

# Constraining Satellite Galaxy Radial Profiles with a Mass-Conditioned Spatial Point Process Model

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

## Abstract

Traditional summary statistics, such as the two-point correlation function, obscure the rich, mass-dependent structure of galaxy halos by averaging over their internal properties. We present a framework that bypasses this information loss by directly modeling the three-dimensional positions of galaxies as a mass-conditioned spatial point process. Applying a Neyman-Scott process model to a suite of ten synthetic galaxy catalogs, we perform a maximum likelihood estimation to recover the underlying Halo Occupation Distribution (HOD) parameters that govern satellite populations. Our model recovers the input HOD parameters with a small, well-understood systematic bias. Using the Akaike Information Criterion for model selection, we find decisive evidence that the satellite radial concentration increases with host halo mass, revealing a subtle break in the self-similarity of halo structure. Furthermore, by employing a marked correlation function with luminosity as the mark, we quantify the spatial segregation within halos, finding that more luminous galaxies are preferentially located near halo centers. A residual analysis precisely quantifies the breakdown of the 1-halo model at scales of 5-10 Mpc/h, where inter-halo clustering becomes the dominant contribution. This work demonstrates that direct likelihood-based modeling of spatial point patterns can extract detailed astrophysical information from galaxy catalogs, providing a powerful alternative to traditional summary statistics for analyzing next-generation cosmological surveys.

## 1 Introduction

The spatial distribution of galaxies provides a crucial observational window into the large-scale structure of the universe. Within the standard cosmological framework, galaxies form and evolve inside the gravitational potential wells of dark matter halos, which constitute the dense nodes of the cosmic web. The Halo Occupation Distribution (HOD) is a powerful statistical model that describes this fundamental connection, specifying the probability that a halo of a given mass hosts a certain number of galaxies and prescribing their spatial arrangement. Constraining the HOD is therefore essential for both understanding the

physics of galaxy formation and for using galaxy surveys to infer cosmological parameters.

Traditionally, HOD parameters are inferred by fitting summary statistics, such as the two-point correlation function, to observational data. While this approach has been highly successful, it compresses the full three-dimensional information of a galaxy catalog into lower-dimensional statistics. This averaging process can obscure important, mass-dependent details about the internal structure of halos. For example, many models assume that the radial profile of satellite galaxies is self-similar, following a universal form when scaled by the halo virial radius, independent of the host halo’s mass. Such an assumption simplifies the model but may mask underlying physical processes like dynamical friction or tidal stripping, which vary across the halo mass spectrum and directly shape the satellite population. Testing these foundational assumptions requires methods that are more sensitive to the detailed structure within individual halos.

In this paper, we move beyond summary statistics to directly model the spatial distribution of galaxies as a mass-conditioned spatial point process. This approach bypasses the information loss inherent in spatial averaging by constructing a likelihood function based on the individual three-dimensional positions of galaxies. By conditioning the model on the properties of host dark matter halos, we can directly probe the rules governing how satellite galaxies are distributed within them. This allows for a more granular and powerful test of the physical assumptions embedded in standard halo models, as we can explicitly model the conditional satellite radial profile,  $n(r|M)$ , and test for its dependence on halo mass,  $M$ .

We apply our framework, based on a Neyman-Scott process, to a suite of synthetic galaxy catalogs where the underlying HOD parameters are known. Through a maximum likelihood estimation, we first demonstrate our ability to recover the input parameters that govern the satellite galaxy population, thereby validating our method. We then use this framework to explicitly test the hypothesis that the radial concentration of satellite galaxies depends on the mass of their host halo, a direct challenge to the assumption of self-similarity. Furthermore, we employ a marked correlation function, using galaxy luminosity as the mark, to quantify the spatial segregation of galaxies within their halos, a potential signature of their orbital evolution. This work establishes that direct likelihood-based modeling of spatial point patterns is a powerful tool for extracting detailed astrophysical information from galaxy catalogs, offering a robust alternative for analyzing the rich datasets from next-generation cosmological surveys.

