $z = 0.57$ (Alam et al. 2016). SDSS is an example of how spectroscopic surveys can provide strong constraints on our knowledge of the Universe.

Bright quasars constitute the best targets to sample the matter field at high redshift with a small exposure time. Indeed, quasars bear an active galactic nucleus (AGN) that generates light which outshines the entire host galaxy. SDSS I/II published a sample of ∼100 000 confirmed quasars (Schneider et al. 2010) and SDSS-III observed ∼170 000 quasars with redshift $2.1 < z < 3.5$ as faint as $g = 22$ (Pâris et al. 2014). Using both samples, the BAO feature was measured to a few per cent in the Lyman α (Ly α) forest (Font-Ribera et al. 2014; Delubac et al. 2015). Despite the large sample of quasars observed by the SDSS programs, there is still a large region in redshift ($1 < z < 2.1$) that ought to be studied by targeting quasars fainter than $i = 19.1$ in the SDSS imaging. Recent data from other experiments (Wright et al. 2010, e.g. *WISE*) provides additional information to best target quasars. A cutting-edge target selection algorithm was implemented in Myers et al. (2015) and is being observed by the extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016), part of the SDSS-IV program. It will increase the number of quasars found by SDSS I/II in the redshift range $0.9 < z < 2.2$ by a factor of 5. This new sample will cover ∼7500 deg^2, increasing both the volume and the low number density of the previous samples. It is designed to measure the BAO scale with quasars as tracers of the matter field. In this study, we consider the eBOSS First Year QSO data (hereafter Y1Q). For more details, please see Section 2.1.

Different models have been used to analyse the clustering of quasars. In the literature, many studies focus on the linear regime (large scales). At these scales, correlation function can be described by a power law (e.g. Chehade et al. 2016), mostly due to the intrinsic low density of quasars. A more sophisticated method used to model the galaxy clustering and generate mock catalogues is the halo occupation distribution (HOD; Jing, Mo & Börner 1998; Peacock & Smith 2000; Scoccimarro et al. 2001; Berlind & Weinberg 2002; Cooray & Sheth 2002; Zheng et al. 2005). The HOD model recovers the quasar clustering, but its parameters are largely degenerate, producing poor constraints on the host halo masses and satellite fraction (Richardson et al. 2012; Shen et al. 2013). Galaxy samples have also been studied with another method, namely halo abundance matching (HAM), which reproduces the clustering of complete galaxy samples with a reasonable agreement (e.g. Kravtsov et al. 2004; Conroy, Wechsler & Kravtsov 2006; Behroozi, Conroy & Wechsler 2010; Guo et al. 2010; Trujillo-Gomez et al. 2011; Nuza et al. 2013; Reddick et al. 2013). By including the stellar mass distribution (or luminosity distribution), the HAM also accounts for incomplete samples (e.g. Rodríguez-Torres et al. 2016). HAM requires knowledge of the stellar mass function, the scatter in the stellar mass to halo mass relation and the incompleteness of the sample. In the case of quasars, obtaining such information is not an easy task. However, modifications of the standard method can be implemented to describe the quasar population.

In this study, we generate light-cones based on the BigMultiDark Planck simulation (BigMDPL; Klypin et al. 2016), using a modified HAM technique to reproduce the Y1Q clustering properties. The BigMDPL is an N-body simulation with box size 2.5 h^{-1} Gpc and 3840^3 particles, which yields a volume large enough to encompass Y1Q. A variety of mocks, which model different populations of galaxies, has already been constructed using the BigMDPL simulation. They predict, with a good agreement, the observed two-point and three-point statistics (Guo et al. 2015; Favole et al. 2016; Rodríguez-Torres et al. 2016).

This paper is structured as follows. In Section 2, we describe the data used in our analysis. Section 3 presents the different steps to construct the BigMDPL eBOSS quasar mocks, including how we populate dark matter haloes using a modified HAM algorithm. A set of predictions from our model is shown in Section 4. Subsequently, we discuss and summarize the most relevant results in Sections 5 and 6. In this paper, we assume a fiducial Λ cold dark matter (ΛCDM) cosmology with the Planck-I parameters $\Omega_{\rm m} = 0.307$, $\Omega_B = 0.048$, $\Omega_\Lambda = 0.693$ (Planck Collaboration XVI 2014).

2 DATA

2.1 eBOSS QSO survey and clustering

The eBOSS (Dawson et al. 2016) is part of a six year SDSS-IV programme (fall 2014 to spring 2020). It combines the potential of SDSS-III/BOSS and new photometric information to optimize target selection and extend BAO studies to higher redshift. eBOSS uses the 2.5-m Sloan Foundation Telescope at Apache Point Observatory (Gunn et al. 2006) and the same fibre-fed optical spectrograph as BOSS, where each fibre subtends a 2 arcsec diameter of the sky (Smee et al. 2013). This survey will provide redshifts for 300 000 luminous red galaxies (LRG) in the redshift range $0.6 < z < 1.0$, a new sample of ∼200 000 emission line galaxies (ELG) at redshift $z > 0.6$, more than 500 000 spectroscopically confirmed quasars at $0.9 < z < 2.2$ and ∼120 000 new Ly α forest quasars at redshift $z > 2.1$.

eBOSS dedicates 1800 plates to cover an area of 9000 deg^2: 1500 plates to measure LRG and QSO redshifts on 7500 deg^2 and 300 plates to measure ELG redshifts on 1000 deg^2. The first two years, observations were dedicated to the QSO and LRG samples. In order to maximize the tiling completeness and fibre efficiency in the LRG/QSO sample, a tiered priority is adopted (Dawson et al. 2016), where the QSO targets have maximal priority and are assigned to fibres first.

eBOSS has adopted two approaches to target quasars for redshift >0.9 (Myers et al. 2015). In the first approach, 'Clustering' quasar targets (QSO_CORE) are used as a direct tracer of the large-scale structure in the redshift range $0.9 < z < 2.2$. The second approach consists in detecting quasars at $z > 2.1$ to map the large-scale structure via absorption of the Ly α forest (Palanque-Delabrouille et al. 2016).

(i) The CORE quasar sample is constructed combining optical selection in *ugriz* using a likelihood-based routine called XDQSOz (Bovy et al. 2011), with a mid-IR–optical colour cut. eBOSS CORE selection (to $g < 22$ or $r < 22$) should obtains ∼70 quasars deg^{-2} at redshifts $0.9 < z < 2.2$ and about 7 quasars deg^{-2} at $z > 2.2$.

(ii) The Ly α quasar selection is based on variability in multi-epoch imaging from the Palomar Transient Factory (Palanque-Delabrouille et al. 2016). It recovers an additional 3 or 4 quasars deg^{-2} at $z > 2.2$ to $g < 22.5$. A linear model of how imaging systematics affect target density recovers the angular distribution of eBOSS CORE quasars over 96.7 per cent (76.7 per cent) of the SDSS North (South) Galactic Cap area (Myers et al. 2015).

Table 1. Distribution of the Y1Q sample in four redshift bins. $\bar{n}$ represents the comoving number density of QSO, N is the number of QSO and V is the comoving volume of the redshift bin subtended by 1168 deg^2. The last line shows the values for the complete sample.

Redshift	$\bar{n}$ (10^{-5} Mpc^{-3} h^3)	N	V (10^9 h^{-3} Mpc^3)
$0.9 < z < 1.2$	1.36	13 484	0.99
$1.2 < z < 1.5$	1.48	17 578	1.19
$1.5 < z < 1.8$	1.36	17 778	1.31
$1.8 < z < 2.2$	1.05	19 429	1.84
$0.9 < z < 2.2$	1.28	68 269	5.34

Busca et al. (2013) measure the BAO scale using Ly α quasars from the BOSS data. Font-Ribera et al. (2014) also give measurements of this scale using the cross-correlation between visually confirmed quasars with the Ly α forest absorption. One of the goals of eBOSS is to provide a first detection of the BAO scale using only the CORE quasar sample.

In this context, we focus our study on the spectroscopically confirmed QSO using the Y1Q data which includes 68 269 objects that cover 1168 deg^2 of the sky. Table 1 shows the abundance of CORE QSO at different redshift ranges.

2.2 Redshift error and statistical weights

eBOSS expects a redshift precision better than 300 s^{-1}km rms for the QSO CORE at $z < 1.5$ and better than [300+400(z −1.5)] km s^{-1} at $z > 1.5$ (Myers et al. 2015). It corresponds to redshift errors of the order of 1×10^{-3} for $z < 1.5$ and $\sim 5 \times 10^{-3}$ for larger redshift. These errors have an important impact on scales smaller than 10 h^{-1}Mpc (see Appendix B). For this reason, we add redshift errors to the mock catalogues using these upper limits. In addition, less than 1 per cent of the sample is expected to have catastrophic redshift errors.

In order to include the observed redshift precision in the light-cones, we model redshift errors using a Gaussian distribution with mean value $z_{\rm true}$ and width Δz,

$$z = z_{\rm true} + \Delta z \mathcal{N}(0, 1), \tag{1}$$

where $\mathcal{N}(0, 1)$ is a random number coming from a Gaussian distribution with mean 0 and standard deviation 1 and

$$\Delta z = \begin{cases} 300 \text{ km s}^{-1}c^{-1} & \text{if } z<1.5 \\ [300 + 400(z - 1.5)] \text{ km s}^{-1}c^{-1} & \text{if } z \geq 1.5, \end{cases} \tag{2}$$

c represents the speed of light. We also include 1 per cent of catastrophic redshift errors, which introduces a reduction in the amplitude of the correlation function of $\sim$1 per cent at all scales (Appendix B). In order to include these errors, we randomly select 1 per cent of the mock galaxies and replace their redshift by a random value within the range of the catalogue.

A correct estimation of redshift errors is important in order to understand the behaviour of the clustering at small scales. The monopole of the correlation function is affected by over 50 per cent at scales below 10 h^{-1} Mpc. The impact is larger on the quadrupole, where the effects are detected at scales below 40 h^{-1} Mpc (Reid & White 2011). In Appendix B, we explore with more detail the impact of these errors on clustering measurements. Nevertheless, even if we model the redshift errors, this is still an approximation that can introduce unphysical effects. This can result in a wrong estimation of the model's parameters if scales affected by errors are included in the fitting procedure. For this reason, we fix the parameters using the monopole of the correlation function between 10 and 40 h^{-1} Mpc, where the impact of redshift errors decreases and the effects of the cosmic variance and shot noise become smaller (Appendix B).

In addition to redshift measurement, the 5σ detection limit for point sources (also called depth) of the SDSS photometric survey varies across the footprint and differs for each band. The amplitude of the variations implies that faint targets end up very close to the detection limit. These targets are then more likely to be missed by the target selection algorithm. eBOSS corrects this effect by applying a depth-dependent weight, called 'systematics weight' $w_{\rm sys}$ to each quasar (see Laurent et al., in preparation for a detailed description).

Finally, eBOSS takes fibre collisions and redshift failures into account by using weights for each, w_{cp} and w_{zf}, respectively. Those quantities are initialized to one for all objects. Then, if a quasar has a nearest neighbour with a redshift failure or its redshift was not obtained because it was in a close pair, w_{zf} or w_{cp} are increased by one (Ross et al. 2012). Including all these effects, the total weight for each quasar in the observed data is given by

$$w_{\rm Q} = w_{\rm FKP} w_{\rm sys}(w_{cp} + w_{zf} - 1), \tag{3}$$

where $w_{\rm FKP}$ is the density weight applied for an optimal estimation of the two-point function and is defined by the expression (Feldman, Kaiser & Peacock 1994)

$$w_{\rm FKP} = \frac{1}{1 + n(z)P_{\rm FKP}}, \tag{4}$$

where $n(z)$ is the number density at redshift z and $P_{\rm FKP} = 6000\ h^{-3}$ Mpc^3.

Corrections for fibre collisions using close pair weights do not provide an accurate clustering signal at small scales (Guo, Zehavi & Zheng 2012; Hahn et al. 2017). However, in the quasar sample the distribution of objects is disperse and the number of collided pairs is very small. Additionally, our analysis does not use scales below 10 h^{-1}Mpc, so the close pair correction is good enough for our purpose. In the case of the simulated quasars, we include FKP weights but do not simulate the effects that require any of the additional weights applied to the data sample.

2.3 The eBOSS BigMultiDark light-cone

The suite of MultiDark[1] Planck (MDPL) simulations adopts a flat ΛCDM model with PLANCK-I cosmological parameters (Planck Collaboration XVI 2014): $\Omega_{\rm m} = 0.307$, $\Omega_B = 0.048$, $\Omega_\Lambda = 0.693$, $\sigma_8 = 0.829$, $n_s = 0.96$ and a dimensionless Hubble parameter $h = 0.678$. We only use two of the N-body simulations described in Klypin et al. (2016). The BigMultiDark (BigMDPL) has a box length of 2.5 h^{-1} Gpc with 3840^3 particles of mass 2.4 $\times$ 10^{10} h^{-1}M$_\odot$ and the MDPL has a box length of 1.0 h^{-1} Gpc with 3840^3 particles with a mass of 1.5×10^9 h^{-1} M$_\odot$. Both were built with GADGET-2 (Springel 2005) using initial Gaussian fluctuations generated with the Zel'dovich approximation at redshift 100.

