FERMILAB-PUB-20-409-E

Studying the effects of overlapping objects in dark energy

Katarzyna Krzyżańska, SULI
Supervisor: Javier Sanchez

Observing the clustering of galaxies allows us to calculate cosmological parameters necessary for understanding dark energy. However, as the density of observed objects increases, the probability of these objects blending likewise increases, causing multiple galaxies to be observed as one. This affects the inferred values of parameters such as the galaxy bias (b) and the matter energy density (Ω_M). To see whether the bias from incorrectly inferring the galaxy count is significant, we compare the correlation functions in simulated data for "true" and "observed" data sets with one-to-one and multiple-to-one correspondences, respectively. For each data set, we create two correlation functions: one "measured" function directly relying on the galaxies' positions using the TreeCorr python library, and one "model" derived mathematically from the galaxies' power spectrum using the Cosmological Core Library (CCL). By minimizing the residual between these two functions, we compute the ideal values for b and Ω_M across the various possible redshifts that position the galaxies in three dimensional space. This minimization is done with an Markov chain Monte Carlo (MCMC) estimate that finds one value of Ω_M and ten values for b corresponding to the ten redshift bins ranging from z = 0.2 to z = 1.2. We find that neither b nor Ω_M is particularly affected by inclusion of blended galaxies. Though there is room for improvement, the data suggests that the fluctuations we found are a result of noise or limitations on the modeling rather than blending explicitly.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

I. INTRODUCTION

In order to determine certain parameters describing the distribution of dark energy, we can observe the clustering of galaxies. The Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) will probe more deeply than current surveys, with the expectation of reaching limiting magnitude i ≈ 27 after 10 years of operation as compared to, for example, the DES with i ≈ 24 (Melchior et al. 2018). However, partly as a result of Olber's paradox, as the density of objects in the sky increases, the probability of these objects overlapping and blending likewise increases. The HSC survey with a limiting magnitude of i ≈ 26 found that 58% of measured objects are affected by blending, as compared to the 30% from DES (Melchior et al. 2018). This suggests that i ≈ 27 will find similarly more blended objects (~60% of the objects, Sánchez et al, in prep.).

The authors in Dawson et al. 2015 find that 14% of objects in ground based imaging are blended to the degree that two or more objects are mistaken as a single object. Due to this, the perceived clustering might significantly differ from the true clustering, affecting the calculated values of parameters such as the galaxy bias (b) and the matter energy density (Ω_M). The impact of this shape-affecting blending on photometric redshift and weak lensing is out of the scope of this work. It is unclear whether the sample incompleteness due to our inability to see all dim galaxies at a given redshift or the biases resulting from the measured counts and centroid positions dominates at small scales.

To see whether the bias from incorrectly inferring the number of galaxies is significant, we examined simulated data representing blended objects in the sky and compared it to the "true" data set. We then calculated the correlation functions for the data points that had a one-to-one correspondence between the observed blended galaxies and the true ones as well as those data points with a multiple-to-one correspondence. These correlation functions are used to determine the bias in the optimal b and Ω_M values. This document is structured as follows: in Section II, we describe the data used for these calculations. In Section III, we provide the methods used for calculating the correlation functions and for analysis. Finally, in Section IV we present some concluding remarks.

II. DATA

The data used in calculations was obtained from the second data challenge (DC2) simulations prepared for the analysis of the by the LSST Dark Energy Science Collaboration (DESC), which is well documented in DESC Collaboration et al., in prep. As it specifies, DC2 generates and processes images created from the cosmoDC2 (Korytov et al. 2019) cosmological catalog using Rubin's LSST Science Pipelines[1] to account for weak gravitational lensing correlations, large-scale structure statistics, galaxy cluster abundance, and inference of ensemble redshift distributions for samples based on photometric redshifts. This simulated data covers 400 sq-deg, compared to the 18,000 of the full survey, and accurately portrays an output catalog.

What is relevant to our study is the creation of "truth" catalogs. These catalogs contain the true measurable properties produced by the LSST Science Pipelines which can be used to assess the output catalogs after image processing. We can then compare the truth and the object output catalogs to determine which galaxies have a one-to-one correspondence and which have a many-to-one correspondence between catalogs. Ideally, each "true" galaxy should be observed as one "object" galaxy, though blending impacts this fit. Figure 1 provides a 2D histogram showing the correlation of the true and observed datasets. One-to-one correlations have the single greatest count of matches; we found that of the 131,118,359 galaxies in the whole matched dataset, 81,652,106 have a one-to-one fit. The entirety of the data takes up enough memory that analysis had to be run on separate tracts of the data for it to be manageable. A majority of the analysis was carried out at NERSC through the Jupyter hub interface.[2] In principle, it offers up to 512 GB of memory with 64 cores, though the maximum recommended usage is 40 GB and 32 cores.