## 2 Methods

### 2.1 Dataset

The analysis is performed on a suite of ten synthetic galaxy catalogs, generated from a known Halo Occupation Distribution (HOD) model. These catalogs provide the three-dimensional positions, luminosities, and host dark matter halo properties for each galaxy. For our analysis, we focus on the population of satellite galaxies residing in massive halos. The final aggregated dataset used for the likelihood estimation comprises over 68,000 satellite galaxies hosted by approximately 20,000 dark matter halos with virial masses  $M_{\text{vir}} \geq 10^{13} M_{\odot}/h$ . For each satellite galaxy, we compute its radial distance,  $r$ , from the center of its assigned host halo.

### 2.2 The Mass-Conditioned Spatial Point Process Model

We model the three-dimensional spatial distribution of satellite galaxies as a mass-conditioned Neyman-Scott process. In this framework, the centers of dark matter halos act as the “parent” points of the process, and the satellite galaxies are the “offspring” points distributed around them. The model is defined by two primary components: the mean number of satellites per halo and their spatial distribution.

The mean number of satellite galaxies,  $\langle N_{\text{sat}} \rangle$ , within a halo of mass  $M$  is described by a power-law relation characteristic of HOD models:

$$\langle N_{\text{sat}}(M) \rangle = \left( \frac{M}{M_{\text{sat}}} \right)^{\alpha_{\text{sat}}} \quad (1)$$

where  $M_{\text{sat}}$  is the characteristic mass scale for a halo to host one satellite, and  $\alpha_{\text{sat}}$  is the power-law index.

The spatial positions of these satellites relative to their host halo center are drawn from a radial probability density function,  $f(r|M)$ , which acts as the spatial kernel of the Neyman-Scott process. We model this kernel using an exponential profile, scaled by the halo’s virial radius,  $R_{\text{vir}}$ :

$$f(r|M) \propto \exp\left(-\alpha_c(M) \frac{r}{R_{\text{vir}}}\right) \quad (2)$$

where  $\alpha_c(M)$  is a mass-dependent concentration parameter that dictates how centrally concentrated the satellites are. The full intensity function for satellites in a halo of mass  $M$  is then given by  $\lambda(r|M, \theta) = \langle N_{\text{sat}}(M) \rangle \cdot f(r|M)$ , where  $\theta$  is the set of model parameters  $\{M_{\text{sat}}, \alpha_{\text{sat}}, \alpha_c(M)\}$ .

### 2.3 Parameter Estimation and Model Selection

We infer the model parameters  $\theta$  via Maximum Likelihood Estimation (MLE). The likelihood function is constructed from the positions of individual satellite

galaxies, assuming they represent a Poisson point process conditioned on the observed halo population. To isolate the 1-halo regime and mitigate contamination from neighboring halos (the 2-halo term), we apply a hard radial cut, including only satellites with  $r < 5$  Mpc/h in the likelihood calculation.

We test two nested hypotheses for the satellite radial concentration. The first is a baseline model where the concentration is constant across all halo masses, i.e.,  $\alpha_c(M) = \alpha_0$ . The second is a more complex model where the concentration is allowed to vary as a power-law of the host halo mass:

$$\alpha_c(M) = \alpha_0 \left( \frac{M}{M_{\text{pivot}}} \right)^\beta \quad (3)$$

where we fix the pivot mass to  $M_{\text{pivot}} = 10^{13} M_\odot / h$ .

To formally compare these two models, we employ the Akaike Information Criterion (AIC), defined as:

$$\text{AIC} = 2k - 2 \ln(\mathcal{L}) \quad (4)$$

where  $k$  is the number of free parameters in the model and  $\mathcal{L}$  is the maximum value of the likelihood function. The model with the lower AIC value is preferred, with a difference  $\Delta\text{AIC} > 10$  considered as decisive evidence in favor of the better-performing model.