From the dark matter catalogues of the simulation, haloes are defined with the Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement halo finder (ROCKSTAR; Behroozi, Wechsler & Wu 2013). Spherical dark matter haloes and subhaloes are identified using an approach based on adaptive hierarchical refinement of friends-of-friends groups in six-phase space dimensions

[1] http://www.multidark.org/

Table 2. Deviation from the mass function at redshift 0 for the MDPL and the BigMDPL simulations. The masses and the maximum circular velocities are the threshold above which the completeness in this box relative to the mass function is higher than the percentage given in the header (see equation 6). The corresponding number of particles is provided in brackets.

	$\log(M_{200c}(z)/\mathrm{M}_\odot)$				$V_{\rm max}$			
Fraction	80 per cent	90 per cent	95 per cent	97 per cent	80 per cent	90 per cent	95 per cent	97 per cent
				Central haloes				
MDPL	11.04 (71)	11.10 (82)	11.26 (119)	11.61 (266)	57.3	68.6	98.3	121.9
BigMD	12.22 (69)	12.28 (79)	12.32 (87)	12.36 (98)	131.0	145.9	201.6	299.3

and one time dimension. ROCKSTAR computes halo mass using spherical overdensities of a virial structure (Bryan & Norman 1998). Before calculating halo masses and circular velocities, the halo finder performs a procedure that removes unbound particles from the final mass of the halo.[2] We include observational effects and construct a catalogue with similar volume to the eBOSS sample, by making light-cones based on different snapshots of the BigMDPL simulation.

We perform the modified HAM by using the maximum circular velocity of the halo ($V_{\rm max}$) in order to link dark matter haloes and quasars. The maximum circular velocity is one of the best candidates for matching dark matter haloes and galaxies (Reddick et al. 2013). $V_{\rm max}$ can be related to the virial mass of the halo through a power law given by

$$V_{max} = \beta(z)[M_{vir}E(z)/(10^{12}h^{-1}M_\odot)]^{\alpha(z)} \quad (5)$$

where, $E(z) = \sqrt{\Omega_{\Lambda,0} + \Omega_{\rm m,0}(1+z)^3}$, $\log_{10}\beta(z) = 2.209 + 0.060a - 0.021a^2$ and $\alpha(z) = 0.346 - 0.059a + 0.025a^2$, with $a = 1/(1+z)$ the scalefactor (see Rodriguez-Puebla et al. 2016). There are better candidates to perform the matching between dark matter haloes and galaxies, such as, the maximum circular velocity along the whole history of the halo ($V_{\rm peak}$). However, the BigMDPL simulation has a small number of snapshots (4) in the quasar redshift range thus preventing a good estimation of quantities that are computed by tracing haloes between snapshots. For this reason, we use $V_{\rm max}$ to implement our model. Differences between $V_{\rm peak}$ and $V_{\rm max}$ become important in case of substructures, while the selection of host haloes is similar with both quantities. Reddick et al. (2013) show a significantly larger amount of subhaloes when $V_{\rm peak}$ is used rather than other quantities. However, in our model the impact of choosing $V_{\rm max}$ can be compensated by using the fraction of satellites as a free parameter. Furthermore, the poor information of the one halo term in the quasar sample and the large errors in observations will not allow us to distinguish which quantity performs the matching better.

Table 2 presents the deviation of each simulation from a model of the complete mass function (Comparat et al. 2017), which is obtained by fitting a data set that contains the complete part of each of the MultiDark Planck simulation (SMDPL, MDPL, BigMDPL, HMDPL). Masses in Table 2 fulfil the condition given by

$$N_{\rm sim}(M_{200} > M_i)/N_{\rm mod}(M_{200} > M_i) < \text{percentage}, \quad (6)$$

where $N_{\rm sim}$ is the number of objects in the simulation with M_{200} smaller than the threshold mass M_i and $N_{\rm mod}$ is the corresponding number of haloes in the model. Previous works showed that quasars live in haloes with masses of the order of $\log(M/\mathrm{M}_\odot) \sim 12.5$ (Shen et al. 2013; Chehade et al. 2016). Both simulations mentioned above are complete for this mass as is shown in Table 2. But depending on the dispersion of the distribution of haloes hosting QSO, a small fraction of haloes coming from the incomplete part of the simulation enter in the final mock. We quantify the effect of the resolution in our catalogues with the MDPL, where this effect is negligible thanks to its higher resolution. MDPL has enough resolution to cover the halo mass range for the QSO population. However, its volume is smaller than the one covered by eBOSS, so one cannot construct a complete light-cone without box replications. Furthermore, the shot noise from a mock using this volume is very large, due to the low number density of the observed sample. In Appendix A, we show this effect by comparing the mocks generated from both simulations.

We include the redshift evolution in the number density and of the clustering when constructing light-cones from the BigMDPL simulation. These light-cones cover the redshift range $0.9 < z < 2.2$ and 1,481.75 deg^2 of the sky, which is comparable with the area of Y1Q. The mocks are built with the SUrvey GenerAtoR code (SUGAR; Rodríguez-Torres et al. 2016). In this procedure, we use all available snapshots from the BigMDPL simulation, $z = 2.145$, 1.445,1, 0.8868. In order to analyse the effects of the incompleteness, we select only the closest snapshots from the MultiDark simulation ($z = 1.425$, 0.987, see Appendix A). We present results from three different light-cones, the first one uses a single set of parameters to describe the Y1Q (BigMDPL-QSO). The second one is obtained by fitting the clustering in four redshift bins with a different set of parameters (BigMDPL-QSOZ). The last light-cone uses a single set of parameters, but only host haloes are included (the fraction of substructures is equal to zero, BigMDPL-QSO-NSAT).

2.4 Galaxy mocks for QSO (GLAM)

In order to estimate the uncertainties in the clustering measurements, we use the GaLAxy Mocks (GLAM) scheme for the eBOSS quasar sample. For this application, GLAM implements a new parallel particle mesh method (Klypin & Prada 2017) to construct the dark matter density field and an optimization to populate the simulation with quasars (Comparat et al., in preparation). We run the SUGAR code to construct light-cones (Rodríguez-Torres et al. 2016). Errors are extracted from the covariance matrix of 1000 GLAM-QSO mocks which cover the same area as the data. They are computed using the diagonal terms, $\sigma_i(x_i) = \sqrt{C_{ii}}$, thus these errors correspond to one standard deviation (1σ) away from the mean value of the mocks. We use the covariance matrix estimator given by

$$C_{ij} = \frac{1}{n_s - 1}\sum_{k=1}^{n_s}\left(x_i^k - \mu_i\right)\left(x_j^k - \mu_j\right), \quad (7)$$

where n_s is the total number of mocks and the mean of each measurement is

$$\mu_i = \frac{1}{n_s}\sum_{k=0}^{n_s} x_i^k. \quad (8)$$

[2] http://www.cosmosim.org/

Using the covariance matrix from these mocks we perform the fitting with the χ^2 statistics,

$$\chi^2 = \sum_{ij} \left[x_i^d - x_i^m\right] C_{ij}^{-1} \left[x_j^d - x_j^m\right], \quad (9)$$

where x_i^m and x_i^d are the measurements from the model and the data in the bin i, respectively. χ^2 values presented in this work are computed from the monopole of the correlation function.

3 CLUSTERING MODEL

One of the best ways to study the observed clustering of a survey is to simulate not only the effect of the gravity on the dark matter but also on the baryonic matter. In this case, stellar physics should be included to provide a direct prediction of the relation between dark matter haloes and the galaxies and their evolution in time. This approach is undertaken by hydrodynamical simulations, that include galaxy formation processes, stellar physics and AGN feedback. EAGLE (Rahmati et al. 2015) and ILLUSTRIS (Sijacki et al. 2015) are two of the most recent realizations which predict a realistic distribution of galaxies and quasar populations. However, these simulations are constructed in rather small boxes of $\sim 75\,h^{-1}$ Mpc and this impedes studies of the large-scale structure. The large amount of computational resources required for a hydrodynamic simulation is prohibitive and the computation of volumes comparable to observations nearly infeasible.

An alternative approach, cheaper in computational time, is to use the dark matter only simulations and add galaxies in a statistical way. There are two widely used models based on these statistical relations. The first one is the HOD (e.g. Guo et al. 2014), which gives the probability, $P(N|M_{\rm h})$, that a halo of mass $M_{\rm h}$ hosts N galaxies. This probability is described by a fitting formula, which is fixed using the clustering measurements from the observational data. The second method to populate the dark matter haloes is the HAM (e.g. Reddick et al. 2013). This model assumes that the most massive galaxies populate the most massive haloes.

3.1 The modified SHAM model

Favole et al. (2016) introduced a modified (sub)halo abundance matching (SHAM), designed to reproduce the clustering of the BOSS ELG sample. They select haloes from the simulation using a probability function which is the sum of two terms corresponding to host and satellite haloes. This probability is a Gaussian function described by three parameters: the mean mass, the width of the distribution and the satellite fraction. This method is useful to describe incomplete samples, such as the Y1Q, which is not complete in halo mass or stellar mass whatsoever. In this paper, we use a similar model to study the clustering of quasars. Favole et al. (2016) use the virial mass of haloes to implement their method. Instead of that, we use $V_{\rm max}$ and assume that the distribution of haloes hosting quasars has a Gaussian shape. The most general model is split in central and satellite haloes as done in Favole et al. (2016). When a QSO is located in the centre of a host halo, it is denoted as a central QSO. The satellite fraction refers to the fraction of QSO living in a subhalo. This fraction does not represent systems of binary quasars. The central halo which is the counterpart of a satellite QSO can host another kind of galaxy.

In the case of quasars, we do not use the luminosity or the stellar mass of the observed sample. Our model only uses the $V_{\rm max}$ distribution of haloes, as done by Nuza et al. (2013). Rodríguez-Torres et al. (2016) extend the HAM technique implemented by Nuza et al. (2013) using the stellar mass function and modelling the incompleteness of the sample. In that study, galaxies are assigned to haloes via a standard HAM and then they are downsampled to obtain the observed stellar mass distribution. Here, we assume that the intrinsic scatter between quasars and dark matter haloes, plus the incompleteness of the sample will produce a $V_{\rm max}$ distribution with a Gaussian shape. Then, the model orders haloes by $V_{\rm max}$ and downsamples objects as done by Rodríguez-Torres et al. (2016).

3.2 Implementation

Assuming that the final $V_{\rm max}$ distribution of the simulated quasar catalogue is Gaussian, we need to construct a probability distribution function that selects haloes from the complete simulation based on this condition. In a general case, the $V_{\rm max}$ distribution of the final catalogue will be

$$\begin{aligned}\phi_{\rm QSO}(V_{\rm max}) &= \phi^s_{\rm QSO} + \phi^c_{\rm QSO} \\ &= P_s(V_{\rm max})\phi^s_{\rm sim}(V_{\rm max}) + P_c(V_{\rm max})\phi^c_{\rm sim}(V_{\rm max}) \\ &= \mathcal{G}_s(V_{\rm max}) + \mathcal{G}_c(V_{\rm max}),\end{aligned}$$

where $\phi^c_{\rm sim}$ and $\phi^s_{\rm sim}$ represent the $V_{\rm max}$ distribution of host haloes and subhaloes, respectively, $\mathcal{G}_c$ and $\mathcal{G}_s$ are Gaussian functions with mean $V_{\rm mean}$, standard deviation $\sigma_{\rm max}$ and each one is normalized using

$$\int \mathcal{G}_s(V_{\rm max}, z){\rm d}V_{\rm max} = N_{\rm tot}(z) f_{\rm sat}$$

$$\int \mathcal{G}_c(V_{\rm max}, z){\rm d}V_{\rm max} = N_{\rm tot}(z)(1 - f_{\rm sat}),$$

where $N_{\rm tot}(z)$ is the total number of quasars per redshift bin given by the observed number density.

In order to construct the probability distribution, we sort all haloes in the simulation and compute the maximum circular velocity function ($V_{\rm max}$) for subhaloes and host haloes separately. Using the fraction of satellites as a free parameter and the observed number density, we normalize the Gaussian distribution for central and satellite haloes. We split all haloes of the simulation in bins of $V_{\rm max}$ and compute the probability of assigning a quasar to a dark matter halo (central or satellite) per bin as

$$P_{s/c}(V_{\rm max}) = \frac{N^{\rm gaus}_{s/c}}{N^{\rm tot}_{\rm sub/host}}, \quad (10)$$

where $N^{\rm tot}_{\rm sub/host}$ is the total number of subhaloes/host haloes in the range $[V_{\rm max} - \Delta V_{\rm max}/2, V_{\rm max} + \Delta V_{\rm max}/2]$ and $N^{\rm gaus}_{s/c}$ is the number of satellite/central quasars necessary to produce the final Gaussian shape. Using equation (10), we downsample all haloes in the simulation to obtain the QSO mock catalogue.

Our model consists of five different parameters, the mean and standard deviation values for satellite and central distributions and the fraction of satellites. However, we assume the same mean and standard deviation for central and satellite quasars thus decreasing the number of parameters. In addition, the current data do not provide enough information at small scales ($<1.0\ h^{-1}$ Mpc) to extract precise information about the standard deviation of the distribution and the satellite fraction of the eBOSS QSO sample. For these reasons, our unique parameter to fit the clustering is the mean value of the distribution ($V_{\rm mean}$).

