

The best fit for the observed galaxy counts-in-cell distribution function

Lluís Hurtado-Gil^{1,2,3}, Vicent J. Martínez^{2,3}, Pablo Arnalte-Mur^{2,3}, María-Jesús Pons-Bordería⁴,
Cristóbal Pareja-Flores⁵, and Silvestre Paredes⁶

¹ Departamento de Matemática Aplicada y Estadística, Universidad CEU San Pablo, C/ Julian Romea 23, 28003 Madrid, Spain
e-mail: lluis.hurtadogil@ceu.es

² Observatori Astronòmic, Universitat de València, C/ Catedrático José Beltrán, 2, 46980 Paterna (València), Spain

³ Departament d'Astronomia i Astrofísica, Universitat de València, 46100 Burjassot, València, Spain

⁴ Departamento de Matemática Aplicada, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, 28040 Madrid, Spain

⁵ Departamento de Sistemas Informáticos y Computación, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, 28040 Madrid, Spain

⁶ Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, C/Dr. Fleming s/n, 30202 Cartagena, Spain

Received 10 June 2016 / Accepted 28 February 2017

ABSTRACT

Context. The Sloan Digital Sky Survey (SDSS) is the first dense redshift survey encompassing a volume large enough to find the best analytic probability density function that fits the galaxy counts-in-cells distribution $f_V(N)$, the frequency distribution of galaxy counts in a volume V .

Aims. Different analytic functions have been proposed that can account for some of the observed features of the observed frequency counts, but fail to provide an overall good fit to this important statistical descriptor of the galaxy large-scale distribution. Our goal is to find the probability density function that best fits the observed counts-in-cells distribution $f_V(N)$.

Methods. We have made a systematic study of this function applied to several samples drawn from the SDSS. We show the effective ways to deal with incompleteness of the sample (masked data) in the calculation of $f_V(N)$. We used LasDamas simulations to estimate the errors in the calculation. We tested four different distribution functions to find the best fit: the gravitational quasi-equilibrium distribution, the negative binomial distribution, the log normal distribution, and the log normal distribution including a bias parameter. In the two latter cases, we apply a shot-noise correction to the distributions assuming the local Poisson model.

Results. We show that the best fit for the counts-in-cells distribution function is provided by the negative binomial distribution. In addition, at large scales the log normal distribution modified with the inclusion of the bias term also performs a satisfactory fit of the empirical values of $f_V(N)$. Our results demonstrate that the inclusion of a bias term in the log normal distribution is necessary to fit the observed galaxy count-in-cells distribution function.

Key words. large-scale structure of Universe – galaxies: clusters: general – methods: statistical

1. Introduction

One of the first statistics applied to study the galaxy clustering in the very early galaxy catalogs (that were built on the projected celestial sphere) was the counts-in-cells (CiC) method. Hubble (1934) was the first to notice that the distribution of galaxy counts in two-dimensional cells could be well approximated by a lognormal distribution. This technique permits us to describe the spatial distribution of galaxies in a way that is complementary to other descriptors of the galaxy clustering such as the correlation function or the power spectrum (White 1979; Peebles 1980). The count probability distribution function $f_V(N)$ gives the probability that a randomly placed volume in the universe will contain exactly N galaxies. For $N = 0$, this function is known as the void-probability function (Maurogordato & Lachize-Rey 1987) and it is of particular interest, since is related with higher order correlation functions (White 1979) and provides a simple approach to establishing hierarchical scaling relations (Balian & Schaeffer 1989; Croton et al. 2004). Fry & Gaztanaga (1994) have shown that correlation functions can be measured from the moments

of $f_V(N)$. Several distribution functions have been used to model the counts-in-cells galaxy distribution (Yang & Saslaw 2011; Bel et al. 2016) and the counts-in-cells matter distribution (Clerkin et al. 2017), or used this distribution to estimate galaxy-dark matter bias (Sigad et al. 2000; Marinoni et al. 2005; Kovač et al. 2011; Di Porto et al. 2016).

However, the count probability distribution function is still missing a model that fits the observed distribution of galaxies well in the three dimensional space. This is the goal of the present paper. We plan to measure the galaxy counts-in-cells distribution on the New York University – Value Added Galaxy Catalog Blanton et al. (2005) and to assess what is the statistical distribution that fits best the observational frequency distribution. This blind fit of the observational galaxy CiC distribution will be accompanied of a model selection analysis, which will help us to determine the best fitting distribution.

The paper is organized as follows. In Sect. 2, we describe the data from SDSS Blanton et al. (2005) used for the analysis. In Sect. 3, we explain the procedure used to measure the observed

counts-in-cells (CiC) distribution $f_N(V)$, and describe how we have estimated the errors. In Sect. 4, we introduce different statistical distribution functions to model the observed galaxy CiC distribution. In Sect. 5, we present the observed results and its comparison with the statistical models introduced in Sect. 4. We summarize our conclusions in Sect. 6. In the Appendix A we use mock catalogs extracted from the LasDamas simulations [McBride et al. \(2011\)](#) to assess the significance of the internal error estimates. We use the same values for the cosmological parameters as were used by LasDamas simulation ([McBride et al. 2011](#)) in preparation for the error analysis of Appendix A: $\Omega_m = 0.25$, $\Omega_k = 0.0$, $\Omega_\Lambda = 0.75$ and $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$.

2. Data catalogs

The data analyzed in this work is a galaxy sample from the main catalog of the Sloan Digital Sky Survey (SDSS) Data Release 7 (DR7, [Abazajian et al. 2009](#)). This catalog is provided by The New York University – Value Added Galaxy Catalog ([Blanton et al. 2005](#)), presented in the next section. In addition to this, we also make use of LasDamas simulation catalog ([McBride et al. 2011](#)) to estimate the uncertainties in $f_V(N)$.

2.1. The SDSS – New York University – Value Added Galaxy Catalog

The NYU-VAGC ([Blanton et al. 2005](#)) is composed of data from the Sloan Digital Sky Survey DR7 ([Abazajian et al. 2009](#)) and the 2-Micron All-Sky Survey (2MASS) ([Skrutskie et al. 1997](#)), though here we make use of only the former. The SDSS-DR7 mapped one quarter of the entire sky and performed a redshift survey of galaxies, quasars and stars. It consists of a series of three interlocking imaging and spectroscopic surveys, carried out over an eight-year period with a dedicated 2.5 m telescope located at Apache Point Observatory in southern New Mexico.

The NYU-VAGC survey provides us with the position and redshift of more than 550 000 galaxies with corrected extinction and K-corrected absolute magnitudes for eight bands, of which the u , g , r , i , and z bands come from the SDSS. In addition, NYU-VAGC catalog also contains a survey geometry catalog, which define the window of the galaxy population and can be operated using the software MANGLE. This survey includes carefully constructed large-scale structure samples useful for calculating power spectra, correlation functions, for example.

In this work we have used two different samples selected in the r band, corresponding to the DR72 catalogs of the SDSS ([Abazajian et al. 2009](#)) included in NYU-VAGC with an apparent r -band magnitude limit of 17.6. These two samples are located in the same angular region. In this work we reproduce and compare our analysis with data from the LasDamas simulations ([McBride et al. 2011](#)), and for this reason we selected comparable datasets. Our first population is at low redshift, between 0.05 and 0.106. The low redshift limit ensures that the sample is within the *Hubble* flow, and excludes the Coma and Virgo clusters. We also took a second sample with redshifts between 0.075 and 0.165. To determine a suitable absolute magnitude cut, we define a faint limit M_r where the limiting magnitude has been reached to ensure luminosity completeness. In this way, we can assure a similar comoving number density of galaxies within the redshift limits. These limits are $M_r < -20$ for the first population and $M_r < -21$ for the second population. The samples are summarized in Table 1.

Table 1. SDSS selected samples.

Sample	Redshift	Magnitude $M - 5 \log(h)$	Density \bar{n} $h^3 \text{ Mpc}^{-3}$	Galaxies
Pop1	0.05–0.106	$M_r < -20$	5.7×10^{-3}	109046
Pop2	0.075–0.165	$M_r < -21$	1.04×10^{-3}	84445

Table 2. LasDamas selected mock catalogs.