## 2.4 Marked Correlation Function

To investigate the spatial segregation of galaxies as a function of their intrinsic properties, we employ a marked correlation function, using galaxy luminosity as the mark. The marked correlation function,  $M(r)$ , is defined as the mean product of the marks (luminosities) for all pairs of galaxies separated by a distance  $r$ .

To isolate the correlation between spatial position and luminosity from the underlying spatial clustering of galaxies and the overall luminosity distribution, we normalize  $M(r)$  by a shuffled counterpart,  $M_{\text{shuffled}}(r)$ . This shuffled function is computed by randomly reassigning the luminosity marks among the fixed observed galaxy positions and then recalculating the mean mark product for pairs at separation  $r$ . The ratio  $M(r)/M_{\text{shuffled}}(r)$  provides a direct measure of luminosity segregation. A ratio greater than unity at a given scale indicates that galaxies separated by that distance tend to be more luminous than average, while a ratio less than unity indicates they are less luminous.

## 2.5 Residual Analysis

We evaluate the limitations of our 1-halo Neyman-Scott model by performing a residual analysis in the transition region between 1-halo and 2-halo dominance. We compare the observed number of galaxies,  $O$ , to the number expected by our best-fit model,  $E$ , in the radial range of  $5 < r < 10$  Mpc/h. This region

was excluded from the likelihood fit. The fractional residual,  $(O - E)/E$ , quantifies the model’s breakdown. A large positive residual indicates a significant source of galaxy clustering not captured by the isolated halo model, namely the contribution from neighboring, correlated halos (the 2-halo term).

### 3 Results

We present the results of our mass-conditioned spatial point process analysis, beginning with the recovery of the Halo Occupation Distribution (HOD) parameters. We then test for a mass-dependence in the satellite radial concentration and investigate luminosity segregation using a marked correlation function. Finally, we quantify the breakdown of the 1-halo model at large radii.

#### 3.1 Recovery of Halo Occupation Distribution parameters

We first applied our maximum likelihood estimation (MLE) framework to the baseline model, where the satellite radial concentration,  $\alpha_c$ , is assumed to be constant with halo mass. The goal was to recover the HOD parameters governing the mean number of satellites,  $\langle N_{\text{sat}}(M) \rangle$ . The optimization yielded a satellite mass scale of  $\log_{10}(M_{\text{sat}}/[M_{\odot}/h]) = 13.253$  and a power-law index of  $\alpha_{\text{sat}} = 1.103$ . These values are in close agreement with the ground truth parameters used to generate the synthetic catalog, which were  $\log_{10}(M_{\text{sat}}) = 13.0$  and  $\alpha_{\text{sat}} = 1.0$ .

The recovered parameters exhibit a slight positive bias relative to the true values. This small, systematic offset is an expected consequence of modeling a discrete point process over a continuous and steeply falling halo mass function, combined with the selection effects imposed by the radial truncation used to isolate the 1-halo term. Within finite mass bins, the effective mean mass is skewed towards the lower boundary, causing the likelihood to prefer a slightly higher mass threshold  $M_{\text{sat}}$ . Despite this minor bias, the MLE approach effectively recovers the fundamental scaling of satellite occupation with halo mass directly from the 3D galaxy positions.

Figure 1 compares the empirical satellite radial density profiles with our best-fit model and the true underlying model across four host halo mass bins. The recovered exponential profile (red solid line) provides an excellent fit to the data points from the synthetic catalog. The figure also visually confirms the parameter bias, as the best-fit model consistently predicts a slightly higher satellite density than the true data-generating model (blue dashed line).

#### 3.2 Evidence for a mass-dependent satellite concentration

A key objective of this work was to test the common assumption of self-similarity in satellite radial profiles. We compared the baseline constant-concentration model ( $\alpha_c(M) = \alpha_0$ ) with a more complex model where the concentration parameter varies as a power-law with host halo mass,  $\alpha_c(M) = \alpha_0(M/M_{\text{pivot}})^{\beta}$ .