3.3 Parameters

The most general model is defined by three parameters. However, due to the poor information at small scales, we only use one free parameter ($V_{\rm mean}$) to describe the Y1Q sample. Fig. 2 presents the χ^2 maps we obtain for different combinations of the three parameters $V_{\rm mean}$, $\sigma_{\rm max}$ and $f_{\rm sat}$. We find the satellite fraction, $f_{\rm sat}$, to be degenerate with $V_{\rm mean}$ (left-hand panel of Fig. 2) and this degeneracy could be broken only with information from the one halo term. However, the current Y1Q data do not allow going to those scales. For this reason, we do not fix the number of satellites in two of the three mocks presented, which means that host haloes and subhaloes are not distinguished when the selection is implemented. In addition, just as Favole et al. (2016), we do not find a dependency of the clustering with the width of the Gaussian distribution ($\sigma_{\rm max}$). $\sigma_{\rm max}$ cannot be constrained with the current data as is shown in the right-hand panel of Fig. 2. In the mass regime where QSOs live, $\sigma_{\rm max}$ impacts the clustering at small scales (<0.5 h^{-1} Mpc), so it is not possible to constrain this parameter.

In the case of quasars, at scales larger than 1.0 h^{-1} Mpc, the clustering amplitude only depends on $V_{\rm mean}$. In order to fix $\sigma_{\rm max}$, we use previous results in the literature. The model shown in Chehade et al. (2016) is consistent with a width in $V_{\rm max}$ of $\sigma_{\rm max} = 45\,{\rm km\ s^{-1}}$. However, due to the resolution of BigMultiDark, we decrease this value to $\sigma_{\rm max} = 30\,{\rm km\ s^{-1}}$. If we use larger values of $\sigma_{\rm max}$, we will include a larger fraction of haloes from the incomplete mass region of the simulation. Fixing $\sigma_{\rm max} = 30\,{\rm km\ s^{-1}}$, we ensure that the BigMDP light-cones have only ∼2 per cent of haloes selected from regions where the incompleteness is greater than 10. Thus, we avoid including any unphysical effects coming from the low resolution of the simulation.

Thus, our model describes the quasar sample with a single parameter which is fixed by minimizing the χ^2 distribution. As mentioned previously, we use the monopole of the correlation function between 10 and 40 h^{-1} Mpc (10 data points shown in Fig. 1), thereby avoiding systematic effects that influence the clustering measurements at small scales. Varying $V_{\rm max}$, we find that the χ^2 distribution is well described by a quadratic function. This is used to find the parameter that best represents the data.

4 RESULTS

We compare the Y1Q 2-point correlation function (2PCF) with that of the mocks using the χ^2 statistics with 9 degrees of freedom (10 data points and 1 parameter). In order to compute the 2PCF, we use a modified version of the Correlation Utilities and Two-point Estimation code (CUTE; Alonso 2012). We first analyse the complete sample, using the clustering measurements in the redshift range $0.9 < z < 2.2$. We find the best value for the parameter $V_{\rm mean} = 341.2\,{\rm km\ s^{-1}}$, which corresponds to a sample of mock QSO with mean mass $\log[M_{200}/{\rm M}_\odot] = 12.66 \pm 0.16$. Fig. 1 presents the clustering measurements (2PCF and power spectrum) along with the prediction of the best-fitting mock light-cone. We find an excellent agreement between the data and the model for the studied scales.

When fitting is performed using the clustering of the complete redshift range, the evolution of the mass distribution is not taken into account. In order to investigate this effect, we divide the sample in four redshift bins and find the best parameter to match the clustering in each individual redshift range. It slightly improves the quality of the fits, presented in Table 3 which gives the best-fitting values of $V_{\rm mean}$ and their corresponding reduced χ^2.

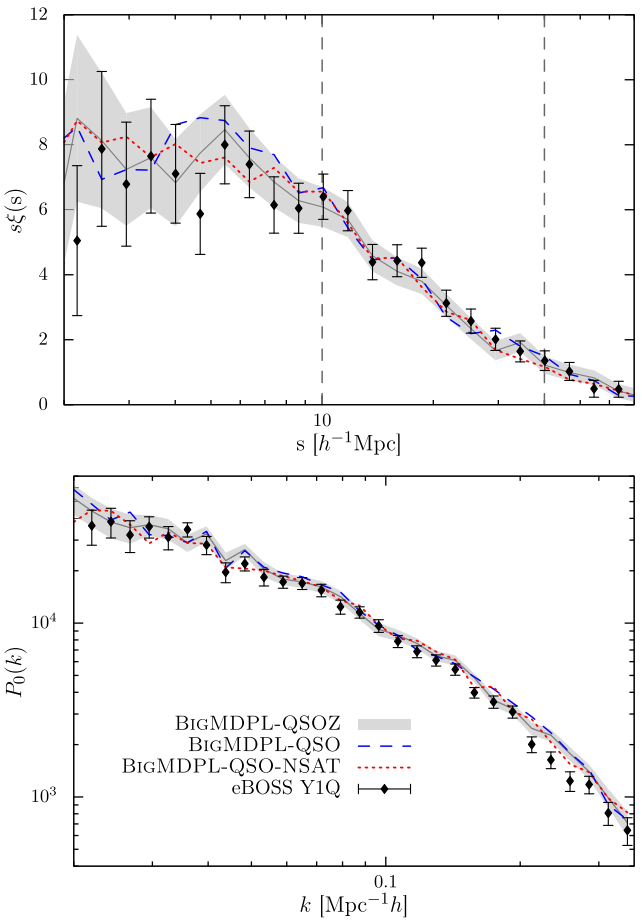


Figure 1. Top panel: monopole of the correlation function in configuration space of Y1Q (points with error bars). The shaded area represents the BigMDPL-QSOZ light-cone fitted in four different redshift bins. The dashed line represents the BigMDPL-QSO light-cone fitted on a single redshift bin and the dotted line is the BigMDPL-QSO-NSAT. The vertical lines represent the limit values used for fitting the parameters. Bottom panel: monopole of power spectrum of the Y1Q (points with error bars) and the three BigMDPL light-cone. The agreement between the best model and the data is remarkable. Error bars and dashed areas are computed using 1000 GLAM catalogues and correspond to 1σ deviation from the mean value. Differences at high k are due to redshift errors.

Table 3. Results of the fit per redshift bin. A gives the area in deg^2 subtended by the mock light-cone. z bin gives the lower and upper boundary of the redshift bin. $V_{\rm mean}$ is the best-fitting parameter found. $\log_{10}(M_{200}/{\rm M}_\odot)$ is the corresponding mean ± standard deviation of the halo mass of the population selected. χ^2_r is the reduced χ^2 per 9 degrees of freedom. We fixed $\sigma_{\rm max} = 30\,{\rm km\ s^{-1}}$ and $f_{\rm sat}$ is percentage of satellites in the catalogue.

A (deg^2)	z bin	$V_{\rm mean}$ (s^{-1}km)	$\log_{10}\frac{M_{200}}{{\rm M}_\odot}$	χ^2_r	$f_{\rm sat}$
		BigMDPL-QSO			
1481.75	0.9–2.2	341.2 ± 30.0	12.66 ± 0.16	1.78	5.3
		BigMDPL-QSOZ			
3275.06	0.9–1.2	282.8 ± 30.2	12.53 ± 0.17	1.47	9.0
2371.81	1.2–1.5	324.1 ± 30.1	12.63 ± 0.14	1.85	5.0
1879.13	1.5–1.8	339.5 ± 29.9	12.69 ± 0.14	1.70	4.3
1481.75	1.8–2.2	353.5 ± 29.7	12.60 ± 0.13	2.24	3.3
		BigMDPL-QSO-NSAT			
1481.75	0.9–2.2	349.5 ± 30.3	12.70 ± 0.16	1.52	0.0

Table 4. Mean halo mass and satellite fraction prediction from the BigMDPL light-cones.

Light-cone	$V_{\rm mean}$ (s^{-1}Km)	$\log_{10}[M_{200}/M_\odot]$	$f_{\rm sat}$
BigMDPL-QSOZ	326.9	12.61	0.048
BigMDPL-QSO	341.2	12.66	0.053
BigMDPL-QSO-NSAT	349.5	12.70	0.0

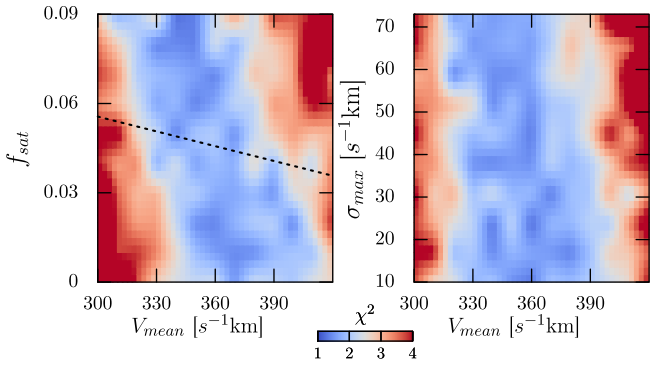


Figure 2. χ^2 maps for the three parameters of the model implemented on the BigMDPL-QSO. The left-hand panel shows the satellite fraction versus $V_{\rm mean}$. It is possible to note a degeneracy between both parameters. This is why we use the $f_{\rm sat}$ given by the simulation. The dashed line shows the satellite fraction given by the simulation for different values of $V_{\rm mean}$. The right-hand panel presents $\sigma_{\rm max}$ versus $V_{\rm mean}$. $\sigma_{\rm max}$ cannot be constrained using the current data.

Comparing the values of M_{200} presented in Table 3 with those of Table 2, we infer that the best-fitting mocks have less than 1 per cent of objects taken from a bin where the completeness is lower than 90 per cent. The effect of the resolution on the clustering is discussed in more detail in Appendix A.

Table 3 shows the values of satellite fractions of the BigMDPL light-cones. As we explained in Section 3.3, we do not use $f_{\rm sat}$ as a parameter so the fraction of satellites in the mock has the same dependency with $V_{\rm max}$ as the complete simulation. The third light-cone is the only catalogue where we fix $f_{\rm sat} = 0$. We include it to show the impact of removing all substructures from our analysis. The second parameter of the model, $\sigma_{\rm max}$ is also not constrained (see Fig. 2). A similar problem was found by Shen et al. (2013), their HOD parameters are largely degenerate and the fraction of satellites is not well constrained. For these reasons, we only vary the mean value of the Gaussian distribution ($V_{\rm mean}$) to fix the clustering of the model.

4.1 Trends of the QSO clustering with redshift

The signal of the quasar clustering does not have an important evolution, as shown in Fig. 3. The monopole varies mildly in the linear regime in all four redshift bins. If we assume a constant distribution of $V_{\rm max}$ for the whole redshift range, the evolution of the dark matter field will produce a non-constant signal of clustering in the different redshifts. In order to reproduce the observed evolution and predict a most realistic linear bias, we divide the complete redshift range into four regions, fitting the clustering of the light-cone in each bin. Table 3 presents the redshift range and the best-fitting parameters found to match the observed data. We use different areas for each redshift bin to maximize the volume

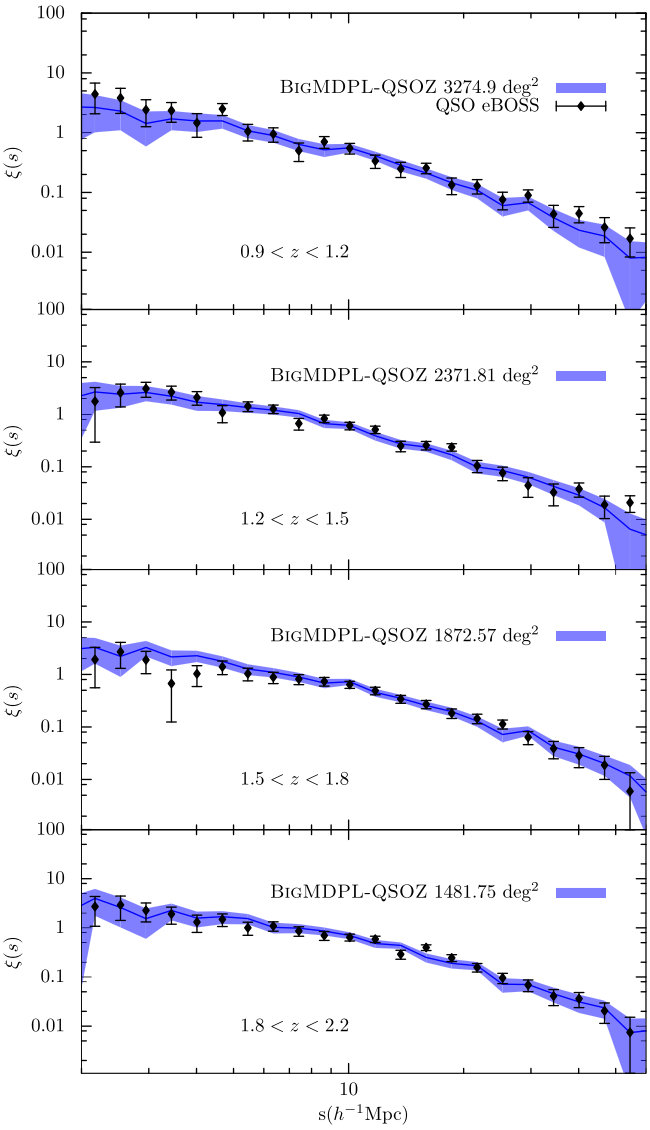


Figure 3. Monopole 2PCF versus redshift. We show the Y1Q (points) and the best-fitting mock (shaded area) of the BigMDPL-QSOZ light-cone (see Table 3). Each panel corresponds to a different redshift bin. Error bars and dashed areas are computed using 1000 GLAM catalogues and correspond to 1σ deviation from the mean value.

used from the simulation. These larger areas increase the statistics and reduce the shot noise in the 2PCF of the mocks as seen in Table 3.