Parent Simulation	Redshift	Magnitude $M - 5 \log(h)$	Density $h^3 \text{ Mpc}^{-3}$	Average n° of particles
Esmeralda	0.05–0.106	$M_r < -20$	6.01×10^{-3}	121838.9
Carmen	0.075–0.165	$M_r < -21$	1.11×10^{-3}	81733.8

2.2. LasDamas simulations

To obtain reliable estimations of the CiC distribution error bars, we will make use of the multiple realizations of the Large Suite of Dark Matter Simulations (LasDamas) [McBride et al. \(2011\)](#), [Swanson \(Accessed: 2016-03-25\)](#), a project that ran a large suite of cosmological N -body simulations that follow the evolution of dark matter in the universe. Results provide us an adequate resolution in many large boxes, rather than a single realization at high resolution. The enormous volume of generated data is appropriate for statistical studies of the distribution of galaxies and halos. The LasDamas simulations are designed to model the clustering of Sloan Digital Sky Survey (SDSS) galaxies in a wide luminosity range, with the goal of assisting in the modeling of galaxy clustering measurements. Specifically, the simulations are used to construct detailed mock galaxy catalogs by placing artificial galaxies inside dark matter halos using the Halo Occupation Distribution (HOD) with parameters fitted to reproduce the galaxy number density and projected correlation function of the respective SDSS galaxy samples. The HOD describes the distribution of galaxies within the dark matter halos. It uses three related properties of the halo model: the probability distribution relating the mass of a dark matter halo to the number of galaxies that form within that halo, the distribution in space of galactic matter within a dark matter halo and the distribution of velocities of galaxies relative to the dark matter within the halo. LasDamas is also designed to reproduce the SDSS DR7 geometry. Altogether, we can see that LasDamas simulations are specially adequate for NYU-VAGC comparison.

Data used consist of two different sets of mocks with different properties, matched to our SDSS samples. They are summarized in Table 2. Simulation *Esmeralda* contains 120 realizations and is started at an initial redshift of $z = 99$; while *Carmen* contains 164 realizations and is started at $z = 49$. Mocks are generated with the same initial power spectrum, but a different random seed.

As mentioned above, we have used LasDamas simulations to estimate the error bars in the CiC distribution. A brief analysis on the reliability of this method and other alternatives can be found in Appendix A.

3. Estimation of the CiC distribution

For our counts-in-cells process we use spherical cells in three-dimensional redshift space with constant radius. These cells are described by their center locations (three coordinates) and a radius r , and randomly distributed in space allowing them to

intersect. The higher the number of cells used in the calculation, the more precise our calculation of the counts-in-cells process will be.

A randomly distributed population of cells is generated in the studied survey volume. However, a galaxy survey window can be highly irregular, masked by stars and other objects. This affects not only the distribution of the cells but also their effective volume, since even if the center of the cell is inside the window, a significant part of the sphere might be outside. Once the centers of the galaxies are determined, we estimate the effective volume of the cell contained in the window. For this purpose was created the software MANGLE (Blanton et al. 2005; Hamilton & Tegmark 2004; Swanson et al. 2008). MANGLE performs several operations with complex angular windows, mapping them with sky-projected polygons and allowing one to find the polygons containing a given point on the sphere.

The effective volume of our cells is estimated with a Monte Carlo integration, populating a large number of random distributed points in our window and counting the number of points that lie inside our cells. This is exactly a CiC calculation over a random process. Given a window of volume $v(W) = V$ and a cell radius r , we determine the total number of necessary random points to estimate the effective volume with typical rms uncertainty a as

$$N = 1/a^2 \cdot t, \quad (1)$$

where $t = 4\pi r^3/3V$ is the ratio between the cell volume and the total volume. The number of points in a non masked cell has expectancy $N \cdot t$, and therefore the relative effective volume of each cell is estimated as

$$v_{\text{eff}} = N_{\text{obs}}/N \cdot t. \quad (2)$$

We estimate this effective volume for all cells using a number of random points corresponding to $a = 1\%$. Only cells with $v_{\text{eff}} \geq 95\%$ will be accepted.

With our accepted cells we can proceed to perform the counts-in-cells. The counting is done using the Euclidian metric and we are interested in knowing how many galaxies are found inside every cell of radius r . Then we multiply the number of counts by the inverse of the effective volume to obtain our counts-in-cells distribution. With this correction we estimate the number of galaxies unseen by the masking effects or redshift limits assuming uniform distribution of galaxies for cells dimensions.

The counts-in-cells distribution is binned in a histogram of frequencies, each bin containing the number of cells with n galaxies. If we normalize it by the total number of used cells we obtain the probability density function $f_V(N)$ of finding N galaxies in a cell of volume V . In our case, we will compute the CiC distribution $f_V(N)$ for our galaxy samples using five different cell radii: 8, 12, 16, 20 and $24 h^{-1}$ Mpc.

Estimation of the errors: LasDamas populations

The estimation of the uncertainty in our CiC measurements are done using the presented LasDamas populations. These samples have been selected so they can be compared with the SDSS populations. Using our accepted cells and their volume correction, we repeated the CiC calculations over LasDamas realizations, obtaining a reliable sampling of the distribution of cells containing a given number of galaxies. We used the standard deviation of these quantities as an error estimation for our CiC distribution over the SDSS populations. We computed the mean and the

variance with

$$\begin{aligned} \bar{x}_j &= \frac{1}{N_s} \sum_{i=1}^{N_s} f_V^j(i) \\ \sigma_j^2 &= \frac{1}{N_s - 1} \sum_{i=1}^{N_s} (f_V^j(i) - \bar{x}_j^j(i))^2, \end{aligned} \quad (3)$$

where N_s is the number of LasDamas samples and $f_V^j(i)$ is the i th realization distribution evaluated at the number of galaxies per cell $N = i$. We opted for using these quantities instead of errors estimated from intrinsic methods (i.e., the delete-one jackknife method, Norberg et al. 2009, or the bootstrap method, Efron & Tibshirani 1994) due to an overestimation of the found errors. An analysis of this issue can be found in Appendix A.

The standard deviation values σ_j found with LasDamas realizations provide error estimation for a large quantity of number of galaxies per cell N . However, for radii 20 and $24 h^{-1}$ Mpc we will see how, for every realization, no cells were found with a certain small number of galaxies inside. These values have $f_V(N) = 0$ with no estimation of their uncertainty, for which we assume the closest positive σ_j value. This is specially important for the case of the void-probability function, the probability of finding an empty cell ($f_V(0)$). These values might affect the fitting results (Sect. 5.1) and must be included in the best fit estimation. For large values of N this problem can be ignored.

As we will see in Sect. 5.1, we have made use of these standard deviations as weights for our χ^2 fitting procedure (see Eq. (17)). This implies assuming a Gaussian distribution for the observed abundances of cells containing N galaxies. However, the Shapiro-Wilk test for Gaussianity (Royston 1982) reveals how these distributions are not Gaussian when $f_V(N)$ is close to zero. We have not been able to find any known probability density satisfactorily modeling these distributions at all scales and further analysis is needed. As explained in Sect. 5.1 we perform the χ^2 fit twice, first using all the obtained σ_i values and then only with those satisfying Gaussianity. Since most of the non-Gaussian distributed errors correspond to high values of N , where the CiC distribution describe an asymptotic tail, negligible differences were found between both best fit parameters. However, the number of bins used can be crucial for the χ^2 value and the parameter error bars. The chosen solution consists of increasing the size of those bins where Gaussianity cannot be assured including the cell counts from the first Gaussian bin on the right. This creates a wide bin for the smaller N values in large cell radius samples, and another wide bin for the right tail. Altogether, Gaussianity of all bins can be assured to be above 0.9 in the Shapiro-Wilk test.

4. Frequency distribution models

In this section we present the probability distribution functions used to model the observed counts-in-cells distributions from the SDSS. Some of these functions have been commonly used in the cosmological literature to fit the CiC distribution, and some are an original contribution of this work. It is therefore of high interest to discriminate the best fitting model and their best fitting parameter values. Despite the information contained in a CiC distribution does not fully characterize the galaxy distribution, it is a useful constraint for N -body simulations or models of the galaxy distribution. Next we present the used probability density functions.