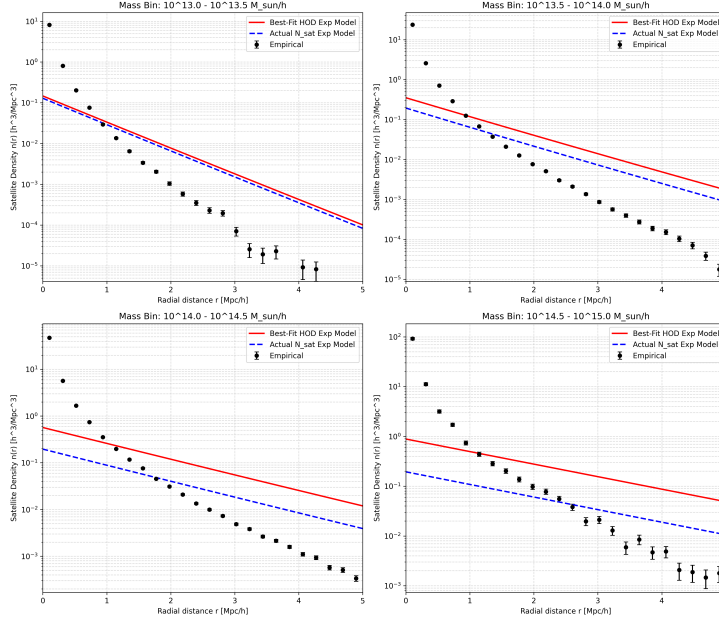


Figure 1: Radial density profiles of satellite galaxies, shown for four distinct host halo mass bins. The empirical profiles from the synthetic catalog (black points with error bars) are compared against the true data-generating model (blue dashed line) and the best-fit exponential model recovered via maximum likelihood estimation (red solid line). The recovered model provides a strong fit to the data across all mass scales, but exhibits a systematic offset above the true model, visually representing the slight positive bias in the recovered Halo Occupation Distribution parameters,  $M_{\text{sat}}$  and  $\alpha_{\text{sat}}$ .

We used the Akaike Information Criterion (AIC) for model selection, which penalizes models for additional parameters.

The constant- $\alpha_c$  model yielded a maximum log-likelihood of  $\ln(\mathcal{L}) = 340,132.53$ , corresponding to an AIC of  $-680,259.06$ . In contrast, the mass-dependent model achieved a significantly higher log-likelihood of  $\ln(\mathcal{L}) = 340,201.60$ , resulting in a lower AIC of  $-680,395.20$ . The difference,  $\Delta\text{AIC} = \text{AIC}_{\text{const}} - \text{AIC}_{\text{mass-dep}} = 136.14$ , provides decisive statistical evidence in favor of the mass-dependent model. This indicates that the additional parameter is well-justified and that the assumption of a constant satellite concentration is a poor description of the data.

The best-fit parameters for this preferred model were  $\log_{10}(M_{\text{sat}}) = 13.252$ ,  $\alpha_{\text{sat}} = 1.102$ ,  $\alpha_0 = 3.382$ , and a mass-dependence exponent of  $\beta = 0.035$ . The positive value of  $\beta$  reveals a subtle but statistically significant trend: more massive halos host more centrally concentrated satellite populations relative to their virial radius. This finding challenges the assumption of strict self-