Fig. 1 shows the monopole of the correlation function and the power spectrum of the three different mocks (BigMDPL-QSO/QSOZ/QSO-NSAT) compared to the observed data for the whole redshift range. All light-cones can reproduce the eBOSS data with a good agreement. We underline that the BigMDPL light-cones have shot noise and cosmic variance similar to the data. Due to these large errors in the model and the data, it is difficult to distinguish which light-cone reproduces the data better in the complete redshift range. However, if the model reproduces the clustering at different redshifts, we can estimate the evolution of the bias with better accuracy.

In order to quantify the difference between two models, we compare them using the Bayes factor. We can compute it with the

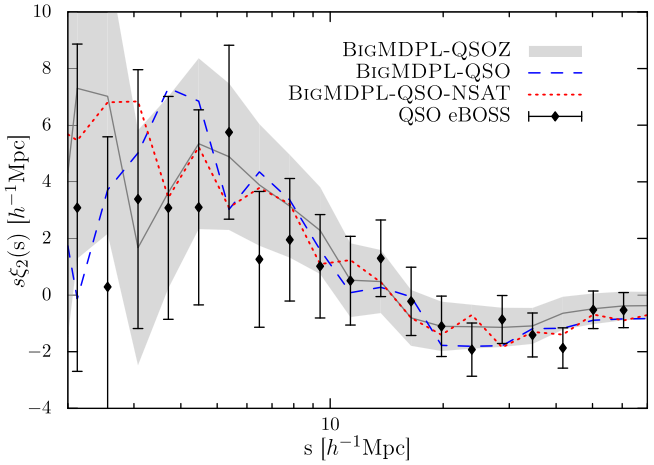


Figure 4. Quadrupole versus comoving scale in redshift space predicted by the BigMDPL-QSOZ (shaded region), BigMDPL-QSO (dashed line) and BigMDPL-QSO-NSAT (dotted lines) compared to the Y1Q (black points). All mocks are in agreement with observations. Error bars and shaded areas are computed using 1000 GLAM catalogues and correspond to 1σ deviation from the mean value.

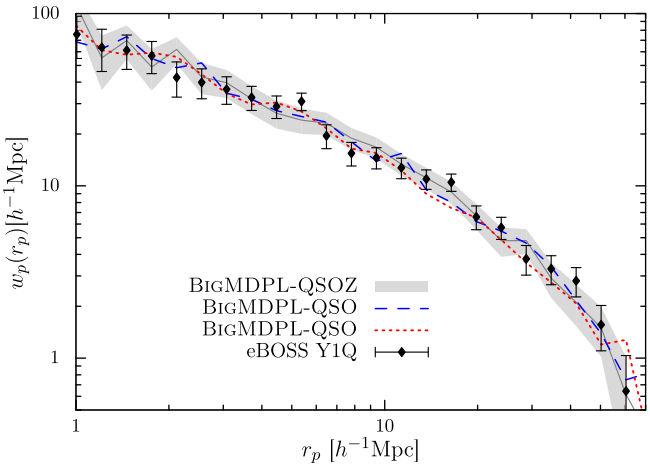


Figure 5. Projected correlation function predicted by the BigMDPL-QSOZ (shaded region), BigMDPL-QSO (dashed line) and BigMDPL-QSO-NSAT (dotted line) compared to the Y1Q (black points). The width of the shaded area represents 1σ errors computed with 1000 GLAM catalogues and correspond to 1σ deviation from the mean value. Our model reproduces the clustering for all relevant scales.

maximum likelihood

$$P(\mathbf{x}|\mathbf{p}) = \frac{|\tilde{\mathbf{C}}^{-1}|}{(2\pi)^p} \exp\left[-\frac{1}{2}\sum_{ij}\left(x_i^d - x_i(\mathrm{p})\right)\tilde{C}_{ij}^{-1}\left(x_j^d - x_j(\mathrm{p})\right)\right] \tag{11}$$

where x^d represents the data and x(p) the model. We estimate the inverse covariance matrix using equation (7) and correcting for bias using the Hartlap factor (Hartlap, Simon & Schneider 2007)

$$\tilde{C}_{ij}^{-1} = \frac{N_{\mathrm{mock}} - N_{\mathrm{p}} - 2}{N_{\mathrm{mock}} - 1} C_{ij}^{-1}, \tag{12}$$

where N_{p} represents the number of data points used. The Bayes factor between the BigMDPL-QSO and the BigMDPL-QSOZ model is

$$K = \frac{P(\xi_{\mathrm{data}}|\xi_{\mathrm{QSOZ}})}{P(\xi_{\mathrm{data}}|\xi_{\mathrm{QSO}})} = 5.45. \tag{13}$$

This result suggests that BigMDPL-QSOZ model is more substantially supported by the data than BigMDPL-QSO. The Bayes factor between the BigMDPL-QSOZ and the BigMDPL-QSO-NSAT is $K = 1.67$. In this case, we cannot conclude which model better reproduces the data. Furthermore, the BigMDPL light-cones have an important variability between realizations when the random seed is changed and it is not possible to construct a sufficient number of independent light-cones to make a definitive statement about the two models. In terms of χ^2 both light-cones are in agreement with the current data, though including a model with more parameters will improve the fitting of the data.

4.2 Checking $\xi_2(s)$ and $w_p(r_p)$

The quadrupole is very sensitive to processes affecting the small scales. Effects due to fibre collisions have an important impact at scales beyond the fibre size. However, the effect of fibre collisions is very small in the QSO sample. The most important observational effect is due to redshift errors, as shown in Appendix A. Fig. 4 shows the quadrupole of the BigMDPL-QSO, BigMDPL-QSOZ and BigMDPL-QSO-NSAT light-cones compared to the observations. All light-cones reproduce the data within 1σ error. This agreement suggests that we are using a reasonable model to account for redshift errors. We note that the BigMDPL-QSOZ light-cone reproduces the quadrupole better than the other two light-cones.

We compared the projected correlation function for the three light-cones and the observed data, finding a good agreement shown in Fig. 5.

The clustering predicted by the best-fitting model, which is mainly determined by the V_{mean}, reproduces with good agreement the two-point statistics of the observed data. We do not find significant differences between the three light-cones presented, all of them can reproduce the two-point statistics of the complete Y1Q sample with good agreement.

4.3 Bias

The Y1Q data allows for accurate measurements of the correlation function $\xi(r)$ and of the quasar bias b_{Q}, within the redshift range $0.9 < z < 2.2$. Laurent et al. (in preparation) obtain $b_{\mathrm{Q}} = 2.45 \pm 0.05$, when averaged over separations between 10 and 90 h^{-1} Mpc. This value is compatible with previous SDSS measurements, $b_{\mathrm{Q}}(z = 1.58) = 2.42 \pm 0.40$, by Ross et al. (2009).

We estimate the bias using the dark matter counterpart of the QSO mock light-cone. Using the autocorrelation of the dark matter sample, and the correlation function of the QSO mock in real space, we estimate the bias using

$$b(r)^2 = \frac{\xi(r)}{\xi_{\mathrm{DM}}(r)}. \tag{14}$$

Fig. 6 presents the bias of the BigMDPL-QSOZ and the BigMDPL-QSO compared to previous studies.

The bias measurements presented in Fig. 6 come from spectroscopically confirmed quasars in the two degree field (Porciani & Norberg 2006) at $0.8 < z < 2.1$, SDSS-I/II (Ross et al. 2009) at $z < 2.2$, the Quasar Dark Energy Survey pilot (2QDESp; Chehade et al. 2016) for redshift between 0.8 and 2.5 and the BOSS sample (Eftekharzadeh et al. 2015) at $2.2 < z < 2.8$. All these studies parametrize the real space correlation function by a power law,

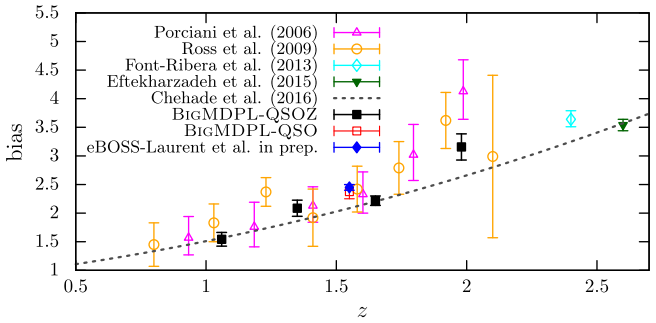


Figure 6. QSO bias as a function of redshift. The bias is computed using BigMDPL-QSOZ and BigMDPL-QSO light-cones. We include results from Chehade et al. (2016), Eftekharzadeh et al. (2015), Font-Ribera et al. (2014), Ross et al. (2009) and Porciani & Norberg (2006). eBOSS bias measurements are in agreement with previous results and about 10 times more precise. Results of eBOSS from Laurent et al. (in preparation) are also included.

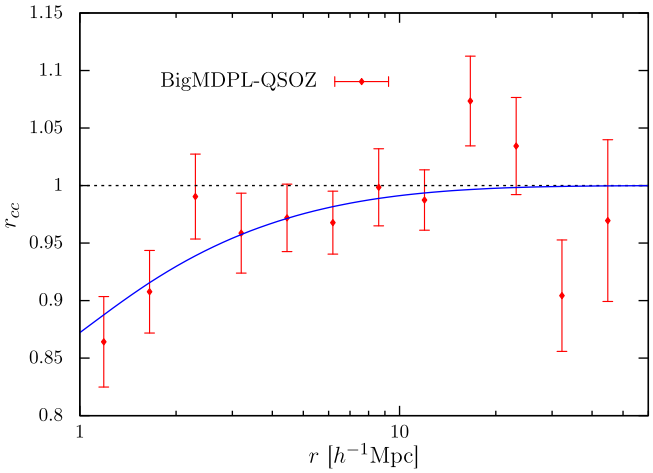


Figure 7. Cross-correlation coefficient between the dark matter field and the BigMDPL-QSOZ light-cone. The best model from (17) is shown with a solid line.

$\xi(r) = (r/r_0)^{\gamma}$, which can be related with the observed correlation function (redshift space) by

$$\xi(s) = \left(b_{\mathrm{Q}}^2 + \frac{2}{3}b_{\mathrm{Q}}f + \frac{f^2}{5}\right)\xi(r), \tag{15}$$

where $f = [\Omega_{\mathrm{m}}(z)]^{0.56}$ is the gravitational growth factor. In addition, we include measurements of quasars via Lyman α absorption at redshift 2.4 from the BOSS sample (Font-Ribera et al. 2014). Eftekharzadeh et al. (2015) also show a comparison between different estimations of the bias. At the redshifts studied, the bias measurements obtained in our study are in good agreement (see Fig. 6) and they are a factor 5 to 10 times more precise than previous studies.

4.4 Cross-correlation coefficients

The linear bias provides a good description of the relationship between dark matter and QSO mock in the linear regime. However, a single parameter b_{Q} is not enough to understand the link between galaxies and dark matter at all scales. To parametrize this relationship, we use the second-order bias, which is related to scales smaller than 10 h^{-1} Mpc. The second-order bias is inferred from the cross-correlation coefficient. It gives an estimation of the correlation between the positions of quasars and the dark matter field (Dekel & Lahav 1999). The cross-correlation, denoted r_{cc}, between quasars and the dark matter field is defined as

$$r_{\mathrm{cc}}(r) = \frac{\xi_{qm}(r)}{\sqrt{\xi_{qq}(r)\xi_{mm}(r)}}, \tag{16}$$

where q denotes the quasar sample and m the dark matter. r_{cc} is sensitive to the non-linear stochastic bias of the sample. Fig. 7 shows the cross-correlation coefficient between BigMDPL-QSOZ and the dark matter field. For scales larger than 10 h^{-1} Mpc, the cross-correlation function is consistent with 1. As expected, in this regime, we have $\xi_{gm} = b_{\mathrm{Q}}\xi_{mm}$ and $\xi_{gg} = b_Q^2\xi_{mm}$. At smaller separations, r_{cc} becomes smaller than one. This tendency is described in perturbation theory (Baldauf et al. 2010), where r_{cc} is described with the second-order bias by

$$r_{\mathrm{cc}}(r) \approx 1 - b_2^2\frac{\xi_{\mathrm{lin}}(r)}{4}, \tag{17}$$

where b_2 is the second-order bias and ξ_{lin} is the linear correlation function. The cross-correlation coefficient fit directly to the clustering by $b_2 = 0.314 \pm 0.030$. This relation is sufficient for the scales studied ($1 < rh^{-1}$ Mpc < 10), see the solid line in Fig. 7.

4.5 Halo occupation distribution

Table 5 shows the mean mass of haloes hosting quasars, the satellite fraction characterizes how quasars populate dark matter haloes and the mean value of V_{max} for all light-cones built in this study.

If the satellite fraction is not fixed (no distinction between haloes and subhaloes), we obtain a non-negligible fraction of satellites, ∼5 per cent. This value is consistent with Shen et al. (2013) which finds a satellite fraction of 6.8 per cent. However, due to the degeneracy between V_{mean} and f_{sat}, our model could also match the clustering with a negligible fraction (Fig. 2), as presented in Richardson et al. (2012).

Another way to formulate how QSO populate the density field is the probability of finding N quasars in a halo of mass M ($\langle N(M)\rangle$), namely the HOD model. This method describes how quasars would statistically populate haloes using a set of parameters fitted directly on the clustering. In SHAM models, $\langle N(M)\rangle$ is given by the halo catalogue by counting the total number of host haloes and the number of QSO per bin of mass. Fig. 8 shows the HOD predicted by the BigMDPL-QSO light-cone. We use this light-cone rather than the other as it has a negligible fraction of objects from the incomplete

Table 5. Mass prediction of haloes hosting quasars for different samples. It is presented with the name of the method used to analyse the sample and the used redshift range.