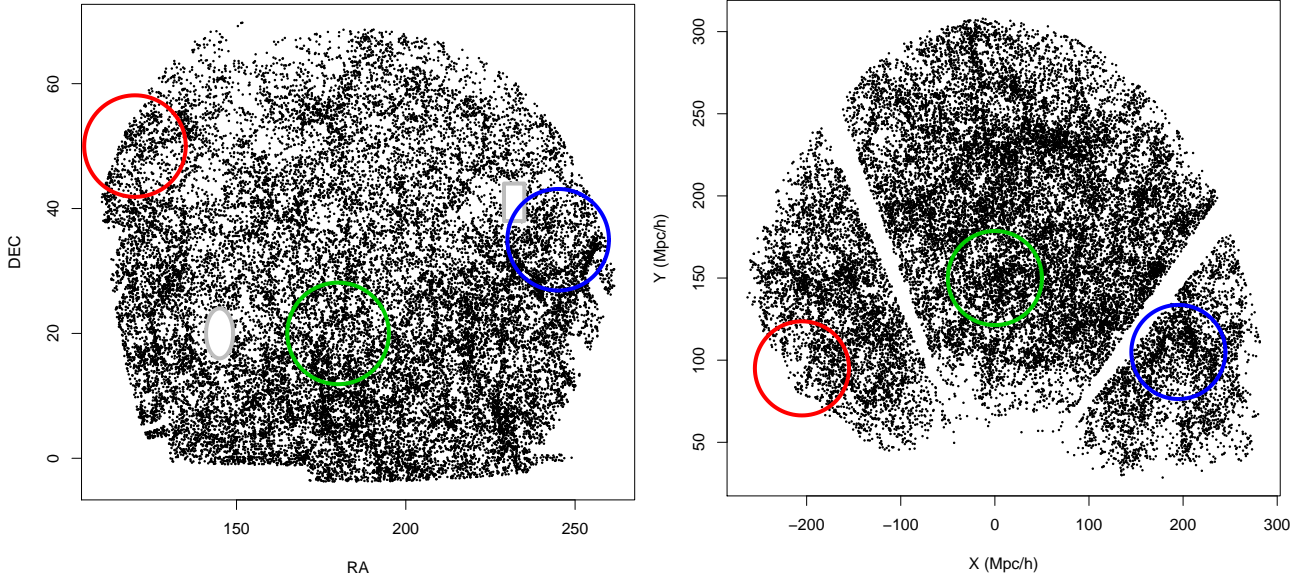


Fig. 1. Illustration of the calculation of the counts-in-cells. On *the left panel*, we show the footprint of about 10% of the galaxies in the SDSS survey used here (Population 1) in right ascension and declination. We see the irregular sky coverage in the borders outlined by the points. In addition, two masked regions have been added to illustrate the real mask. These masked regions are also illustrated on *the right panel* where we represent an edge-diagram in Cartesian comoving coordinates. In both panels, the green circle represents a cell which falls completely within the surveyed region, the blue circle represent a cell, still accepted for the counts-in-cells algorithm, that has part of its volume outside of the window, but at least 95% of it within the surveyed regions. In this case, counts are compensated to account for the galaxies unseen by the masked regions. Finally, the red circle represents a rejected cell because more than 5% of its volume falls out of the surveyed region.

4.1. Gravitational quasi-equilibrium distribution

The GQED (Saslaw & Hamilton 1984) is derived from a thermodynamical description of the galaxy fluid, though it can be also derived from the statistical mechanics (Ahmad et al. 2002). One of its principals assumptions is to accept that the galaxy accretion is succeeded through series of quasi-equilibrium states, a basic condition to start the thermodynamical approximation. This is possible if we assume an infinite number of galaxies in the universe. With this assumption we say clusters and large structures remain stable along large cosmological times before changing into a new quasi-equilibrium state.

We will also assume that galaxies are formed without dynamical interactions with the exterior and then they interact gravitationally as punctual masses with the rest of the universe. As the clusters dimensions are smaller than the curvature radius of the universe and the velocities involved are much lower than the speed of light, we can assume Newtonian gravity with potential $\phi = r^{-1}$.

After these assumptions, the GQED can be derived as the pdf

$$f_V(N) = \frac{\bar{N}(1-d)}{N!} [\bar{N}(1-d) + N \cdot d]^{N-1} e^{-[\bar{N}(1-d) + N \cdot d]}, \quad (4)$$

where the expected value of N is the product of the cell volume by the mean density of the galaxies, $\bar{N} = \bar{n}V$. Details of the deduction and properties of this distribution are described in full in Saslaw & Hamilton (1984), Sheth & Saslaw (1996) and Sheth (1995).

Parameter d (named b in Yang & Saslaw 2011) allows us to study both physical and statistical properties of this distribution. For the limit $d = 0$ we have a random distribution, where galaxies are uniformly distributed. At large scales, where fluctuations are small, the density function becomes Gaussian. Since this d value can be determined from physical magnitudes of the galaxies, we have a free parameter model of the galaxy distribution.

As we find in Ahmad et al. (2002), d represents a measure of the state of aggregation and we can express it as

$$d = \frac{3/2(Gm^2)^3 \bar{n} T^{-3}}{1 + 3/2(Gm^2)^3 \bar{n} T^{-3}}. \quad (5)$$

This quotient relates d to galaxy mass m , mean density \bar{n} , galaxies kinetic temperature T and gravitational constant G .

In the fitting process we will show the resulting parameters obtained directly from the CiC distribution (Eq. (6)) and parameters \bar{N} and d when both are fitted. The calculation of this distribution and the function Gamma for large values of N includes considerable difficulties, since the evaluation of the factorial function requires dealing with large numbers and needs a high precision to avoid the propagation error along the loops. As we will see in the following sections, this is the case for large cell radius distributions. This difficulty has been overcome using continuous fractions as representation of real numbers with infinite (i.e., arbitrarily large) precision and exploiting the model of lazy evaluation in the Haskell programming language (Peyton Jones et al. 2003), which allows us to perform heavy calculations and it is easy to integrate in C++.

In addition, we will compare these results with the estimated values of \bar{N} and d obtained directly from the CiC distribution. As explained in Yang & Saslaw (2011), the dependence of d on V can be obtained empirically from the variance of the number of counts in a cell of volume V ($\langle(\Delta N)^2\rangle_V$) and the mean \bar{N} .

$$d(V) = 1 - (\bar{N} \bar{\xi}_2(V) + 1)^{-1/2}, \quad (6)$$

where ξ_2 is the two-point correlation function.

4.2. Negative binomial bistribution

The negative binomial distribution (NBD) was first proposed in a cosmological context by Carruthers & Duong-van (1983) and

derived later by [Elizalde & Gaztanaga \(1992\)](#). With this model we study the probability of distributing N galaxies in m disconnected boxes. The probability of finding a galaxy in one of these boxes is proportional to the number of galaxies already located in the box. Expressing the probability function in terms of galaxies per cell instead of its conventional form, we have:

$$f_V(N) = \frac{\Gamma(N + \frac{1}{g})}{\Gamma(\frac{1}{g})N!} \frac{\bar{N}^N (\frac{1}{g})^{\frac{1}{g}}}{(\bar{N} + \frac{1}{g})^{N + \frac{1}{g}}}, \quad (7)$$

where Γ is the gamma function. As we had with the value d for the GQED, g is also an aggregation parameter. We can obtain it theoretically for the NBD with

$$g = \frac{\langle (\Delta N)^2 \rangle - \bar{N}}{\bar{N}^2}. \quad (8)$$

In this model, g depends both on the volume and on the shape of the cells.

This function is widely used in statistics and was firstly introduced in the cosmological context as a phenomenological model without physical explanation. Some authors ([Betancort-Rijo 2000](#)) claim that this function appear as a limit case in a variety of processes, in particular [Carruthers & Duong-van \(1983\)](#) found that approximates rather well the distribution of Zwicky clusters. Others, however, found that violates the second law of thermodynamics ([Yang & Saslaw 2011](#); [Saslaw & Fang 1996](#)). However, it provides a fair agreement with the observational distribution and it is thought to be related with the hierarchical universe properties. A deeper study of the implications of using such distribution suggests that the NBD assumes galaxies to be formed where there is already a galaxy cluster. Hence, this model does not take infall into account, but can describe the case where galaxies form from the merging of less massive objects. As before, we will show the resulting parameters obtained from Eq. (8) and parameters \bar{N} and g when both are fitted. Similarly to the GQED distribution, the NBD involves the calculation of factorials for large N values and infinite precision, which requires versions of these functions parameterized with the precision required, as we have already seen. We have also compared the results of Haskell programming mentioned above with the use of the Stirling approximation for the factorials of big numbers, by means of the MAXIMA program¹. The differences of the results are negligible.

4.3. Log normal distribution

The log normal distribution (LND) was first used by [Hubble \(1934\)](#) but it was not formally proposed until [Coles & Jones \(1991\)](#). It was one of the first fully described stochastic models applied to the distribution of matter density. This model allows us to calculate many complex properties of the CiC distribution.

While the Gaussian distribution provides us a valid description of linear and weak density perturbation fields, the log normal distribution represents the same case for the non-linear regime and is one of the few non-Gaussian random fields for which interesting properties are calculable analytically. In [Coles & Jones \(1991\)](#) one may find further motivation for this model, such as agreements with observational data. This has turned the log normal distribution into one of the most well known and widely applied models ([Wild et al. 2005](#); [Kitaura et al. 2010](#); [Kayo et al. 2001](#)).