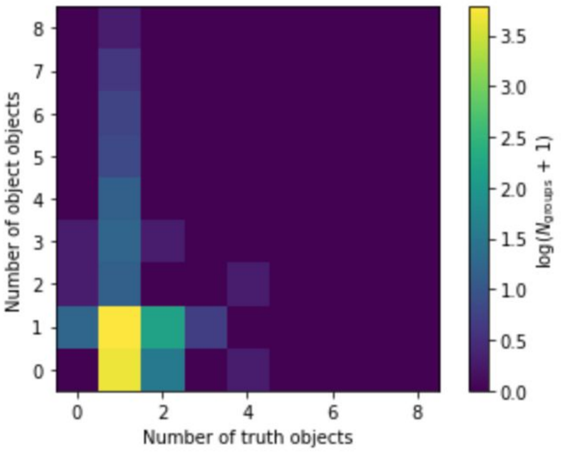

FIG. 1. Histogram relating the number of galaxies matched between the truth and object catalogs. The color represents the logarithm of the number of groups found with N objects in the truth (input) catalog, and N' objects in the object catalog (output). The groups were created using a Friends-of-friends algorithm[3] with one arcsecond linking length.

[1] pipelines.lsst.io

[2] jupyter.nersc.gov

[3] https://github.com/yymao/FoFCatalogMatching

III. METHODS

The two-point correlation function describes the probability of finding a pair of galaxies a given distance apart compared to if their distribution was random (Peebles, 1980). If we consider the residual between the measured correlation function and a model, we can fit for the ideal b and Ω_M values that minimize this residual. We calculated the measured correlation function using the TreeCorr (Jarvis et al. 2004) python library and the model correlation with the Core Cosmological Library, or CCL (Chisari et al. 2019).

With TreeCorr, estimating the correlation function involved just taking a data catalog and comparing it to a catalog of randomly distributed data. For every angle θ specified, TreeCorr finds a corresponding w(θ), where w is the two-point count-count (number density) correlation function in two dimensions. These two dimensions are given by the declination and right ascension angles; the redshift, *z,* provides a third dimension as a galaxy's redshift is correlated with its distance from Earth. Therefore, in order to accurately calculate how galaxies are clustered, the data is divided into ten redshift slices, ranging from z = 0.2 to z = 1.2 in steps of 0.1 following the analysis choices in the LSST Science Requirements Document (DESC Collaboration et al., 2018). The redshift for the data is the most reliable in this range. Greater values of z are too noisy, and the main features of galaxies at smaller z fall in the u- and g-filters that are more difficult to process than r- and i-, making the redshift inference more troublesome. To a lesser extent, at small z there may also be interference from overlaying stars. The spacing of the redshift bins is narrow enough to allow us to treat each one as though it were a two dimensional plane, where the galaxies within it can accurately be described by w(θ).

We also found the covariance matrix for the data using a jackknife approximation (Quenouille 1949) with TreeCorr. The approximation splits the data into 100 bins, and recalculates the correlation function that would have been computed if a given patch had been excluded. The covariance becomes relevant when minimizing the residual.

CCL, on the other hand, provides a theoretical model of the correlation function derived from the (dark) matter power spectrum, given a certain set of cosmological parameters and integrating over contributions observed fluctuations of the tracers representing the number of counts in a redshift bin. Here, dn(z) is the redshift distribution, $b^2(z)$ is the bias (assumed to be constant in each redshift slice), H is the Hubble parameter, and P(k,z) is the matter power spectrum. For k, we use the Limber approximation k ≈ l+1/r where r is the comoving distance at redshift z.

$$C_l^{TH} = \frac{2}{2l+1} \int \mathrm{d}z \left(\frac{\mathrm{d}n(z)}{\mathrm{d}z} \right)^2 b^2(z) H^2(z) \times P\left(k = \frac{l+1}{r}, z \right) \tag{1}$$

As given in equation (2), this is then summed over the angular wavenumbers used in calculating the power spectrum, which here ranged from 1 to 7500 (an approximation with error < 5%). $\mathscr{P}_l$ represents the Legendre polynomials. Full documentation concerning the inner workings of CCL is provided in Chisari et al. 2019

$$w(\theta) = \sum_{l \geq 0} (\frac{2l+1}{4\pi}) \mathcal{P}_l(\cos\theta) C_l \tag{2}$$

Calculating the model correlation function for the different tracers requires cosmological parameters including the matter energy density. The galaxy bias, on the other hand, is used as a scaling factor when calculating the residuals.