Sample	N_{QSO}	z	Method	$\log_{10}(M_h/\mathrm{M}_\odot)$
eBOSS[a]	68 269	0.9–2.2	HAM	12.5–12.82
SDSS-I/II[b]	8198	0.3–0.9	Power-law fit	12.75
SDSS-I/II[c]	48 000	0.4–2.5	HOD	12.70–12.77
BOSS[d]	27 129	2.2–2.8	Power-law fit	12.59–11.65
BOSS[e]	55 826	2.2–2.8	Power-law fit	11.63-12.63
2QDESp[f]	10 000	0.8–2.5	Power-law fit	12.17–12.64

Notes. [a]This work; [b]Shen et al. (2013); [c]Richardson et al. (2012); [d]White et al. (2012); [e]Eftekharzadeh et al. (2015); [f]Chehade et al. (2016).

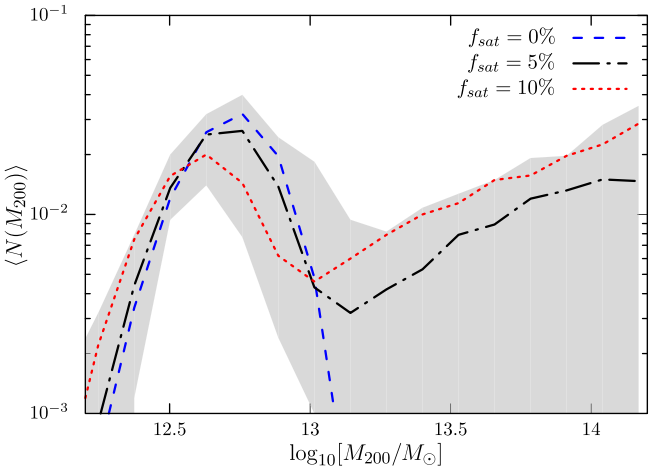


Figure 8. HOD for central plus satellites predicted from the BigMDPL-QSO light-cone. We present three light-cones using different fraction of satellites. The shaded area is computed adding 1σ error in the $V_{\rm mean}$ parameter for each light-cone. In addition, we vary the width of the distribution from 10 to 60 s^{-1} km to see the impact of this parameter in the HOD. $f_{\rm sat}$ is also changed from 0 to 0.12

part of the BigMDPL simulation. It also allows $\sigma_{\rm max}$ and $f_{\rm sat}$ to vary in a wide range, letting us show the dependency of $\langle N(M)\rangle$ on these parameters reflected in the different lines of Fig. 8.

Additionally, we construct light-cones with different $V_{\rm mean}$ including variations of 1σ from the best fit. We also vary the width of the distribution between 10 and 60 s^{-1} km. We do not use a larger $\sigma_{\rm peak}$, because we do not want to include a large fraction of objects coming from the incomplete part of the simulation. $f_{\rm sat}$ also varies between 0 and 10 per cent. The shaded area in Fig. 8 represents all HODs encompassed by these parameter variations.

Compared to previous HOD results (Shen et al. 2013), our model puts new constraints for masses below 10^{13} $M_{\odot}$. We find a distribution dominated by the mean halo mass of the sample. However, $\langle N(M)\rangle$ has a strong dependency with the other two parameters of the model, which we cannot constrain with the current data. An improvement on small scales of the QSO clustering or the cross-correlation between ELG and QSO in future surveys would constrain $\sigma_{\rm peak}$ and $f_{\rm sat}$ and therefore provide better HOD predictions.

5 DISCUSSION

Previous HOD analysis of the SDSS QSO sample combined different data sets to get more information about the distribution of QSOs inside haloes. However, due to large uncertainties in the data, the parameters of the HOD remain degenerate. eBOSS will greatly increase the statistical size of quasar samples, giving an excellent opportunity to learn more about this population and its connection with the dark matter. What we do here is to present the first study of the Y1Q clustering introducing a modified HAM that allows us to predict the HOD, masses of the dark matter haloes and the bias of the sample.

Several studies have provided information about quasars at different redshifts using their clustering measurements. Richardson et al. (2012) study the clustering of the 48 000 QSO from the SDSS sample in the redshift range $0.4 < z < 2.5$. They interpret the measurements of the projected correlation function at redshift 1.4. In addition, 4426 spectroscopically identified quasars in the redshift interval $2.9 < z < 5.4$ (Shen et al. 2007) are used to study the small-scale clustering. However, they use a regular HOD without including a duty cycle. For this reason, their parameters reproduce the clustering, but most of them are unphysical. Shen et al. (2013) study the two-point cross-correlation function of 8198 SDSS QSO and 349 608 BOSS CMASS galaxies in the redshift range $0.3 < z < 0.9$. They provide predictions of the HOD from quasars. However, the large degeneracies of the parameters make it impossible to have a well-constrained HOD. The BOSS sample provides a set of CORE QSO which is studied by Eftekharzadeh et al. (2015). They extend the analysis of the projected correlation function of the BOSS sample done by White et al. (2012). In that analysis, ∼70 000 quasars in the redshift range 2.2–3.4 are studied. In a more recent study, Chehade et al. (2016) combine the optical photometry of the 2dF Quasar Dark Energy Survey pilot (2QDESp) and the bands of the *Wide-field Infrared Survey Explorer* (*WISE*) to provide a sample of ∼10 000 QSO in the redshift range 0.8–2.5. Our study uses a larger and wider QSO sample than in previous works. It allows us to have a good estimation of the clustering in the redshift range $0.9 < z < 2.2$.

The mean mass of haloes hosting quasars has been measured by different methods finding a reasonable agreement between their results. However, the range of masses cover by quasars is still not well constrained. Richardson et al. (2012) predict a mean halo mass for central haloes $M_{\rm cen} \sim 10^{12.77}$ $M_{\odot}$ with a small fraction of QSO satellites, 7.4×10^{-4}. This result is in agreement with the BigMDPL-QSO-NSAT, which provides host halo masses for quasars of $10^{12.7\pm0.16}$ $M_{\odot}$. Shen et al. (2013) model the cross-correlation between CMASS galaxies and QSO by a power law, $\xi_{\rm QG} = (r/r_0)^{\gamma}$, with $r_0 = 6.61 \pm 0.25\ h^{-1}$ Mpc and $\gamma = 1.69 \pm 0.07$ for scales $r = 2$–$25\ h^{-1}$ Mpc. They find a characteristic mean halo mass of $10^{12.8}$ $M_{\odot}$. In contrast to Richardson et al. (2012), a non-negligible satellite fraction is predicted by Shen et al. (2013). They find that 6.8 per cent of QSO are hosted by subhaloes. This result is in better agreement with our mocks without fixing the fraction of satellites, which predict ∼5 per cent of quasars living in subhaloes. The halo masses predicted by this HOD are also in agreement within 1σ errors with our measurements. Nevertheless, they have larger degeneracies between their parameters. From the BOSS sample, White et al. (2012) find the quasar halo masses covering a wide mass range between $10^{11.59}$ and $10^{12.65}$ $M_{\odot}$. Just as in the previous cases, these values of masses are still in agreement with our results shown in Table 3. The Chehade et al. (2016) results are compared with other surveys (SDSS, 2QZ and 2SLAQ). As in previous works, they find no evidence of a dependency between the clustering and the luminosity of the QSO. In addition, they show that quasar clustering depends on redshift, in particular, when BOSS data are included. They describe the clustering of the sample using a power law, where $r_0 = 7.3 \pm 0.1\ h^{-1}$ Mpc at redshift 2.4, while the correlation scale for the whole redshift range is $r_0 = 6.1 \pm 0.1\ h^{-1}$. Their measurements are consistent with host haloes masses of $\sim 10^{12.46}$. Future observations will allow cross-correlations between ELGs and quasars, which will enable a better understanding of the distribution of quasars within the dark matter halo. These measurements could fix the satellite fraction of quasars. However, the width of the distribution is more difficult to constrain. In the similar case of ELG, Favole et al. (2016) faced an equivalent problem to describe their clustering. They use constraints from lensing measurements to understand the clustering on the smallest scales. Unfortunately, such measurements are not available for quasars.

Using our model, the signal of the clustering in the linear regime is dominated by the mean halo mass of the distribution. This is clear in the HOD (Fig. 8), where the distribution has a strong peak near the mean halo mass of the sample. We find a more constrained HOD region for quasars than Shen et al. (2013). However, more information from small scales is needed to have better constraints in the satellite fraction and width of the distribution in order to provide more realistic uncertainties. We find a bias equal to 2.37 ± 0.12 for the redshift range $0.9 < z < 2.2$, which is in good agreement with previous analysis and with eBOSS data from Laurent et al. (in preparation, Fig. 6). We provide measurements for the evolution of the bias using the BigMDPL-QSOZ light-cone, finding that the eBOSS quasars are in agreement with b_Q=1.54, 2.08, 2.21, 3.15 for redshift 1.06, 1.35, 1.65, 1.98. Furthermore, to give a complete parametrization of the scales studied in this work, we calculate the second-order bias from the cross-correlation coefficients, finding $b_2 = 0.314 \pm 0.030$. Table 5 presents a comparison of the halo mass predictions of previous studies and our result.

6 SUMMARY

We modelled the clustering of ∼70 000 optical quasars from the eBOSS Y1Q CORE sample in the redshift range $0.9 < z < 2.2$. We used a modified HAM that takes into account the incompleteness of the QSO sample and the intrinsic scatter between QSOs and dark matter haloes. This model was implemented in a light-cone constructed from a 2.5 h^{-1} Gpc simulation, covering an area comparable to the eBOSS Y1Q sample.

Our main results can be summarized as follows.

(i) We assume that the $V_{\rm max}$ distribution of haloes hosting QSOs is described by a Gaussian function which is defined by its mean and width plus one parameter for the satellite fraction. The current observations do not bear information on small-scale clustering. For this reason, we cannot constrain the fraction of satellites. Hence, we do not distinguish between host and subhaloes when the selection is done. The final mock thus has the same fraction of satellites as the complete simulation in the mass range used.

(ii) We model the clustering of the Y1Q using a single free parameter ($V_{\rm mean}$). The width of the Gaussian distribution is fixed to 30 s^{-1} km and we only impose a value to the satellite fraction in the BigMDPL-QSO-NSAT light-cone, for the other light-cones we do not fix this parameter.

(iii) The prediction of our model is in a good agreement with the 2PCF and the monopole of the power spectrum of the Y1Q data. The light-cone is constructed assuming Gaussian redshift errors given by Dawson et al. (2016). Their modelling improves the agreement between our model and the data. It provides a good description of the observed clustering on small scales, which is very sensitive to variations caused by these errors.

(iv) We construct three kinds of light-cones: one including the evolution of the parameters with redshift (BigMDPL-QSOZ), another describing the whole redshift range with a single parameter (BigMDPL-QSO) and a third one fixing the satellite fraction to zero (BigMDPL-QSO-NSAT). The mean halo masses are $10^{12.61}$, $10^{12.66}$ and $10^{12.70}$ $M_\odot$, respectively.

(v) Using the Bayes factor, we find a strong evidence that the BigMDPL-QSOZ (four parameters) reproduces the data better than the BigMDPL-QSO (one parameter). However, we cannot make the same conclusion with the model without satellites, which reproduces the data with a similar agreement to the BigMDPL-QSOZ model.

(vi) We find a mean bias of the Y1Q sample equal to 2.37 ± 0.12 and a second-order bias $b_2 = 0.314 \pm 0.030$, which both describe the relation between the dark matter and the QSO mock for the studied scales.

BigMDPL-QSOs and GLAM-QSO eBOSS mocks are publicly available through the *Skies and Universes* website.[3]

ACKNOWLEDGEMENTS

SRT is grateful for support from the Campus de Excelencia Internacional UAM/CSIC.

SRT, JC, FP acknowledge support from the Spanish MICINN Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064 MINECO Severo Ochoa Award SEV-2012-0249 and grant AYA2014-60641-C2-1-P.

GY acknowledges financial support from MINECO/FEDER (Spain) under research grants AYA2012-31101 and AYA2015-63810-P.

The BIGMULTIDARK simulations have been performed on the SuperMUC supercomputer at the Leibniz-Rechenzentrum (LRZ) in Munich, using the computing resources awarded to the PRACE project number 2012060963. The authors want to thank V. Springel for providing us with the optimised version of GADGET-2.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University and Yale University.

SRT thanks Sylvie Adenis for her help improving the grammar and the style of this paper.