¹ <http://maxima.sourceforge.net/>

Given a Gaussian random field $X(r)$, the probability of finding a certain number density from our field in a given position is given by $X \sim N(\mu, \sigma^2)$. The log normal field is obtained by applying a non-linear transformation of the Gaussian field: $Y(r) = \exp(X(r))$, where the new one-point probability function in the field Y is

$$f_1(y) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right) \frac{1}{y}. \quad (9)$$

This is the log normal variable $Y \sim \Lambda(\mu, \sigma^2)$ pdf, with the same parameters as the original Gaussian variable. In this work we make use of an expression adapted to the cosmological scenario ([Arnalte-Mur et al. 2016](#))

$$f(\Delta) = \frac{1}{\sqrt{2\pi}H_0} \exp\left(-\frac{y^2}{2H_0}\right) \frac{1}{\Delta}, \quad (10)$$

where

$$\begin{aligned} \Delta &= N/\bar{N} \\ H_0 &= \log(1 + C) \\ y &= \log\left(\Delta \sqrt{1 + C}\right), \end{aligned} \quad (11)$$

and C is the variance in the matter distribution, which we expect to be roughly $C = \sigma_8^2$ for cell radius $r = 8 h^{-1}$ Mpc at $z = 0$. This distribution has proved efficient in the description of the dark matter density field ([Arnalte-Mur et al. 2016](#); [Clerkin et al. 2017](#)), understood as a local non-linear transformation of the initial energy density distribution. However, the galaxy distribution studied in this work presents a bias with respect to the dark matter distribution, and the log normal distribution must be modified conveniently. This bias parameter b can be introduced in the probability density function as shown in [Dekel & Lahav \(1999\)](#).

$$f(\Delta) = \frac{1}{\sqrt{2\pi}H_0} \frac{\exp\left(-\frac{1}{2} \frac{y^2}{H_0}\right)}{\Delta + b - 1}, \quad (12)$$

where the above expressions are conveniently modified into

$$\begin{aligned} H_0 &= \log(1 + C_b) \\ y &= \log\left((\Delta + b - 1) \frac{\sqrt{1 + C_b}}{b}\right), \end{aligned} \quad (13)$$

and Δ remains the same. With this new distribution we expect improved fittings of the galaxy CiC distribution. In this case, we assume the simplest case of a scale-independent and linear bias, so $\delta_g = b\delta$, where δ_g and δ are the contrasts of the galaxy number density and the matter density, respectively.

Shot noise correction

As explained above, the log normal distribution is valid to describe a continuous field. In the cosmological context this is either the dark matter or the galaxy density fields. In our case, therefore, we need to introduce a discrete sampling of that underlying continuous field in order to compare to the observed counts-in-cells. This discrete sampling introduces a modification of the distribution known as shot noise. It is more important at smaller scales, where the typical number of galaxies per cell is low.

We correct for this shot noise term assuming the local Poisson process model ([Layzer 1956](#); [Peebles 1980](#)). This has been done in the literature both at the level of the moments

of the distribution (Gaztañaga et al. 2002; Angulo et al. 2008; Bel & Marinoni 2014; Bel et al. 2014) or, as we are dealing in this paper, at the level of the probability distribution function (Di Porto et al. 2016; Bel et al. 2016; Hoffmann et al. 2017). Following these authors the modified probability of galaxy counts, $f(N)$, can be expressed as

$$f(N) = \int_{-1}^{+\infty} f(\Delta)P(N|\Delta)d\Delta, \quad (14)$$

here $P(N|\Delta)$ is the conditional probability of finding N galaxies in a Poisson process with density Δ . Its expression is

$$P(N|\Delta) = \frac{(\bar{N}\Delta)^N e^{-\bar{N}\Delta}}{\bar{N}!}. \quad (15)$$

Although in this work we limit ourself to the use of a Poisson distribution, different alternatives can be used to correct from shot noise effects. In the following sections, we will use Eq. (14) as our log normal distribution $f_V(N)$ (with or without bias correction).

5. Results

Counts-in-cells calculations have been performed making use of at least as many cells as galaxies in the sample, which we consider a sufficient condition (Yang & Saslaw 2011). In addition, nearly 50% of cells are rejected by mask effects, so we have to double the number of tested cells. In Table 3 one can find the numbers of used cells for each SDSS sample and Las-Damas realizations under column Cells. The counts-in-cells observed probability density functions $f_V(N)$ are shown in Figs. 2 and 3 together with the best fits of the different distributions. The curves show the expected probabilities for our chosen samples and radii (Yang & Saslaw 2011), with higher values of \bar{N} for bigger cells. Used bins variate their width depending on the CiC population. We have used the Freedman-Diaconis rule (Freedman & Diaconis 1981), which fixes the width m of the bin depending on the number of observations and the spread of the population:

$$m = 2 \frac{IQR(X)}{n^{1/3}}, \quad (16)$$

where $IQR(X)$ is the interquartile of population X and n is the number of objects in X . The resulting quantity has been rounded, resulting on $m = 1$ for cell radii 8 and 12 h^{-1} Mpc, $m = 3$ for radius 16 h^{-1} Mpc and $m = 6$ for radius 24 h^{-1} Mpc in Population 1. In Population 2, all bins have width $m = 1$, except for radius 24 h^{-1} Mpc, where $m = 2$.

5.1. Fitting the results to a distribution function

In this section we proceed to fit the probability distribution functions defined in Sect. 4 to our counts-in-cells observed distributions. We calculate the χ^2 goodness of fit

$$\chi^2 = \sum_{N=0}^{N_{\max}} \frac{(f_V^{\text{obs}}(N) - f_V(N, \theta))^2}{\sigma_N^2}, \quad (17)$$

between the observed distribution and the theoretical distribution as a qualitative measure, where N_{\max} is the largest number of galaxies in a cell. The fitting makes use of the diagonal values σ^2 , ignoring correlations between bins. This estimation might

underestimate the obtained best fit χ^2 values and the parameters 1σ error bars, however, we have checked that this decision does not affect the best-fit values. After generating samples of Poisson-distributed points in the used masks with the densities of Table 1, we performed the described effective volume estimation method. The analytical Poisson distribution for the used densities are within the uncertainties of the obtained results and thus, we are confident of our methods. Vector θ contains the parameters of the theoretical distribution.

As introduced in Sect. 3, we have performed our fitting analysis over an observational distribution function with adaptive bin width, with wider bins for those values where $f_V(N)$ was close to zero and the Gaussianity of its error distribution could not be assured. The solution consisted of merging these bins with the closest Gaussian bin on the right. The best fit solution for the full curve analysis for each population can be found respectively in Figs. 2 and 3.

We show in Table 3 the values of θ and χ^2 obtained by performing a blind fit where all parameters are allowed to variate. Fittings have been performed using our own developed routines and also by means of the non-linear least-squares fitting MPFIT routines in IDL Markwardt (2009). Differences between both procedures are negligible. The numbers reported here are the ones by our own routines.

We use the χ^2 values to discriminate the best fitting distribution. These values can be compared with results in Figs. 2 and 3. The top box shows the observational and best fit distributions. In the bottom boxes we can see the residuals, the difference between the best-fit distributions and the observed one (X axis). As expected, curves outside the error bars have higher values of χ^2 .

Regarding the best fit distribution, the negative binomial distribution generally performs the best fit results with lower values of χ^2 . This is specially clear for radius 8 h^{-1} Mpc. However, we note that for higher radii, the log normal distribution with bias obtains remarkably good values as well, specially the higher radii of Population 2. At small scales, the log normal distribution with bias show a worse fit than the NBD. This may be an indication that the correction introduced in Sect. 4.3 does not completely correct for the shot noise term, or that this underlying distribution fails to describe the density field at these scales.

The gravitational quasi-equilibrium distribution generally performs poorer fittings than the NBD or the LN with bias distribution. But its χ^2 values are often close to those obtained by the LN. Finally, the LN distribution, without the bias modification, performs, as expected, worse fittings than any other distribution, clearly showing its unsuitability to describe the galaxy distribution, although it is rather appropriate to describe the dark matter distribution (see, e.g., Clerkin et al. 2017).

Regarding the found parameters, we have obtained similar results to Yang & Saslaw (2011), with generally good fittings (inside the 1σ values). A monotonic trend is found between parameter d and radii r , indicating stronger correlations inside bigger cells, which contain more structure. For the NBD we have $g = \xi_2(V)$, and therefore the opposite trend is found.

For the log normal distribution with bias, mean \bar{N} is, as expected, close to the value obtained in Table 3 or the best fit of GQED and NBD. Parameter C_b , the variance of the matter distribution, strongly varies with r . This quantity is roughly related to the cosmological parameter σ_8 , which at $z = 0$ is measured $\sigma_8 = 0.828 \pm 0.012$ (Planck Collaboration XVI 2014). Therefore, for the case of cells of radius 8 h^{-1} Mpc, we would expect the value of C_b to be close to $\sigma_8^2 = 0.686$, which we actually found for Population 1, with lower redshift, where $C = 0.71 \pm 0.03$. The monotonic evolution of C_b with the cell radius for Population 1

Table 3. Counts-in-cells best fit $f_V(N)$ – all parameters free.