The variation in both the model and measured correlation function across redshift slices is given by figures 2.1 and 2.2. The analysis was ultimately done separately for the one-to-one and the many-to-one data so that after the fact the calculated b and Ω_M values could be compared. By observing the change in the separation between the model and measured correlation functions, we can anticipate how the galaxy bias value will change between redshift bins. Furthermore, just by visual inspection, we can see from figure 3 that the difference in correlation functions for these two datasets is slight.

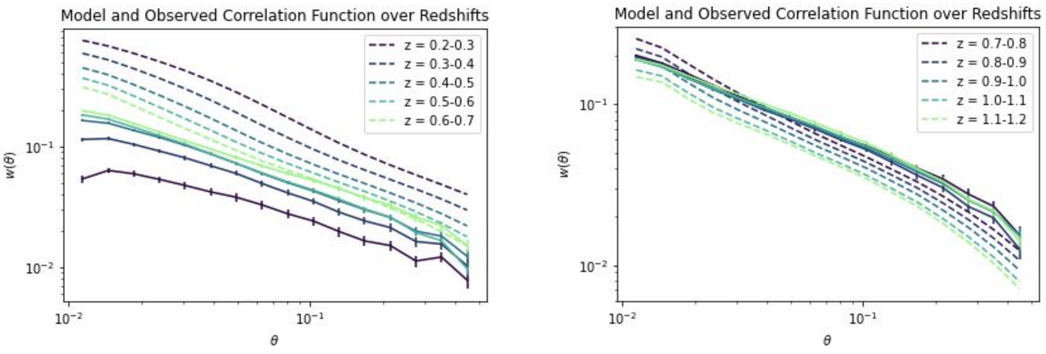

FIG. 2.1, 2.2. The model and observed correlation functions calculated with CCL and TreeCorr, respectively, for the redshift bins between z = 0.2 and z = 1.2. FIG. 2.1 displays 0.2 < z < 0.7, and FIG. 2.2 displays 0.7 < z < 1.2. Each color corresponds to a different bin. The dashed lines represent the model function, and the solid lines with error bars represent the measured function.

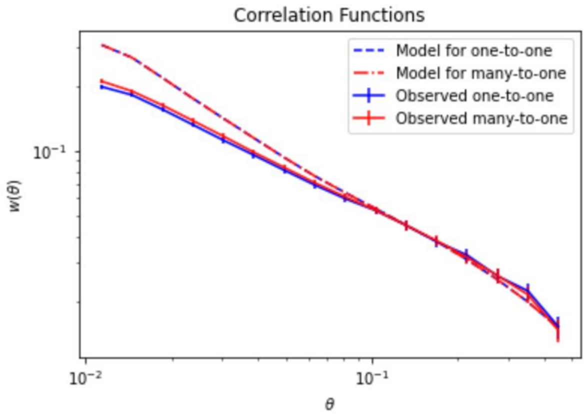

FIG. 3. An example of the model and measured correlation function for both the data with a one-to-one fit and the data with a many-to-one fit, taken from the redshift bin where $0.6 < z < 0.7$. The blue plots are the one-to-one fit, and the red are the many-to-one; the dashed lines represent the model function, and the solid lines with error bars represent the measured function.

An analysis is then done to minimize the residual between these two functions and find optimal values for b and Ω_M. A separate value for b is found for each redshift slice, as it is known that b depends on z. In the lower redshift regime, we are able to observe both bright-massive galaxies that tend to form in the more overdense regions, and smaller galaxies, resulting in an overall lower bias value. However, this is no longer true in the case of high redshift, where, due to observational limitations, we are only able to see the most massive objects, resulting in a large value for the bias. Figure 4 illustrates this trend. On the other hand, we consider Ω_M as $\Omega_{M,0}$, the value of Ω_M at z = 0 where $\Omega_M = \Omega_{M,0}(1+z)^3$, so this parameter as we calculate it should not depend on z. We therefore fit it for a single value that optimizes all the redshift slices at once.

We performed a Markov chain Monte Carlo (MCMC) estimate using the package emcee (Foreman-Mackey et al. 2013) and found b for each bin as well as Ω_M, as shown in figure 5 and table I. Table I provides the numerical estimates for the contours displayed in figure 5. Separate estimations were done for the one-to-one fit, a many-to-one fit that excludes all galaxies with a one-to-one fit, and a many-to-one fit that also includes one-to-one values. This latter match is what was used in previous calculations, and most accurately represents how we would actually observe the data. The many-to-one fit that excludes the one-to-one fit exists for purposes of comparison.