REFERENCES

Alam S. et al., 2016, MNRAS, preprint (arXiv:1607.03155)
Alonso D., 2012, preprint (arXiv:1210.1833)
Baldauf T., Smith R. E., Seljak U., Mandelbaum R., 2010, Phys. Rev. D, 81, 063531

[3] http://projects.ift.uam-csic.es/skies-universes

Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Bovy J. et al., 2011, ApJ, 729, 141
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Busca N. G. et al., 2013, A&A, 552, A96
Chehade B. et al., 2016, MNRAS, 459, 1179
Cole S. et al., 2005, MNRAS, 362, 505
Comparat J., Prada F., Yepes G., Klypin A., 2017, preprint (arXiv: 1702.01628)
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Dawson K. S. et al., 2013, AJ, 145, 10
Dawson K. S. et al., 2016, AJ, 151, 44
Dekel A., Lahav O., 1999, ApJ, 520, 24
Delubac T. et al., 2015, A&A, 574, A59
Eftekharzadeh S. et al., 2015, MNRAS, 453, 2779
Eisenstein D. J. et al., 2005, ApJ, 633, 560
Eisenstein D. J. et al., 2011, AJ, 142, 72
Favole G. et al., 2016, MNRAS, 461, 3421
Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23
Font-Ribera A. et al., 2014, J. Cosmology Astropart. Phys., 5, 027
Gunn J. E. et al., 2006, AJ, 131, 2332
Guo Q., White S., Li C., Boylan-Kolchin M., 2010, MNRAS, 404, 1111
Guo H., Zehavi I., Zheng Z., 2012, ApJ, 756, 127
Guo H. et al., 2014, MNRAS, 441, 2398
Guo H. et al., 2015, MNRAS, 453, 4368
Hahn C., Scoccimarro R., Blanton M. R., Tinker J. L., Rodriguez-Torres S., 2017, MNRAS, 467, 1940
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Jing Y. P., Mo H. J., Börner G., 1998, ApJ, 494, 1
Klypin A., Prada F., 2017, preprint (arXiv:1701.05690)
Klypin A., Yepes G., Gottlober S., Prada F., Hess S., 2016, MNRAS, 457, 4340
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Myers A. D. et al., 2015, ApJS, 221, 27
Norberg P. et al., 2001, MNRAS, 328, 64
Nuza S. E. et al., 2013, MNRAS, 432, 743
Palanque-Delabrouille N. et al., 2016, A&A, 587, A41
Pâris I. et al., 2014, A&A, 563, A54
Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
Planck Collaboration XVI, 2014, A&A, 571, A16
Porciani C., Norberg P., 2006, MNRAS, 371, 1824
Rahmati A., Schaye J., Bower R. G., Crain R. A., Furlong M., Schaller M., Theuns T., 2015, MNRAS, 452, 2034
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Reid B. A., White M., 2011, MNRAS, 417, 1913
Richardson J., Zheng Z., Chatterjee S., Nagai D., Shen Y., 2012, ApJ, 755, 30
Rodriguez-Puebla A., Behroozi P., Primack J., Klypin A., Lee C., Hellinger D., 2016, MNRAS, 462, 893
Rodríguez-Torres S. A. et al., 2016, MNRAS, 460, 1173
Ross N. P. et al., 2009, ApJ, 697, 1634
Ross A. J. et al., 2012, MNRAS, 424, 564
Schneider D. P. et al., 2010, AJ, 139, 2360
Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, 20
Shen Y. et al., 2007, AJ, 133, 2222
Shen Y. et al., 2013, ApJ, 778, 98
Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575
Smee S. A. et al., 2013, AJ, 146, 32
Springel V., 2005, MNRAS, 364, 1105
Trujillo-Gomez S., Klypin A., Primack J., Romanowsky A. J., 2011, ApJ, 742, 16
White M. et al., 2012, MNRAS, 424, 933
Wright E. L. et al., 2010, AJ, 140, 1868
York D. G. et al., 2000, AJ, 120, 1579
Zheng Z. et al., 2005, ApJ, 633, 791

APPENDIX A: SIMULATION RESOLUTION

In order to reproduce the observed clustering of QSO or ELG samples, simulations with large volume and a high resolution are needed to resolve haloes of masses $\sim 10^{12.5}\,M_\odot$. The Y1Q sample covers $\sim$1100 deg^2 of the sky. This area is comparable to the BigMDPL-QSO light-cone. However, a small part of the halo mass range occupied by quasars can be in the incomplete part of the simulation.

We use the 1 h^{-1} Gpc MDPL simulation to quantify the effect of incompleteness of the BigMDPL light-cone. We select two snapshots from each simulation with similar redshift (Table A1). We apply the model using the parameters of Table 3. Table A1 presents a comparison between both simulations. In terms of halo mass, mocks constructed with both simulations provide consistent mean halo masses. Similar results are found for the satellite fraction.

In terms of clustering, both simulations give coherent results with differences of the order of 3 per cent. Fig. A1 shows the difference on the monopole between both simulations. These discrepancies are not a problem for our analysis, where errors from the data are of the order of 15 per cent.

In addition to the large errors in the data, discrepancies between both boxes seem reasonable if we notice the other sources of error.

(i) Both simulations have different initial conditions, this includes variations due to the cosmic variance between simulations.

Table A1. Comparison of the halo mass of mocks constructed with the BigMDPL and MDPL simulations. For comparison, all snapshots of the BigMDPL simulation in the redshift range $0.9 < z < 2.2$ were used. We select snapshots with the nearest redshift from the MDPL simulation.

Box	z	$\log_{10}[M/M_\odot]$	$V_{\rm mean}$	$f_{\rm sat}$
MDPL	0.987	12.41	284.25	0.08
	1.425	12.54	325.95	0.07
BigMDPL	1.000	12.40	284.25	0.11
	1.445	12.55	325.95	0.07

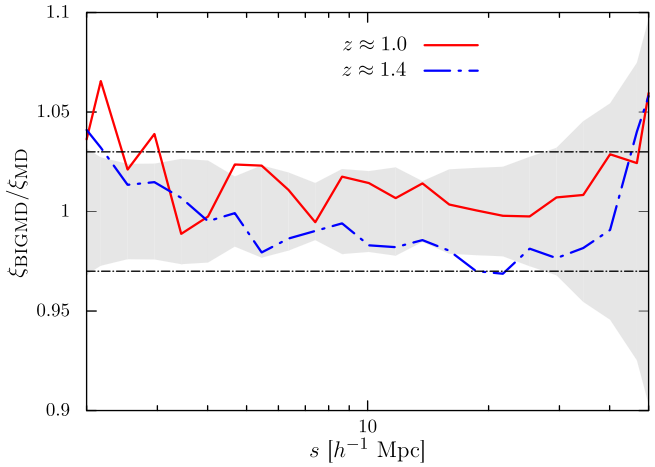


Figure A1. Ratio between BigMDPL and MDPL mocks of the monopole of the correlation function in configuration space. The horizontal lines represent 3 per cent differences. The shaded area shows 1σ dispersion due to the random selection in the MDPL boxes. We use 15 realizations to compute the shaded area.

(ii) The shot noise in the correlation function is larger in the MDPL simulation due to the smaller volume.

(iii) The random selection of our model is another source of errors. The shaded area in Fig. A1 represents the 1σ dispersion of 15 mocks produced with different seeds.

(iv) The BigMDPL simulation includes long waves that are not included in the 1 h^{-1} Gpc box size.

APPENDIX B: EFFECTS OF OBSERVATIONAL ERRORS ON THE CLUSTERING

The model presented in this work includes two observational errors: catastrophic redshift errors and redshift errors. The first errors cause a constant reduction in the clustering amplitude at all the scales. Fig. B1 shows the effect of applying 1 per cent of catastrophic redshifts. We find a reduction of $\sim$1 per cent in all scales of the correlation function in configuration space.

Redshift errors have the strongest impact on the clustering. The selection of QSO implies fixing maximum width (precision) to identify the emission/absorption features of the spectra. We introduce the effect of this tolerance using Gaussian errors with a width given by Dawson et al. (2016). Redshift errors have an important impact at scales $<10\,h^{-1}$ Mpc. In Fig. B1, it is possible to see a disagreement larger than 40 per cent, which cannot be explained by statistical errors of the sample (shaded area in Fig. B1).

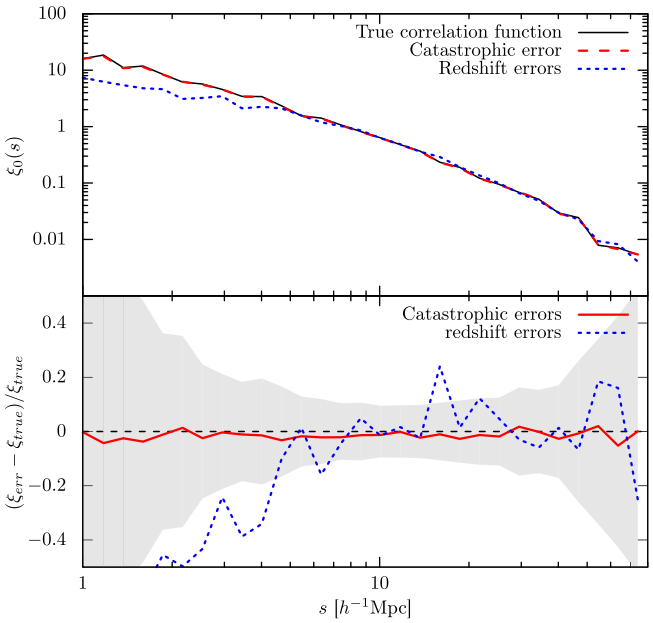


Figure B1. Top panel: impact of catastrophic redshift errors and redshift errors on the monopole of the correlation function. A light-cone reproducing the Y1Q 1-point and two-point statistics is used for this comparison. Bottom panel: normalized differences between mocks including redshift errors (blue dotted line) and catastrophic redshift errors (red line) with a model without errors. The shaded area represents the statistical errors in the light-cone computed from 1000 GLAM catalogues. Differences due to catastrophic redshift errors are $\sim$1 per cent. Redshift errors have an important impact at small scales which cannot be explained by uncertainties from mocks.

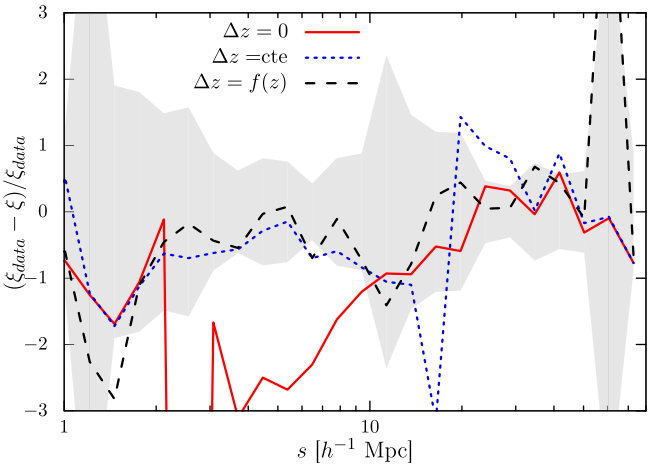


Figure B2. Impact of redshift errors in the quadrupole of the correlation function in configuration space. Lines show the normalized difference between observed data and model without redshift errors (red solid line), constant redshift error $\Delta z = 0.005$ (blue dotted line) and including redshift errors given by equation (2) (black dashed line). Shaded area represent 1σ error computed with 1000 GLAM catalogues for one light-cone.

The impact of redshift error is very important in the monopole of the correlation function. However, the effects on the quadrupole are larger. Fig. B2 shows the ratio of quadrupole from the observed data and the different mocks. The model introduced in this work describes the very large difference found between our mock and the observed data.

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

Conclusions

This thesis was done within the BOSS and eBOSS projects which are part of the SDSS III and IV programs respectively. It is the result of numerous discussions and contributions with other members of the collaborations. In the case of BOSS, we focus our research on the Luminous Red Galaxy population (LRG). However, while working on my PhD I also collaborated in studies on Emission Line Galaxies (ELG). For eBOSS, we focus our work on the clustering of quasars. It is important to recall that our clustering models can also be applied to other eBOSS galaxy populations such as LRG and ELG. Furthermore, as a member of the MultiDark simulation project, I have had access to a set of the top N-body simulations currently available.

Chapter 3 presents the Halo Abundance Matching (HAM) model used to describe the LRG BOSS population. This model has been implemented in the BigMultiDark Planck1 simulation in order to describe the redshift range $0.43 < z < 0.7$ of the CMASS sample. In this project, we include observational effects, such as the stellar mass incompleteness of the data, and features of the observations, such as fiber collisions, angular completeness, radial selection function or geometry of the survey. We also include the evolution of dark matter halos using light-cones constructed from the simulations. All these ingredients enable us to produce simulated galaxy catalogues which reproduce the observed sample with high precision.

Our work results from combining a simulation run with the best estimation of the cosmological parameters and features of the survey that can have a direct impact in the observational measurements. In spite of the simplicity of the model, its results have been very successful because we can generate galaxy catalogues that reproduce the observed two- and three-point correlation functions and their analogues in Fourier-space, the power spectrum and bispectrum.

In comparison with previous studies, our work shows the important impact of the observational effects and the cosmological parameters on the clustering signal. Results of this model

prove that the best ΛCDM model that best fit the CMB anisotropies (Planck Collaboration et al., 2014) predicts with an excellent agreement the clustering of LRG at $z \sim 0.5$. Future surveys will increase the accuracy of the clustering measurements and they will allow us to extend and test our current models including effects such as the assembly bias.

In addition to the clustering measurements, we show the prediction of the stellar to halo mass relation, which is in agreement with weak lensing data. All these results make our model a very useful tool for the study of observations and the analysis of systematic effects on the surveys. Using this model we collaborated with different research groups analysing different observations.

Numerical simulations allow us to compare theoretical models and observations directly including non-linear effects. If a model reproduces the large-scale structure with a good agreement, it is possible to make predictions from the simulation, such as the characteristic halo mass of a given galaxy population. Numerical simulations also help study possible systematic effects presented in observations. Finally, one of the problems in cosmology is the impossibility to repeat experiments in order to estimate errors in the measurements. That is why numerical simulations are important. They can be used as possible realisations of the Universe allowing us to constrain our measurements.