Sample			NBD				GQED				
Pop	Cells	r	\bar{N}	g	χ^2	d.o.f.	\bar{N}	d	χ^2	d.o.f.	
1	195 079	8	13.08 ± 0.09	1.131 ± 0.014	124.56	124	13.12 ± 0.08	0.761 ± 0.0008	417.87	124	
	167 367	12	44.1 ± 0.2	0.59 ± 0.006	97.75	215	45.67 ± 0.18	0.8177 ± 0.0008	326.54	215	
	144 938	16	102.5 ± 0.8	0.373 ± 0.005	17.04	120	$105.6^{+0.6}_{-0.8}$	$0.8482^{+0.0008}_{-0.0017}$	115.66	120	
	123 972	20	200.1 ± 1.6	$0.252^{+0.007}_{-0.005}$	11.01	100	206.4 ± 1.7	0.87 ± 0.0017	60.2	100	
	215 408	24	338 ± 3	0.191 ± 0.005	6.71	115	347 ± 3	0.8812 ± 0.0018	24.65	115	
2	218 339	8	2.81 ± 0.02	1.53 ± 0.03	35.79	34	2.75 ± 0.04	0.575 ± 0.003	51.69	34	
	196 647	12	9.34 ± 0.07	0.848 ± 0.017	34.04	60	9.36 ± 0.09	0.672 ± 0.003	59.87	60	
	177 485	16	22.54 ± 0.18	0.488 ± 0.01	29.17	96	23.02 ± 0.18	0.715 ± 0.003	78.22	96	
	161 357	20	43.7 ± 0.3	0.336 ± 0.007	18.85	132	44.6 ± 0.4	$0.753^{+0.002}_{-0.003}$	22.9	132	
	147 554	24	75 ± 0.6	0.241 ± 0.008	9.62	87	76.6 ± 0.8	0.777 ± 0.004	8.81	87	
Log Normal						Log Normal + bias					
			N	C	χ^2	d.o.f.	\bar{N}	C_b	b	χ^2	d.o.f.
1	195 079	8	13.17 ± 0.09	1.66 ± 0.02	1097.83	124	12.53 ± 0.1	0.71 ± 0.03	1.42 ± 0.02	238.17	123
	167 367	12	46.6 ± 0.2	0.743 ± 0.011	542.29	215	44.5 ± 0.3	0.4005 ± 0.0005	1.297 ± 0.009	96.42	214
	144 938	16	107.4 ± 0.9	0.447 ± 0.011	157.18	120	103.9 ± 0.7	0.241 ± 0.017	1.29 ± 0.03	21.2	119
	123 972	20	213 ± 2	0.308 ± 0.011	71.88	100	200 ± 2	0.113 ± 0.018	1.52 ± 0.1	7.08	99
	215 408	24	351 ± 3	0.215 ± 0.008	30.03	115	338 ± 4	0.09 ± 0.015	1.48 ± 0.15	6.6	114
2	218 339	8	2.16 ± 0.05	5.6 ± 0.3	394.31	34	2.69 ± 0.04	1.1 ± 0.07	1.31 ± 0.03	50.03	33
	196 647	12	9.65 ± 0.2	1.26 ± 0.05	201.82	60	9.18 ± 0.13	0.54 ± 0.04	1.35 ± 0.05	51.8	59
	177 485	16	23.5 ± 0.2	0.547 ± 0.014	156.64	96	22.4 ± 0.3	0.252 ± 0.03	1.45 ± 0.07	29.54	95
	161 357	20	45.3 ± 0.4	0.381 ± 0.011	45.87	132	44.1 ± 0.4	0.21 ± 0.02	1.3 ± 0.07	10.54	131
	147 554	24	77.6 ± 0.9	0.269 ± 0.012	13.34	87	75.9 ± 1	0.16 ± 0.04	1.25 ± 0.1	6.03	86

Notes. Best fit parameters from the counts-in-cells distribution for our four probability functions. Different bin widths have been used, with wider bins for larger cells radii, resulting on different degrees of freedom for the χ^2 fitting. See Sect. 5 for detailed information. Column “Cells” corresponds to the used number of cells in each case.

shows a good description of the variation of the strength of the clustering with the scale. However, we have not been able to find the same results with Population 2. Results on the bias parameter b for different radius of the cells are compatible with expected scale dependent behavior of b (Baugh 2013).

An interesting comparison can be made between the galaxy variance C and the value $b^2 C_b$, the variance of a linearly biased tracer. If we compare these values as found in Table 3, we can see a general similarity of these values except for radii 8 and $12 h^{-1}$ Mpc of Population 2.

5.2. Model selection

The NBD, the GQED and the LND make use of two free parameters, therefore the χ^2 test should be enough to discriminate between them. However, the LN distribution with bias includes a third parameter, and this should be taken into account when comparing its performance with the other distributions. Therefore we use the hypothesis test to discriminate the best fitting distribution for the galaxy counts-in-cells distribution. We use the Bayesian and the Akaike information criteria (BIC and AIC, Schwarz 1978; Akaike 1998). The preference on the AIC or the BIC is widely debated (Lahiri 2001; Konishi & Kitagawa 2008). As usual, we make use of both. These criteria are calculated using the maximum likelihood value L , which can be defined with the χ^2 value from the best fit:

$$\log L = -\frac{1}{2}\chi^2 + K, \quad (18)$$

where K is a constant that only depends on the errors of the data values so it cancels out when comparing different models. And the information criteria

$$\text{BIC} = -2 \log L + |\theta| \log n, \quad (19)$$

$$\text{AIC} = -2 \log L + 2|\theta|, \quad (20)$$

where L is the maximum likelihood for a given sample and model, $|\theta|$ is the number of used parameters in the model and n is the number of observations in the data. Now, we can compare the goodness of fit on any pair of models. We are specially interested in comparing the performance of the three different two-parameters models with the LN distribution with bias (LNb). Then, we will say that a model α is performing a better (or worse) fit than the LN distribution with bias if $\text{BIC}_\alpha < \text{BIC}_{\text{LNb}}$ (or $>$), and similarly with AIC. It is in these comparisons where constant K cancels. Results are shown in Table 4.

The values obtained for the GQED and the LN distributions show that the LN distribution with bias is clearly superior, even with a additional parameter, and it is completely justified to use a modified log normal. However, a comparison between the NBD and LNb distribution shows that the first one is clearly preferable for small radii cells, when the shot noise effect is stronger. For radius $16 h^{-1}$ Mpc and higher, the performance is comparable for both distributions (Koopman 1943; Kass & Raftery 1995). Due to its success among the smaller radii we chose the NBD as the best fit distribution function rather than the LNb distribution, although the satisfactory results at large scales of the later make this distribution still a justifiable election.