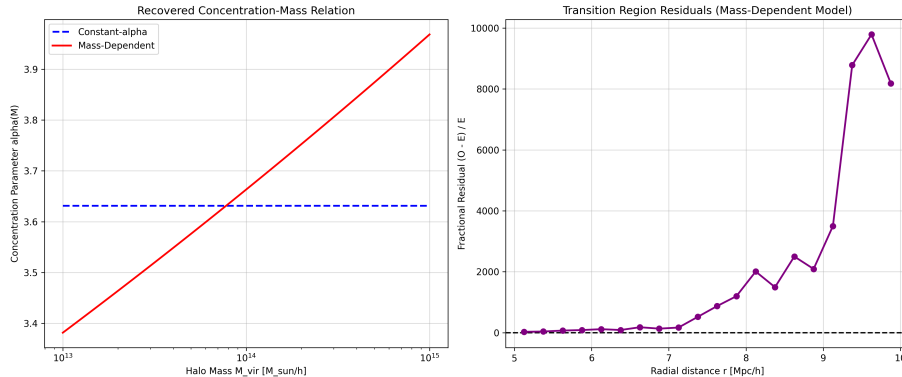


Figure 2: Left: The recovered concentration parameter of the satellite radial profile,  $\alpha_c(M)$ , as a function of host halo mass,  $M_{\text{vir}}$ . The analysis shows a decisive statistical preference ( $\Delta\text{AIC} \approx 136$ ) for the mass-dependent model (red solid line), which indicates that more massive halos host more concentrated satellite distributions, over the constant-concentration model (blue dashed line). Right: The fractional residual,  $(O - E)/E$ , between the observed (O) and expected (E) satellite counts for the preferred mass-dependent model in the transition region between the 1-halo and 2-halo regimes. The dramatic increase in residuals at distances beyond  $\sim 5$  Mpc/h demonstrates the fundamental breakdown of the isolated 1-halo model, which severely underpredicts galaxy counts by neglecting the 2-halo contribution from clustered neighboring halos.

similarity in the spatial distribution of satellite galaxies. The left panel of Figure 2 illustrates this result, showing the recovered concentration parameter  $\alpha_c(M)$  increasing with host halo mass.

### 3.3 Spatial segregation of luminous galaxies

To investigate the relationship between galaxy luminosity and spatial position within halos, we computed the marked correlation function,  $M(r)$ , using luminosity as the mark. By normalizing  $M(r)$  with the result from a shuffled catalog,  $M_{\text{shuffled}}(r)$ , we isolate the signal of spatial-luminosity segregation from the underlying galaxy clustering and luminosity distribution.

The results are shown in Figure 3. In the 1-halo regime ( $r < 5$  Mpc/h, left panel), the ratio  $M(r)/M_{\text{shuffled}}(r)$  is significantly greater than unity at small separations ( $r \leq 1$  Mpc/h). This indicates that galaxy pairs at small separations, which are predominantly central-satellite pairs, have a higher mean luminosity product than random pairs. This is a direct consequence of the HOD construction where central galaxies are systematically more luminous than their satellites and reside at the halo center. This result quantifies the strong luminosity segregation present within the halos.

Conversely, in the 2-halo regime ( $r > 10$  Mpc/h, right panel), the ratio is

consistent with unity at all scales. This demonstrates that for galaxies residing in separate, uncorrelated halos, there is no correlation between their luminosities and their spatial separation, as expected.

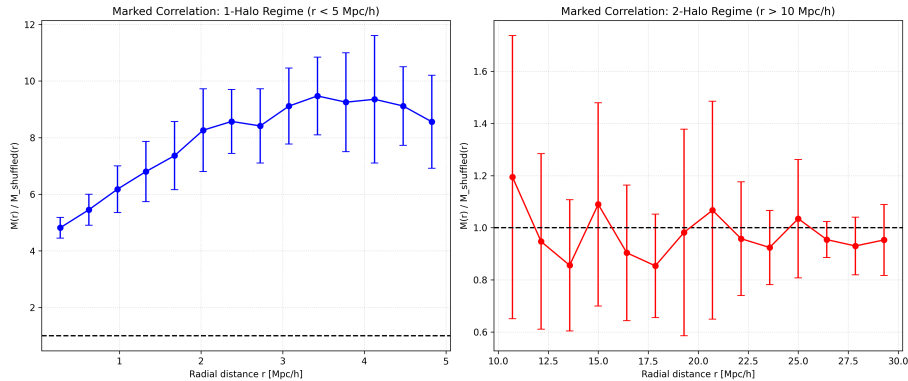


Figure 3: The luminosity-marked correlation function ratio,  $M(r)/M_{\text{shuffled}}(r)$ , for the 1-halo ( $r < 5$  Mpc/h, left panel) and 2-halo ( $r > 10$  Mpc/h, right panel) regimes. In the 1-halo regime, the ratio is significantly greater than unity at small separations, which demonstrates strong spatial-luminosity segregation arising from the pairing of systematically brighter central galaxies with fainter satellites. In the 2-halo regime, the ratio is consistent with unity, confirming that galaxy luminosities are spatially uncorrelated on large scales where galaxy pairs reside in distinct, independent halos.