In this context, Chapter 4 shows the different steps followed to construct the catalogues for covariance matrices. These catalogues are widely used by other studies in the SDSS program. We carefully explained each step and thus allow us to construct thousands of simulated realisations of our Universe. The main goal of this project is to create mock catalogues which describe the distribution of observed galaxies in an accurate way. So, we run simulations using the PATCHY code and the initial conditions from the BigMultiDark simulation. We fix the bias of the galaxy sample using a modified version of the model presented in Chapter 3. Once the simulations are run, halo masses are included with HADRON code. Finally, we build a pipeline using SUGAR code in order to complete the final production of mock catalogues. This pipeline is schematically shown in Figure 1 of Chapter 5.

MD-PATCHY mock catalogues are designed to reproduce the two-point correlation function of the observed sample (monopole and quadrupole). Furthermore, observational effects such as fiber collisions, radial and angular selection function and the incompleteness of the sample are included, just as in the model presented in Chapter 3. The final result is the production of more than 12,000 mock catalogues for BOSS LOWZ and CMASS samples, covering the northern and southern regions. These catalogues reproduce the power spectrum, the three

point correlation function and the bispectrum. MD-PATCHY mocks represent a step forward in the construction of covariance matrices, because their accuracy makes them the best candidates to provide information about the constraints of our measurements.

Future surveys will represent a new challenge for the covariance matrix construction. One of the most important features of the available codes to construct mock catalogues is their low computational cost. However, time and memory needed will dramatically increase for the volumes of the future surveys. Furthermore, these simulations will have to reproduce samples of galaxies with a bias smaller than the one of LRG. This makes that in the current large surveys, such as eBOSS, the available methods can only resolve large scales, which prevents the analysis on intermediate scales. These troubles are also present in numerical simulations. The increasing computational cost due to large volumes and high resolutions is a huge obstacle to run them. However, current simulations still allow us to study distributions of quasars but will represent a problem for ELG and future surveys.

In Chapter 5, we present a modified halo abundance matching which models the clustering of quasars. This method assumes that the circular velocity distribution of halos hosting quasars is Gaussian and can be described by three parameters. This simple model allows us to reproduce observations with a good agreement. Due to the features of the eBOSS quasar sample, it is necessary to fix two parameters by hand. Then, the mean of the Gaussian distribution is the only free paremeter to reproduce the clustering. This simple model is being used as initial conditions in the construction of covariance matrices, just as in the case of LRG. We also use it to find the typical host halo mass of quasars, which varies between $10^{12.5}\ M_\odot$ and $10^{12.8}\ M_\odot$. This wide range is mainly due to the uncertainties in fraction of QSO that live in subhalos which is not fixed by the current data. This model can also be applied to ELG, because they live in dark matter halos with similar masses. Future observations will provide measurements of the cross-correlation between different types of galaxies, which will allow us to better estimate the parameters of our model.

The SUrvey GenerAtoR code will be available soon in the GitHub repository[6].

[6] https://github.com/seroto36/SUGAR

Conclusiones

Los proyectos de esta tesis fueron desarrollados dentro de las colaboraciones BOSS y eBOSS del SDSS III y IV respectivamente. Los resultados presentados son fruto de numerosas discusiones y valiosos aportes de la comunidad científica que rodea estos proyectos. Mi investigación en el poryecto BOSS se enfocó en el estudio del agrupamiento de las galaxias luminosas rojas (LRG), sin embargo, durante el transcurso del doctorado también participé en estudios sobre las galaxias de líneas de emisión (ELG). En el actual proyecto del SDSS, eBOSS, mi trabajo se centró en el estudio de la distribución espacial de los cuásares. Cabe notar que nuestros modelos también pueden ser aplicados a las poblaciones de galaxias de eBOSS (LRG y ELG). Adicionalmente, como miembro del proyecto de simulaciones MultiDark, he tenido acceso a uno de los mejores conjuntos de simulaciones cosmológicas de N-cuerpos de la actualidad.

En el Capítulo 3 se presenta el modelo HAM utilizado para describir la población de galaxias luminosas rojas. Dicho modelo es aplicado en la simulación BigMultiDark para reproducir características del CMASS en el rango $0.43 < z < 0.7$. En este proyecto se incluyen efectos observacionales tales como la incompletitud en masa estelar de la muestra y características propias del telescopio como lo son las colisiones de fibras, la completitud angular, la función de selección radial y la geometría del cartografiado. De igual forma, se incluye la evolución de los halos de materia oscura construyendo conos de luz que dan como resultado catálogos simulados de galaxias que reproducen con muy buena precisión las características de las observaciones.

Nuestro trabajo combina una simulación basada en las mejores mediciones de los parámetros cosmológicos y características del cartografiados que pueden tener un impacto importante en las mediciones del agrupamiento de galaxias. A pesar de la simpleza del modelo utilizado para conectar galaxias y halos de materia oscura, éste genera catálogos de galaxias que reproducen las medidas observacionales de la función de correlación de dos y de tres puntos, así como de su análogo en espacio de Fourier, el espectro de potencias y el biespectro. Siendo éste un

resultado muy relevante para el modelo ΛCDM, ya que los parámetros cosmológicos utilizado son medidos en el fondo cósmico de microondas.

Nuestro trabajo, comparado con estudios anteriores, muestra la importancia de incluir efectos observacionales y de instrumentación, así como el gran impacto que tiene la correcta selección de los parámetros cosmológicos en la distribución de galaxias simuladas. Los resultados obtenidos son una prueba de que el modelo ΛCDM, que describe con una alta precisión el fondo cósmico de microondas (Planck Collaboration et al., 2014), también predice con una excelente exactitud la distribución de LRG para $z \sim 0.5$. Los futuros cartografiados mejorarán la precisión de las medidas del agrupamiento de galaxias, lo que permitirá extender y probar otros modelos que conecten halos de materia oscura y galaxias, los cuales podrán incluir efectos como el sesgo por ensamblaje de los halos (Assambly bias).

Adicionalmente, en el trabajo se muestra la predicción de la relación entre la masa las galaxias y la masa de los halos, la cual se ajusta perfectamente a las medidas observacionales hechas mediante lentes gravitacionales. Todas estas características hacen de nuestros catálogos unas herramientas muy útiles para el estudio de observaciones, así como para el análisis de efectos sistemáticos de los cartografiados. De esta forma, nuestro modelo nos ha permitido trabajar con otros grupos de investigación en el análisis de diferentes muestras de galaxias.

Las simulaciones numéricas son el mecanismo por medio del cual podemos comparar teoría y observaciones incluyendo física no lineal. Cuando un modelo reproduce la estructura a gran escala del Universo, es posible hacer predicciones con las simulaciones de cantidades como la masa característica de los halos para un determinado tipo de galaxias. Las simulaciones también sirven en el estudio de posibles efectos sistemáticos en las observaciones. Finalmente, uno de los problemas dentro de la cosmología es la imposibilidad de repetir experimentos para fijar errores en las mediciones, a diferencia de otros campos de la física. En este sentido, las simulaciones numéricas son utilizadas como posibles realizaciones del Universo que permiten calcular las incertidumbres en las mediciones observacionales.

En este sentido, en el Capítulo 4 se muestra la metodología seguida para construir los catálogos para el cálculo de las matrices de covarianza, las cuales utilizamos para conocer el nivel de precisión de las mediciones observacionales. Estos catálogos han sido utilizados en diferentes estudios dentro de la colaboración SDSS. En este capítulo se muestran los diferentes pasos realizados para poder construir miles de realizaciones simuladas de nuestro Universo. El objetivo fundamental en este proyecto fue la creación de catálogos que pudieran describir de forma precisa la distribución de galaxias observada. Para esto se corrieron simulaciones

con el código PATCHY y se utilizaron las condiciones iniciales de la simulación BigMultiDark. La relación entre galaxias y halos de materia oscura es fijada con una versión especial del modelo presentado en el Capítulo 3. Una vez las simulaciones son generadas, la masa es asignada a los halos usando el código HADRON. Finalmente, se construye un algoritmo basado en el código SUGAR para la producción final de los catálogos. Los pasos seguidos se pueden observar esquemáticamente en la Figura 1 del Chapter 4.

Los catálogos MD-PATCHY fueron diseñados para reproducir la función de correlación de dos puntos, específicamente el monopolo y el cuadrupolo. Adicionalmente, como en el caso presentado en el Capítulo 3, se incluyeron efectos como la colisión de fibras, las funciones de selección radial y angular o la incompletitud de la muestra de galaxias. Como resultado final se realizaron más de 12,000 catálogos para las muestras LOWZ y CMASS, en las regiones localizadas en el norte y en el sur, reproduciendo también el espectro de potencias, la función de correlación de tres puntos y el biespectro. MD-PATCHY ha representado un paso adelante en la generación de catálogos para la construcción de matrices de covarianza, su nivel de precisión hace de éstos unos catálogos que representan mejor las observaciones y por lo tanto pueden darnos una mejor información sobre las incertidumbres en las distintas mediciones.

Los futuros cartografiados van a traer un nuevo reto para la construcción de matrices de covarianza. Todos los códigos utilizados en este tipo de simulaciones son efectivos debido a su bajo consumo computacional, sin embargo, el tiempo y la memoria necesaria aumentan demasiado al incrementar el volumen de los cartografiados. Además, estas simulaciones tendrán que describir muestras de galaxias con un bias mucho menor al utilizado para las LRG. Esto hace que en experimentos recientes como eBOSS, los métodos actuales solo puedan resolver grandes escalas, impidiendo el análisis de escalas intermedias. Este problema también está presente en las simulaciones numéricas, el incremento del tiempo computacional debido al aumento del volumen y la resolución necesaria se vuelven un impedimento a la hora de correr estas simulaciones. Sin embargo, con los cartografiados y las simulaciones actuales todavía podemos realizar estudios sobre objetos como los cuásares, los cuales estarán en el punto de mira de las observaciones en los años venideros.

En el Capítulo 5 presentamos una modificación del modelo de abundancia de halos, con el cual podemos describir el agrupamiento de los cuásares observados por eBOSS. En este método asumimos que la distribución de la velocidad circular de los halos de materia oscura donde habitan cuásares es Gaussiana y puede ser descrita por tres parámetros. Este simple modelo permite importantes variaciones debido a los tres parámetros, lo que permite reproducir

de una forma precisa los datos observacionales. Debido a las características de la muestra de cuásares de eBOSS, es necesario fijar dos de los tres parámetros, siendo la mediana de la distribución Gaussiana el único parámetro libre que utilizamos. Este sencillo modelo está siendo utilizado, como en el caso de las LRG, para definir condiciones iniciales de las simulaciones para las matrices de covarianza, así como poner límites a las masas de halos donde viven los cuásares. En este sentido, se ha encontrado que los cuásares viven en halos de masas típicas entre $10^{12.5}\ M_\odot$ y $10^{12.8}\ M_\odot$, esto dependiendo principalmente de la fracción de subhalos incluidos en el análisis, el cual no arroja ninguna conclusión sobre la fracción de cuásares viviendo en subhalos. El modelo descrito en este trabajo también puede ser aplicado a las galaxias de líneas de emisión, pues éstas habitan en halos con masas similares a la de los cuásares. Futuras observaciones podrán dar mediciones de la cross-correlación entre distintos tipos de galaxias, lo que dará herramientas para mejorar la estimación de los parámetros de nuestro modelo, así como obtener información de la forma en que se seleccionan los halos, es decir, si un proceso aleatorio como el actual es suficiente o se requieren métodos más elaborados.

El código SUrvey GenerAtoR estará próximamente disponible en el repositorio GitHub[7].

[7] https://github.com/seroto36/SUGAR

Bibliography

Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., et al. (2016). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. preprint, (arXiv:1607.03155).

Ata, M., Baumgarten, F., Bautista, J., Beutler, F., Bizyaev, D., Blanton, M. R., et al. (2017). The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: First measurement of Baryon Acoustic Oscillations between redshift 0.8 and 2.2. preprint, (arXiv:1705.06373).

Avila, S., Knebe, A., Pearce, F. R., Schneider, A., Srisawat, C., Thomas, P. A., et al. (2014). SUSSING MERGER TREES: the influence of the halo finder. MNRAS, 441:3488–3501.

Baugh, C. M. (2006). A primer on hierarchical galaxy formation: the semi-analytical approach. Reports on Progress in Physics, 69:3101–3156.

Behroozi, P. S., Conroy, C., and Wechsler, R. H. (2010). A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for $0 < z < 4$. ApJ, 717:379–403.

Behroozi, P. S., Wechsler, R. H., and Wu, H.-Y. (2013a). The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. ApJ, 762:109.

Behroozi, P. S., Wechsler, R. H., Wu, H.-Y., Busha, M. T., Klypin, A. A., and Primack, J. R. (2013b). Gravitationally Consistent Halo Catalogs and Merger Trees for Precision Cosmology. ApJ, 763:18.

Berlind, A. A. and Weinberg, D. H. (2002). The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. ApJ, 575:587–616.

Beutler, F., Blake, C., Colless, M., Jones, D. H., Staveley-Smith, L., Campbell, L., et al. (2011). The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant. MNRAS, 416:3017–3032.

Beutler, F., Seo, H.-J., Saito, S., Chuang, C.-H., Cuesta, A. J., Eisenstein, D. J., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: anisotropic galaxy clustering in Fourier space. MNRAS, 466:2242–2260.

Blake, C., Kazin, E. A., Beutler, F., Davis, T. M., Parkinson, D., Brough, S., et al. (2011). The WiggleZ Dark Energy Survey: mapping the distance-redshift relation with baryon acoustic oscillations. MNRAS, 418:1707–1724.