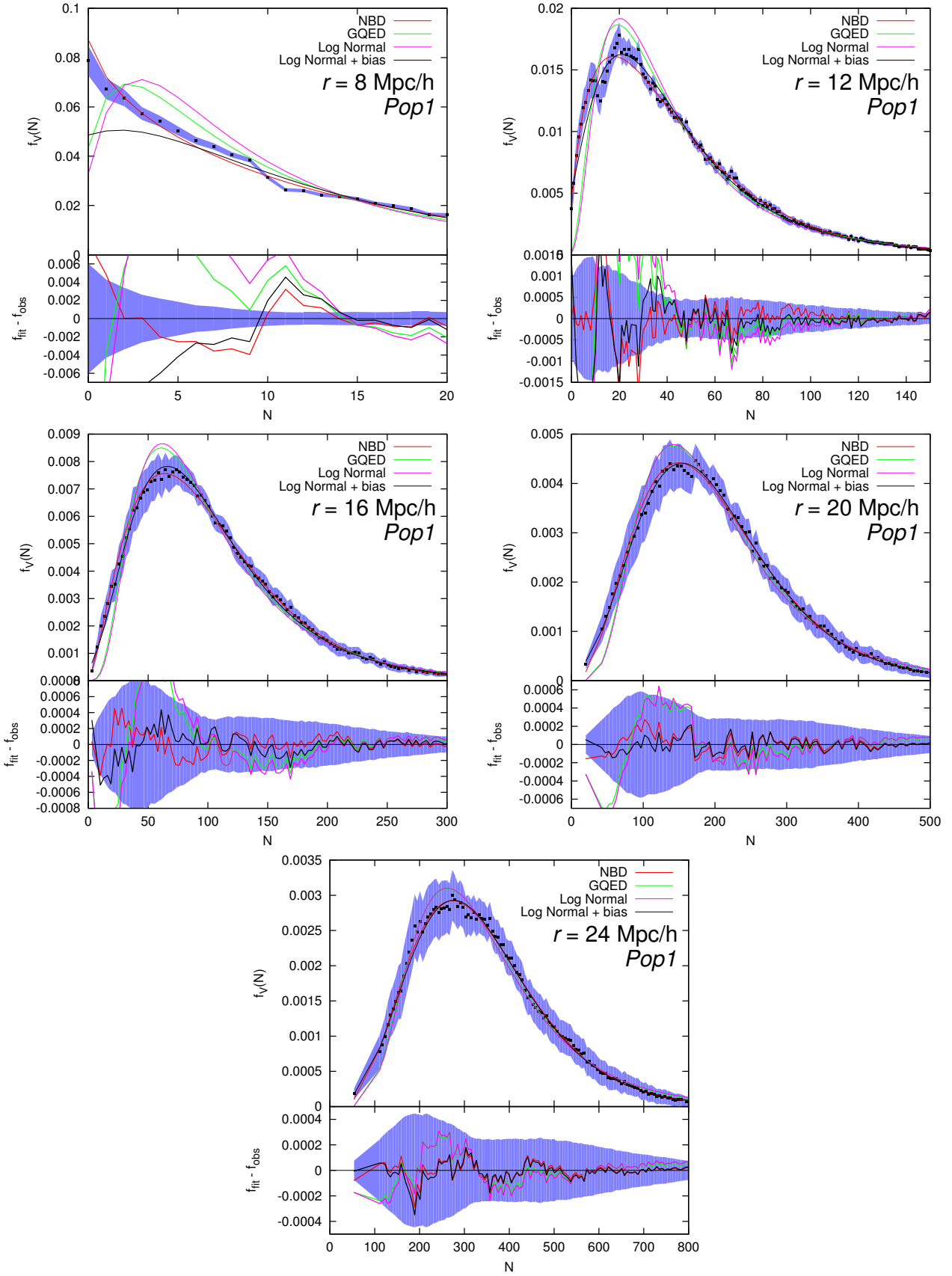


Fig. 2. Count-in-cells results for Population 1. $f_V(N)$ CiC distribution function with best fit models and error bars. *Top to bottom and left to right:* radii 8, 12, 16, 20 and $24 h^{-1}$ Mpc. Best fit plots for bins with Gaussian errors. Boxes: NBD residuals, GQED residuals, log normal distribution residuals, log normal with bias distribution residuals.

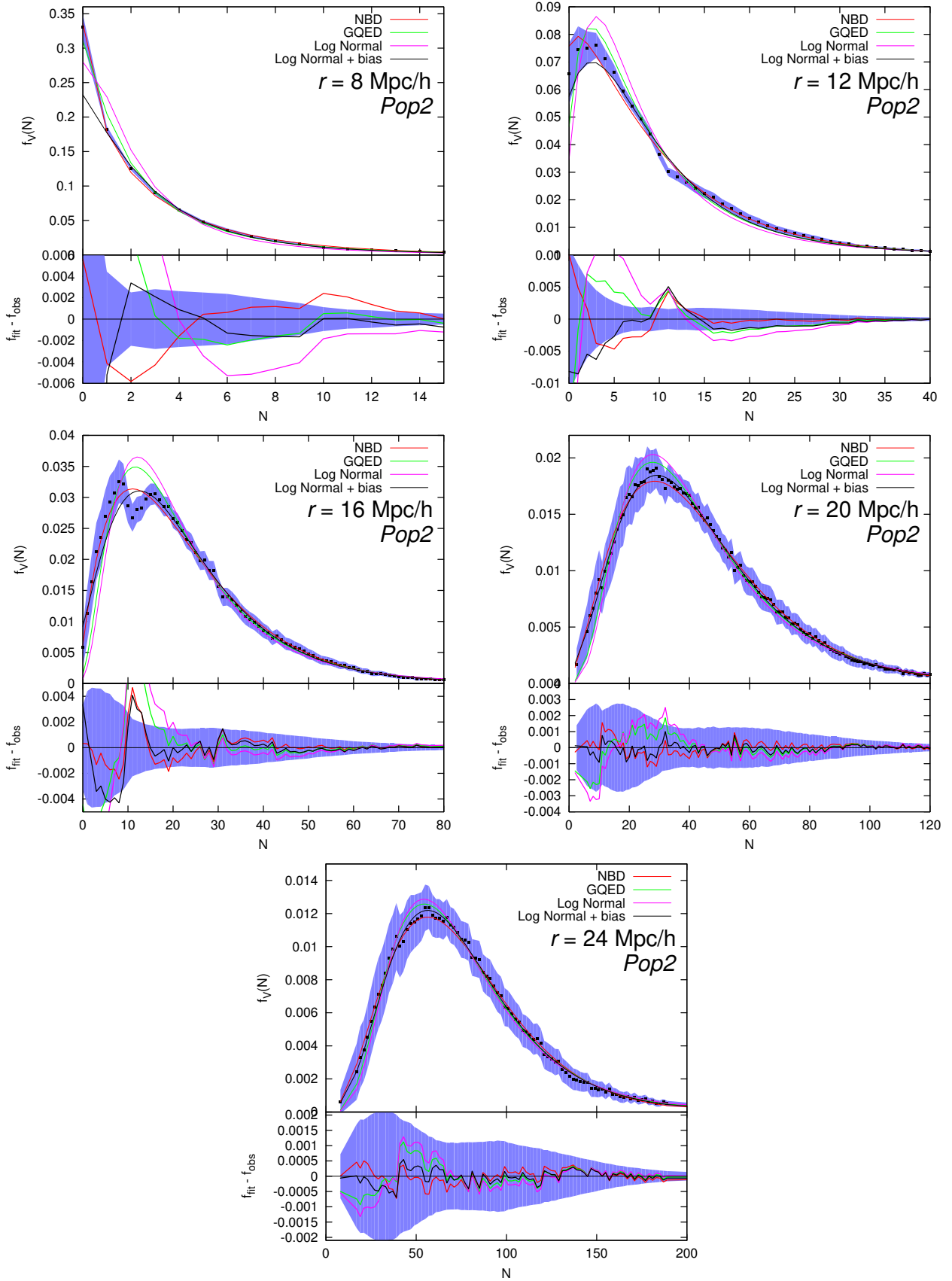


Fig. 3. Count-in-cells results for Population 2. $f_V(N)$ CiC distribution function with best fit models and error bars. *Top to bottom and left to right:* radii 8, 12, 16, 20 and $24 h^{-1}$ Mpc. Best fit plots for bins with Gaussian errors. Boxes: NBD residuals, GQED residuals, log normal distribution residuals, log normal with bias distribution residuals.

Table 4. Model selection.

Sample		Model		
Pop	r	NBD	GQED	LN
BIC $_{\alpha}$ – BIC $_{\text{LNb}}$				
1	8	-118.4234	174.8916	854.8506
	12	-4.039641	224.7505	440.5025
	16	-8.94503	89.67867	131.1977
	20	-0.6724359	48.51936	60.20096
	24	-4.626123	13.31321	18.69431
2	8	-17.74021	-1.896006	340.7819
	12	-21.84225	3.992247	145.9392
	16	-4.927798	44.1163	122.5448
	20	3.435232	7.287432	30.46023
	24	-0.8650677	-1.675998	2.853192
AIC $_{\alpha}$ – AIC $_{\text{LNb}}$				
1	8	-115.611	177.704	857.663
	12	-0.6736	228.1165	443.8685
	16	-6.1657	92.458	133.977
	20	1.92298	51.11478	62.79638
	24	-1.8897	16.04963	21.43073
2	8	-16.2412	-0.397	342.2809
	12	-19.7639	6.0706	148.0176
	16	-2.3736	46.6705	125.099
	20	6.3106	10.1628	33.3356
	24	1.58967	0.77874	5.30793

Notes. Hypothesis contrast between the NBD, the GQED and the LND with the LN distribution with bias. Values shown correspond to the different between the criterion used for one of the three two-parameters models (α) and the LN distribution. Negative values indicate that the two-parameter model performs a better fit than the LN distribution, while positive numbers indicate the opposite.

6. Discussion and conclusions

In this work we proceed with a blind fit of four different probability functions. A blind fit allows us to discriminate the best model of the observed counts-in-cells distribution. This work provides a very necessary tool for many different applications in modern cosmology, such as the generation of galaxy mocks or the testing of N -body simulations.

The GQED and the Log Normal distribution (with no bias parameter) generally show clearly poorer performances and both the χ^2 and the hypothesis test advice against using them as descriptors of the galaxy distribution. We have also fitted the calculated counts-in-cells distribution to the standard log normal distribution (Eq. (10)), and the fits are systematically worse than the fits obtained including the bias parameter. We conclude that this is an important result: the NBD provides the best fit for the galaxy counts-in-cells distribution for all scales, while the log normal distribution modified by the inclusion of the bias term might be a justifiable alternative choice for large-scale cells. Moreover, the estimation of parameters C and b from the log normal distribution incorporates an interesting asset and allows its direct comparison with other methods of bias estimation (López-Sanjuan et al. 2015; Hurtado-Gil et al. 2016). Nevertheless, we might have obtained results for the bias parameter in tension with other works (Zehavi et al. 2011), although our results could be an evidence of the expected scale dependent behavior of the bias parameter Baugh (2013).