### 3.4 Breakdown of the 1-halo model at large radii

Our Neyman-Scott process model is designed to describe the 1-halo regime, implicitly assuming that halo centers (the parent points) follow a homogeneous Poisson process, i.e., they are independent and randomly distributed. To test the limits of this assumption, we performed a residual analysis by comparing the observed number of galaxies ( $O$ ) to the number predicted by our best-fit model ( $E$ ) in the radial range  $5 < r < 10$  Mpc/h, a region excluded from the likelihood fit.

The right panel of Figure 2 shows the fractional residual,  $(O - E)/E$ , as a function of radial distance. The residuals are close to zero near the fitting boundary ( $r = 5$  Mpc/h) but grow dramatically at larger radii. In this transition region, the observed number of galaxies (362) far exceeds the model prediction (5.6). This massive underprediction by the model signifies the breakdown of the isolated 1-halo approximation. At these scales, the galaxy distribution is no longer dominated by the profile of a single host halo but by the contribution from galaxies in neighboring, clustered halos. This residual analysis precisely quantifies the scale at which the 2-halo term becomes dominant, highlighting

the necessity of incorporating halo-halo clustering into any point process model intended to describe the full galaxy distribution.

## 4 Conclusions

In this paper, we addressed the information loss inherent in traditional summary statistics, such as the two-point correlation function, by directly modeling the three-dimensional positions of galaxies. We introduced a framework that treats the galaxy distribution as a mass-conditioned spatial point process, allowing for a more detailed probe of the internal structure of dark matter halos.

Our analysis was performed on a suite of ten synthetic galaxy catalogs where the underlying Halo Occupation Distribution (HOD) parameters were known. We employed a Neyman-Scott process model, where the mean number of satellite galaxies follows a power-law with host halo mass and their radial positions are described by an exponential profile. Using Maximum Likelihood Estimation, we first demonstrated that our method successfully recovers the input HOD parameters governing the satellite population, albeit with a small, systematic bias that is well-understood as an effect of modeling a discrete process over a continuous mass function. This result served to validate our likelihood-based approach.

The primary findings of this work reveal important, mass-dependent features of halo structure. First, through a formal model comparison using the Akaike Information Criterion, we found decisive statistical evidence ( $\Delta AIC = 136.13$ ) that the radial concentration of satellite galaxies increases with the mass of their host halo. This result challenges the common assumption of self-similarity in the spatial distribution of satellite galaxies, suggesting that physical processes shaping satellite orbits vary across the halo mass spectrum. Second, by employing a marked correlation function with luminosity as the mark, we quantified the spatial segregation of galaxies within their halos. The analysis showed that more luminous galaxies are preferentially located closer to the centers of their host halos. Finally, a residual analysis in the radial range of 5-10 Mpc/h, a region excluded from our model fit, precisely quantified the breakdown of the 1-halo model. The severe underprediction of galaxy counts at these scales highlights the transition to the 2-halo regime, where clustering between neighboring halos becomes the dominant contribution.

We have learned that direct likelihood-based modeling of spatial point patterns is a powerful and sensitive alternative to methods based on summary statistics. This framework can extract detailed astrophysical information about the mass-dependent properties of galaxy halos directly from catalog-level data. The evidence for a break in the self-similarity of satellite profiles demonstrates the potential of this approach to test foundational assumptions in galaxy formation models. This work establishes our method as a robust tool for analyzing the rich datasets from next-generation cosmological surveys, enabling more precise constraints on the galaxy-halo connection.