Chuang, C.-H., Kitaura, F.-S., Prada, F., Zhao, C., and Yepes, G. (2015a). EZmocks: extending the Zel'dovich approximation to generate mock galaxy catalogues with accurate clustering statistics. MNRAS, 446:2621–2628.

Chuang, C.-H., Zhao, C., Prada, F., Munari, E., Avila, S., Izard, A., et al. (2015b). nIFTy cosmology: Galaxy/halo mock catalogue comparison project on clustering statistics. MNRAS, 452:686–700.

Cole, S., Percival, W. J., Peacock, J. A., Norberg, P., Baugh, C. M., Frenk, C. S., et al. (2005). The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications. MNRAS, 362:505–534.

Conroy, C., Wechsler, R. H., and Kravtsov, A. V. (2006). Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. ApJ, 647:201–214.

Cooray, A. and Sheth, R. (2002). Halo models of large scale structure. Phys. Rep., 372:1–129.

Dawson, K. S., Kneib, J.-P., Percival, W. J., Alam, S., Albareti, F. D., Anderson, S. F., et al. (2016). The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. AJ, 151:44.

Dawson, K. S., Schlegel, D. J., Ahn, C. P., Anderson, S. F., Aubourg, É., Bailey, S., et al. (2013). The Baryon Oscillation Spectroscopic Survey of SDSS-III. AJ, 145:10.

Delubac, T., Bautista, J. E., Busca, N. G., Rich, J., Kirkby, D., Bailey, S., et al. (2015). Baryon acoustic oscillations in the Lyα forest of BOSS DR11 quasars. A&A, 574:A59.

Dodelson, S. (2003). Modern Cosmology. Elsevier Science.

Drinkwater, M. J., Jurek, R. J., Blake, C., Woods, D., Pimbblet, K. A., Glazebrook, K., et al. (2010). The WiggleZ Dark Energy Survey: survey design and first data release. MNRAS, 401:1429–1452.

Eisenstein, D. J. and Hu, W. (1998). Baryonic Features in the Matter Transfer Function. ApJ, 496:605–614.

Eisenstein, D. J., Weinberg, D. H., Agol, E., Aihara, H., Allende Prieto, C., Anderson, S. F., et al. (2011). SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems. AJ, 142:72.

Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R., Nichol, R. C., et al. (2005). Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. ApJ, 633:560–574.

Favole, G., Rodríguez-Torres, S. A., Comparat, J., Prada, F., Guo, H., Klypin, A., et al. (2016). Galaxy clustering dependence on the [OII] emission line luminosity in the local Universe. preprint, (arXiv:1611.05457).

Font-Ribera, A., Kirkby, D., Busca, N., Miralda-Escudé, J., Ross, N. P., Slosar, A., et al. (2014). Quasar-Lyman α forest cross-correlation from BOSS DR11: Baryon Acoustic Oscillations. "J. Cosmology Astropart. Phys.", 5:027.

Frieman, J. and Dark Energy Survey Collaboration (2013). The Dark Energy Survey: Overview. volume 221 of *American Astronomical Society Meeting Abstracts*, page 335.01.

Genel, S., Vogelsberger, M., Springel, V., Sijacki, D., Nelson, D., Snyder, G., et al. (2014). Introducing the Illustris project: the evolution of galaxy populations across cosmic time. MNRAS, 445:175–200.

Gil-Marín, H., Percival, W. J., Verde, L., Brownstein, J. R., Chuang, C.-H., Kitaura, F.-S., et al. (2017). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies. MNRAS, 465:1757–1788.

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., Scoccimarro, R., Crocce, M., Dalla Vecchia, C., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the Fourier space wedges of the final sample. MNRAS, 467:2085–2112.

Guo, H., Zehavi, I., and Zheng, Z. (2012). A New Method to Correct for Fiber Collisions in Galaxy Two-point Statistics. ApJ, 756:127.

Guo, H., Zheng, Z., Behroozi, P. S., Zehavi, I., Chuang, C.-H., Comparat, J., et al. (2016). Modelling galaxy clustering: halo occupation distribution versus subhalo matching. MNRAS, 459:3040–3058.

Guo, H., Zheng, Z., Zehavi, I., Xu, H., Eisenstein, D. J., Weinberg, D. H., et al. (2014). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: modelling of the luminosity and colour dependence in the Data Release 10. MNRAS, 441:2398–2413.

Guo, Q., White, S., Li, C., and Boylan-Kolchin, M. (2010). How do galaxies populate dark matter haloes? MNRAS, 404:1111–1120.

Hahn, C., Scoccimarro, R., Blanton, M. R., Tinker, J. L., and Rodríguez-Torres, S. A. (2017). The Effect of Fiber Collisions on the Galaxy Power Spectrum Multipoles. MNRAS, 467:1940–1956.

Henriques, B. M. B., White, S. D. M., Thomas, P. A., Angulo, R., Guo, Q., Lemson, G., et al. (2015). Galaxy formation in the Planck cosmology - I. Matching the observed evolution of star formation rates, colours and stellar masses. MNRAS, 451:2663–2680.

Hinshaw, G., Larson, D., Komatsu, E., Spergel, D. N., Bennett, C. L., Dunkley, J., et al. (2013). Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. ApJS, 208:19.

Hu, W. and Sugiyama, N. (1996). Small-Scale Cosmological Perturbations: an Analytic Approach. ApJ, 471:542.

Jackson, J. C. (1972). A critique of Rees's theory of primordial gravitational radiation. MNRAS, 156:1P.

Jing, Y. P., Mo, H. J., and Börner, G. (1998). Spatial Correlation Function and Pairwise Velocity Dispersion of Galaxies: Cold Dark Matter Models versus the Las Campanas Survey. ApJ, 494:1–12.

Kaiser, N. (1987). Clustering in real space and in redshift space. MNRAS, 227:1–21.

Kitaura, F.-S., Yepes, G., and Prada, F. (2014). Modelling baryon acoustic oscillations with perturbation theory and stochastic halo biasing. MNRAS, 439:L21–L25.

Klypin, A. and Holtzman, J. (1997). Particle-Mesh code for cosmological simulations. preprint, (arXiv:astro-ph/9712217).

Klypin, A. and Prada, F. (2017). Dark matter statistics for large galaxy catalogs: power spectra and covariance matrices. preprint, (arXiv:1701.05690).

Klypin, A., Yepes, G., Gottlöber, S., Prada, F., and Heß, S. (2016). MultiDark simulations: the story of dark matter halo concentrations and density profiles. MNRAS, 457:4340–4359.

Knebe, A., Knollmann, S. R., Muldrew, S. I., Pearce, F. R., Aragon-Calvo, M. A., Ascasibar, Y., et al. (2011). Haloes gone MAD: The Halo-Finder Comparison Project. MNRAS, 415:2293–2318.

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B., et al. (2004). The Dark Side of the Halo Occupation Distribution. ApJ, 609:35–49.

Lacey, C. G., Baugh, C. M., Frenk, C. S., Benson, A. J., Bower, R. G., Cole, S., et al. (2016). A unified multiwavelength model of galaxy formation. MNRAS, 462:3854–3911.

Laureijs, R., Amiaux, J., Arduini, S., Augueres, J. ., Brinchmann, J., Cole, R., et al. (2011). Euclid Definition Study Report. preprint, (arXiv:1110.3193).

Leauthaud, A., Bundy, K., Saito, S., Tinker, J., Maraston, C., Tojeiro, R., et al. (2016). The Stripe 82 Massive Galaxy Project - II. Stellar mass completeness of spectroscopic galaxy samples from the Baryon Oscillation Spectroscopic Survey. MNRAS, 457:4021–4037.

Lesgourgues, J. (2011). The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview. preprint, (arXiv:1104.2932).

Lewis, A., Challinor, A., and Lasenby, A. (2000). Efficient computation of CMB anisotropies in closed FRW models. Astrophys. J., 538:473–476.

Liddle, A. R. (1998). Inflation and the cosmic microwave background. Phys. Rep., 307:53–60.

Maraston, C., Pforr, J., Henriques, B. M., Thomas, D., Wake, D., Brownstein, J. R., et al. (2013). Stellar masses of SDSS-III/BOSS galaxies at z $\sim$ 0.5 and constraints to galaxy formation models. MNRAS, 435:2764–2792.

Montero-Dorta, A. D., Bolton, A. S., Brownstein, J. R., Swanson, M., Dawson, K., Prada, F., et al. (2016). The high-mass end of the red sequence at z $\sim$ 0.55 from SDSS-III/BOSS: completeness, bimodality and luminosity function. MNRAS, 461:1131–1153.

Nuza, S. E., Sánchez, A. G., Prada, F., Klypin, A., Schlegel, D. J., Gottlöber, S., et al. (2013). The clustering of galaxies at z $\sim$ 0.5 in the SDSS-III Data Release 9 BOSS-CMASS sample: a test for the ΛCDM cosmology. MNRAS, 432:743–760.

Pâris, I., Petitjean, P., Aubourg, É., Ross, N. P., Myers, A. D., Streblyanska, A., et al. (2014). The Sloan Digital Sky Survey quasar catalog: tenth data release. A&A, 563:A54.

Peacock, J. A. and Smith, R. E. (2000). Halo occupation numbers and galaxy bias. MNRAS, 318:1144–1156.

Peebles, P. (1980). The Large-scale Structure of the Universe. Princeton series in physics. Princeton University Press.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., et al. (2014). Planck 2013 results. XVI. Cosmological parameters. A&A, 571:A16.

Reddick, R. M., Wechsler, R. H., Tinker, J. L., and Behroozi, P. S. (2013). The Connection between Galaxies and Dark Matter Structures in the Local Universe. ApJ, 771:30.

Ross, A. J., Beutler, F., Chuang, C.-H., Pellejero-Ibanez, M., Seo, H.-J., Vargas-Magaña, M., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: observational systematics and baryon acoustic oscillations in the correlation function. MNRAS, 464:1168–1191.

Ross, A. J., Percival, W. J., Sánchez, A. G., Samushia, L., Ho, S., Kazin, E., et al. (2012). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: analysis of potential systematics. MNRAS, 424:564–590.

Saito, S., Leauthaud, A., Hearin, A. P., Bundy, K., Zentner, A. R., Behroozi, P. S., et al. (2016). Connecting massive galaxies to dark matter haloes in BOSS - I. Is galaxy colour a stochastic process in high-mass haloes? MNRAS, 460:1457–1475.

Sartoris, B., Biviano, A., Fedeli, C., Bartlett, J. G., Borgani, S., Costanzi, M., et al. (2016). Next generation cosmology: constraints from the Euclid galaxy cluster survey. MNRAS, 459:1764–1780.

Schaye, J., Crain, R. A., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., et al. (2015). The EAGLE project: simulating the evolution and assembly of galaxies and their environments. MNRAS, 446:521–554.

Schlegel, D. J., Blum, R. D., Castander, F. J., Dey, A., Finkbeiner, D. P., Foucaud, S., et al. (2015). The Dark Energy Spectroscopic Instrument (DESI): The NOAO DECam Legacy Imaging Survey and DESI Target Selection. volume 225 of *American Astronomical Society Meeting Abstracts*, page 336.07.

Scoccimarro, R., Sheth, R. K., Hui, L., and Jain, B. (2001). How Many Galaxies Fit in a Halo? Constraints on Galaxy Formation Efficiency from Spatial Clustering. ApJ, 546:20–34.

Smoot, G. F., Bennett, C. L., Kogut, A., Wright, E. L., Aymon, J., Boggess, N. W., et al. (1992). Structure in the COBE differential microwave radiometer first-year maps. ApJL, 396:L1–L5.

Somerville, R. S. and Primack, J. R. (1999). Semi-analytic modelling of galaxy formation: the local Universe. MNRAS, 310:1087–1110.

Springel, V. (2005). The cosmological simulation code GADGET-2. MNRAS, 364:1105–1134.

Tinker, J. L., Brownstein, J. R., Guo, H., Leauthaud, A., Maraston, C., Masters, K., et al. (2017). The Correlation between Halo Mass and Stellar Mass for the Most Massive Galaxies in the Universe. ApJ, 839:121.

Trujillo-Gomez, S., Klypin, A., Primack, J., and Romanowsky, A. J. (2011). Galaxies in ΛCDM with Halo Abundance Matching: Luminosity-Velocity Relation, Baryonic Mass-Velocity Relation, Velocity Function, and Clustering. ApJ, 742:16.

White, M., Blanton, M., Bolton, A., Schlegel, D., Tinker, J., Berlind, A., et al. (2011). The Clustering of Massive Galaxies at z ˜ 0.5 from the First Semester of BOSS Data. ApJ, 728:126.

White, M., Tinker, J. L., and McBride, C. K. (2014). Mock galaxy catalogues using the quick particle mesh method. MNRAS, 437:2594–2606.

White, S. D. M. and Frenk, C. S. (1991). Galaxy formation through hierarchical clustering. ApJ, 379:52–79.

Zhao, C., Kitaura, F.-S., Chuang, C.-H., Prada, F., Yepes, G., and Tao, C. (2015). Halo mass distribution reconstruction across the cosmic web. MNRAS, 451:4266–4276.

Zhao, G.-B., Wang, Y., Saito, S., Wang, D., Ross, A. J., Beutler, F., et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey:

tomographic BAO analysis of DR12 combined sample in Fourier space. MNRAS, 466:762–779.

Zheng, Z., Berlind, A. A., Weinberg, D. H., Benson, A. J., Baugh, C. M., Cole, S., et al. (2005). Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies. ApJ, 633:791–809.

Zheng, Z., Coil, A. L., and Zehavi, I. (2007). Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. ApJ, 667:760–779.