Regarding the χ^2 results, the NBD is, generally the best model. For the higher radii, 16 to 24 h^{-1} Mpc, the log normal distribution with bias performs a slightly better fit. This probability distribution has three free parameters, which could explain its better performance. Actually, our hypothesis test analysis showed that the NBD (two free parameters) is preferable, since the lower χ^2 values of the log normal distribution do not justify the extra parameter. Therefore, we find that the NBD is the best model describing the counts-in-cells galaxy distribution for all scales, and that a choice for the modified log normal distribution is only justified at large scales.

Regarding the physical motivation of the NBD, here we must say that various authors find it fully justified (Elizalde & Gaztanaga 1992; Betancort-Rijo 2000; Carruthers & Duong-van 1983), while others have found that this model might contradict known physics (Saslaw & Fang 1996), although these authors also recognize that their calculation of the function $f_V(N)$ for 3D cells performed on the SDSS follows better the NBD than the GQED distribution. The same agreement with the NBD is found for the void-probability function, $f_V(0)$ of the SDSS by Conroy et al. (2005). Moreover, Bel et al. (2016) have recently shown that the negative binomial fits rather well the observed Counts-in-Cells distribution function for the VIPERS survey, concluding that the underlying continuous galaxy field should be well described by a Gamma distribution, since the NBD is considered to be the discrete analogue of the continuous Gamma distribution (see Adell & De la Cal 1994, where the approximation between these two distributions is studied).

In sum, we conclude that the negative binomial distribution has proved to be a very efficient descriptor of the counts-in-cells galaxy distribution.

Acknowledgements. We would like to thank an anonymous referee for comments and suggestions that have improved the quality and readability of this paper. This work has been supported by the Spanish Ministerio de Economía y Competitividad projects AYA2013-48623-C2-2 and AYA2016-81065-C2-2-P, including FEDER contributions, by the Generalitat Valenciana project of excellence PROMETEOII/2014/060. We acknowledge the use of data from the Las-Damas simulations, publicly available at <http://lss.phy.vanderbilt.edu/lasdamas/>. We also acknowledge the use of public data from SDSS. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Mombukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is <http://www.sdss.org/>. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

References

- Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, *ApJS*, **182**, 543
- Adell, J. A., & De la Cal, J. 1994, *J. Appl. Prob.*, **31**, 391
- Ahmad, F., Saslaw, W. C., & Bhat, N. I. 2002, *Apl*, **571**, 576
- Akaike, H. 1998, in Selected Papers of Hirotugu Akaike (Springer), 199
- Angulo, R. E., Baugh, C. M., & Lacey, C. G. 2008, *MNRAS*, **387**, 921
- Arnalte-Mur, P., Vielva, P., Martínez, V. J., et al. 2016, *J. Cosmol. Astropart. Phys.*, **3**, 005
- Balian, R., & Schaeffer, R. 1989, *A&A*, **220**, 1
- Baugh, C. M. 2013, *PASA*, **30**, e030

- Bel, J., & Marinoni, C. 2014, *A&A*, **563**, A36
- Bel, J., Marinoni, C., Granett, B. R., et al. 2014, *A&A*, **563**, A37
- Bel, J., Branchini, E., Di Porto, C., et al. 2016, *A&A*, **588**, A51
- Betancort-Rijo, J. 2000, *J. Stat. Phys.*, **98**, 917
- Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005, *AJ*, **129**, 2562
- Carruthers, P., & Duong-van, M. 1983, *Phys. Lett. B*, **131**, 116
- Clerkin, L., Kirk, D., Manera, M., et al. 2017, *MNRAS*, **466**, 1444
- Coles, P., & Jones, B. 1991, *MNRAS*, **248**, 1
- Conroy, C., Coil, A. L., White, M., et al. 2005, *ApJ*, **635**, 990
- Croton, D. J., Colless, M., Gaztañaga, E., et al. 2004, *MNRAS*, **352**, 828
- Dekel, A., & Lahav, O. 1999, *ApJ*, **520**, 24
- Di Porto, C., Branchini, E., Bel, J., et al. 2016, *A&A*, **594**, A62
- Efron, B., & Tibshirani, R. J. 1994, *An introduction to the bootstrap* (CRC press)
- Elizalde, E., & Gaztanaga, E. 1992, *MNRAS*, **254**, 247
- Freedman, D., & Diaconis, P. 1981, *Theory Relat Fields*, **57**, 453
- Fry, J. N., & Gaztanaga, E. 1994, *ApJ*, **425**, 1
- Gaztañaga, E., Fosalba, P., & Croft, R. A. C. 2002, *MNRAS*, **331**, 13
- Hamilton, A. J. S., & Tegmark, M. 2004, *MNRAS*, **349**, 115
- Hoffmann, K., Bel, J., & Gaztañaga, E. 2017, *MNRAS*, **465**, 2225
- Hubble, E. 1934, *ApJ*, **79**, 8
- Hurtado-Gil, L., Arnalte-Mur, P., Martínez, V. J., et al. 2016, *ApJ*, **818**, 174
- Kass, R. E., & Raftery, A. E. 1995, *J. Amer. Statist. Ass.*, **90**, 773
- Kayo, I., Taruya, A., & Suto, Y. 2001, *ApJ*, **561**, 22
- Kitaura, F.-S., Jasche, J., & Metcalf, R. B. 2010, *MNRAS*, **403**, 589
- Konishi, S., & Kitagawa, G. 2008, *Information criteria and statistical modeling* (Springer Science & Business Media)
- Koopman, B. 1943, *J. Symb. Log.*, **8**, 34
- Kovač, K., Porciani, C., Lilly, S. J., et al. 2011, *ApJ*, **731**, 102
- Lahiri, P. 2001, in *Model selection* (Beachwood: Institute of Mathematical Statistics), 1
- Layzer, D. 1956, *AJ*, **61**, 383
- López-Sanjuan, C., Cenarro, A. J., Hernández-Monteagudo, C., et al. 2015, *A&A*, **582**, A16
- Marinoni, C., Le Fèvre, O., Meneux, B., et al. 2005, *A&A*, **442**, 801
- Markwardt, C. B. 2009, in *Astronomical Data Analysis Software and Systems XVIII*, eds. D. A. Bohlender, D. Durand, & P. Dowler, *ASP Conf. Ser.*, **411**, 251
- Maurogordato, S., & Lachieze-Rey, M. 1987, *ApJ*, **320**, 13
- McBride, C., Berlind, A. A., Scoccimarro, R., et al. 2011, in *Am. Astron. Soc. Meet. Abstr.*, *BAAS*, **43**, 249.07
- Norberg, P., Baugh, C. M., Gaztañaga, E., & Croton, D. J. 2009, *MNRAS*, **396**, 19
- Peebles, P. J. E. 1980, *The large-scale structure of the universe* (Princeton, N.J.: Princeton University Press), 435
- Peyton Jones, S., et al. 2003, *J. Functional Programming*, **13**, <http://www.haskell.org/definition/>
- Planck Collaboration XVI. 2014, *A&A*, **571**, A16
- Royston, J. 1982, *Appl. Stat.*, **115**
- Saslaw, W. C., & Fang, F. 1996, *ApJ*, **460**, 16
- Saslaw, W. C., & Hamilton, A. J. S. 1984, *ApJ*, **276**, 13
- Schwarz, G. 1978, *Ann. Statist.*, **6**, 461
- Sheth, R. K. 1995, *MNRAS*, **274**, 213
- Sheth, R. K., & Saslaw, W. C. 1996, *ApJ*, **470**, 78
- Sigad, Y., Branchini, E., & Dekel, A. 2000, *ApJ*, **540**, 62
- Skrutskie, M. F., Schneider, S. E., Stiening, R., et al. 1997, in *The Impact of Large Scale Near-IR Sky Surveys*, eds. F. Garzon, N. Epchtein, A. Omont, B. Burton, & P. Persi, *Astrophys. Space Sci. Libr.*, **210**, 25
- Swanson, M. Accessed: 2016-03-25, LasDamas webpage, <http://lss.phy.vanderbilt.edu/lasdamas/>
- Swanson, M. E. C., Tegmark, M., Hamilton, A. J. S., & Hill, J. C. 2008, *MNRAS*, **387**, 1391
- White, S. D. M. 1979, *MNRAS*, **189**, 831
- Wild, V., Peacock, J. A., Lahav, O., et al. 2005, *MNRAS*, **356**, 247
- Yang, A., & Saslaw, W. C. 2011, *ApJ*, **729**, 123
- Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, *ApJ*, **736**, 59

Appendix A: Reliability of the error estimation

The estimation of error bars for our counts-in-cells distribution is a task that demands an additional analysis. Two kinds of methods can be followed: intrinsic and extrinsic. In the first group we have the jackknife resampling method, which together with the Bootstrap method (Efron & Tibshirani 1994), allows us to estimate the errors in multiple statistics. It can be used when we make use of summary statistics, like the counts-in-cells or the correlation function. With these statistics, we perform an analysis over the entire population, shrinking its information into a single quantity, such as the first and second characteristics. These methods are meant to be used when the limited volume of data in use does not allow us to compare our calculations among different populations, and therefore, we do not have a direct way to obtain its variation. As we only have one single galaxy population for each sample, we have to apply an internal error estimate method over the entire population.

On the other hand, the extrinsic methods make use of alternative samples to estimate the variance of our sample of study. In galaxy surveys analysis, these alternative samples can be obtained from mock catalogs, like those presented in Sect. 2.2 from LasDamas survey (McBride et al. 2011).

In this section we compare the performance of the two methods. To do so, we will compare the error estimations σ_i obtained in Sect. 3 with those obtained with the jackknife method over one of the LasDamas simulations.

The used jackknife method is the “delete one jackknife” method (Norberg et al. 2009). Our data from the chosen LasDamas simulation consists on coordinates in a certain space contained by a window. Dividing this window into different regions we generate N_{sub} subvolumes of different disjoint areas of the sky with the same redshift limits. Each subvolume contains a subsample of cells and deleting one of these subsamples from the parent population we produce a new sample. Cells are included in the corresponding subvolume using their center coordinates, and no further calculation is made regarding their volume overlapping with neighbor subsamples. These resampled data shares its cells and galaxies with the original one but lacks one of the subsamples. We can repeat this procedure N_{sub} times by systematically omitting, in turn, each of the subsamples in which the data has been split. The resampling of the data set consists of $N_{\text{sub}} - 1$ remaining subsamples, with volume $\nu(W) - \nu(V_i)$, where V_i is the volume occupied by the i th subvolume.

Our parent population of cells C occupies an irregular area of the sky due to the rejection process explained in Sect. 3. In order to ensure that all jackknife samples contain the same amount of cells, subsamples are selected in the following way: we sort our cells by declination and create 7 (or other desired number) subgroups of equal amount, in such a way that the groups are also sorted by declination. Then, we perform the same division for each subgroup with the right ascension, sorting each of them and dividing into another seven subgroups, 49 in total. Now we have subgroups of cells located in differently shaped rectangles but containing the same amount of cells.

Let f be the counts-in-cells distribution of our parent population. Once we have generated the new subsamples from the original population, we proceed to calculate the covariance matrix as

$$\Sigma_{ij} = \frac{N_{\text{sub}} - 1}{N_{\text{sub}}} \sum_{k=1}^{N_{\text{sub}}} (f^k(i) - \bar{f}(i))(f^k(j) - \bar{f}(j)), \quad (\text{A.1})$$

where N_{sub} is the number of used subsamples, f^k is the Counts-in-Cells distribution when we omit subsample k , and \bar{f} is the

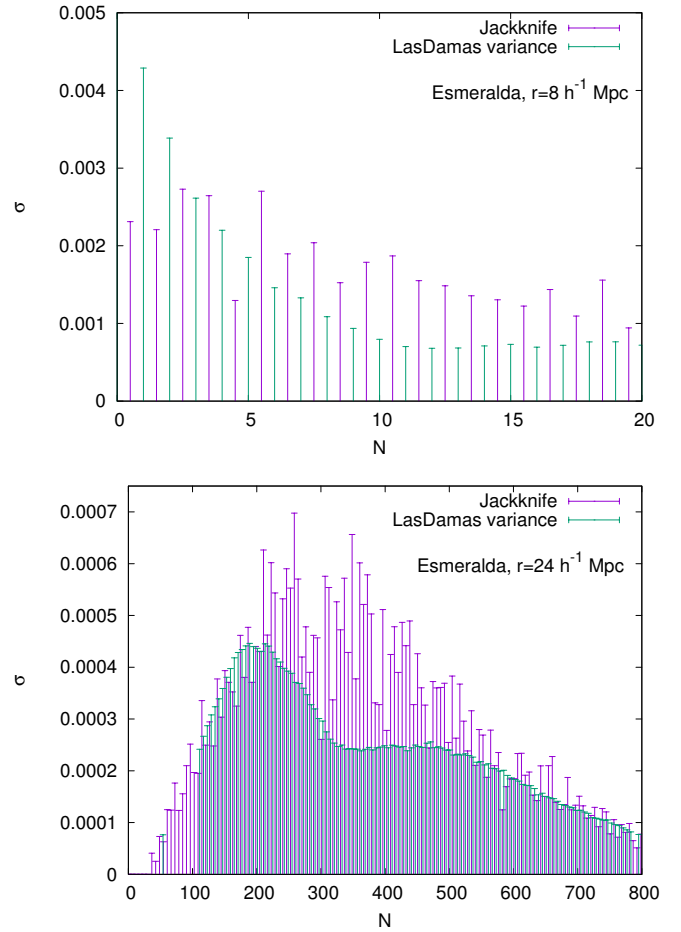


Fig. A.1. Comparison between “delete one jackknife” method error bars for $f_v(N)$ (pink) and its population variance (green) in Esmeralda simulation. *Top:* $8 h^{-1}$ Mpc cell radius, *bottom:* $24 h^{-1}$ Mpc cell radius.

mean of distributions f^k for every subsample $N_{\text{sub}} - k$. We compute these functions in the values i and j corresponding to all the considered values of N , this is, the values of the CiC frequency histogram, from 0 to the maximum number of galaxies contained into a cell. We finally calculate our standard deviations as $\sigma_i = \sqrt{\Sigma_{ii}}$.

Now, we can compare the jackknife error estimates with the standard deviations of the LasDamas simulations (see Sect. 3) and check if jackknife errors are over- or underestimated (Fig. A.1).

After these results, we see that the jackknife errors for LasDamas samples are typically larger than the errors obtained from the sample-to-sample variances in the catalogs. This indicates that the same systematic error should be found for the SDSS samples, with jackknife estimated error bars overestimating the real uncertainty. This proves that the jackknife estimations are unreliable in strong mask conditions, even after our cell rejection process. Hence, we will make use of the variances estimated from the LasDamas simulations (Eq. (3)) when fitting our probability functions. In Fig. A.2 we can see the LasDamas CiC distribution compared with the SDSS distribution. The good agreement between data and simulations allows us to use LasDamas variances as error bars for the SDSS counts-in-cells distributions.

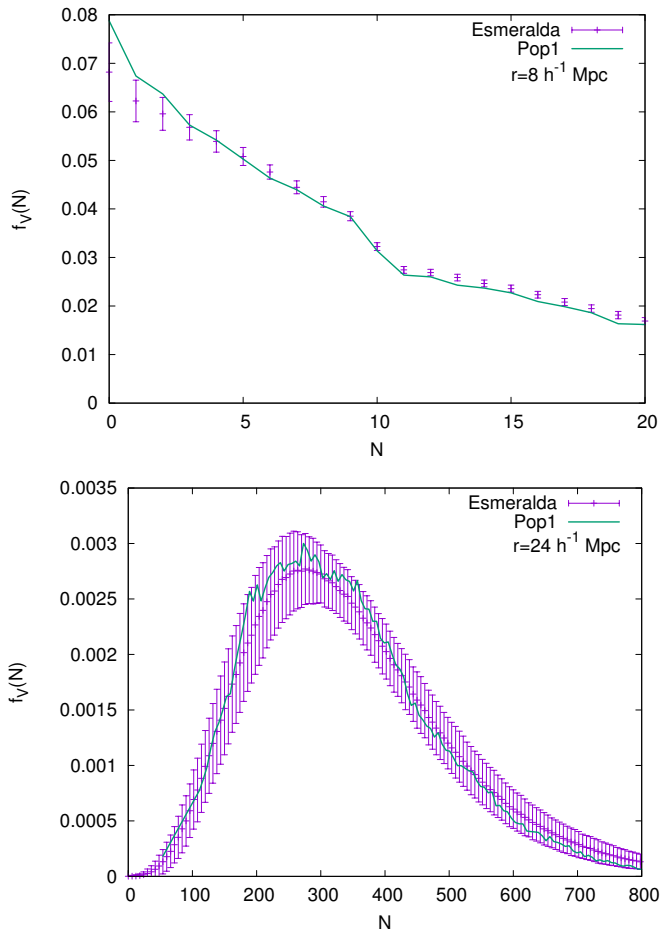


Fig. A.2. Comparison between LasDamas CiC distribution (pink) with the results obtained for the SDSS selected samples. *Top:* $8 \text{ h}^{-1} \text{ Mpc}$ cell radius, *bottom:* $24 \text{ h}^{-1} \text{ Mpc}$ cell radius.