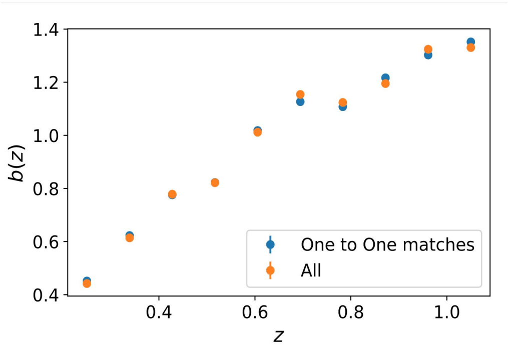

FIG. 4. The value obtained for the galaxy bias b for each redshift bin with a lower bound at z.The many-to-one matches, represented by “All”, closely follow the pattern of the one-to-one matches.

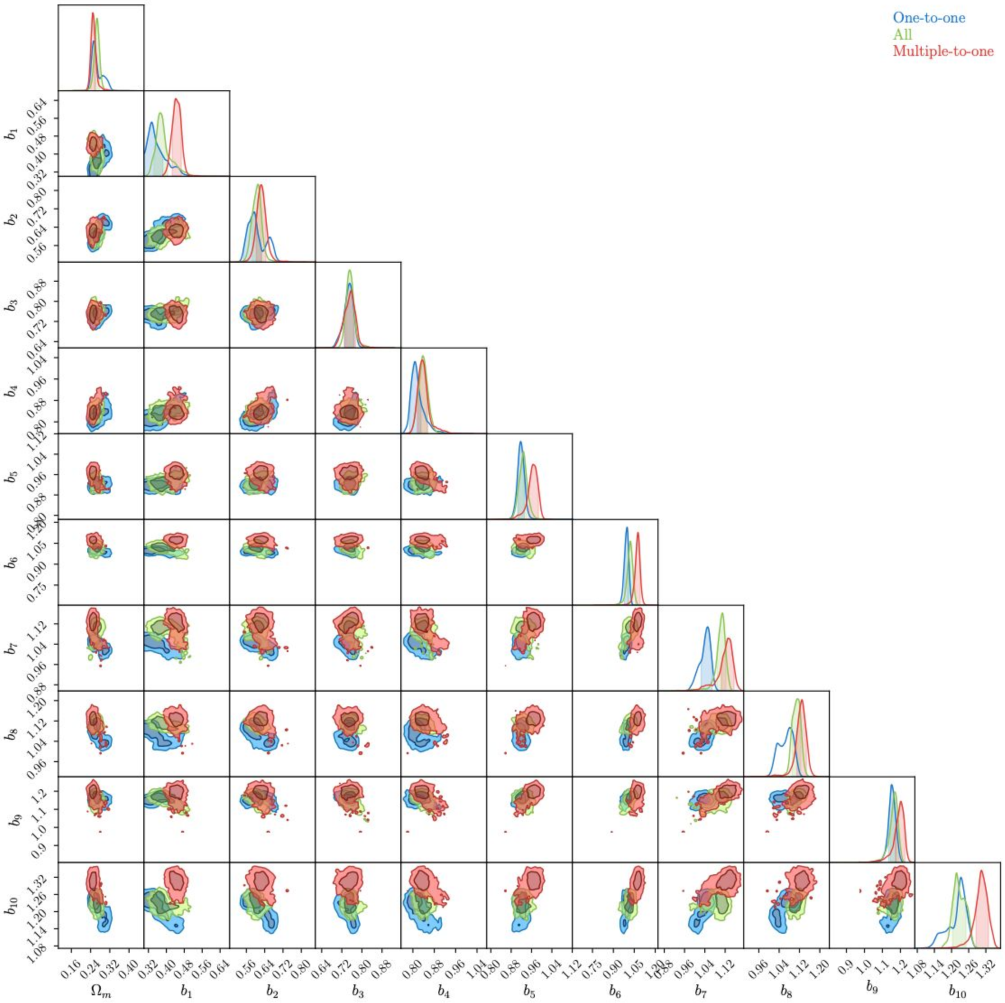

FIG. 5. Markov chain Monte Carlo estimation of Ω_M, calculating a different bias b for each redshift slice. The blue contours represent the one-to-one fit, the red the exclusively many-to-one fits, and the green the entire data set that includes both.

TABLE I. Estimated values for Ω_M and the galaxy bias b_i for each redshift bin *i*.

	Ω_M	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
1 to 1	$0.26^{+0.03}_{-0.01}$	$0.34^{+0.04}_{-0.02}$	$0.59^{+0.05}_{-0.03}$	$0.75^{+0.02}_{-0.02}$	$0.81^{+0.02}_{-0.02}$	$0.92^{+0.01}_{-0.01}$	$1.00^{+0.01}_{-0.02}$	$1.05^{+0.02}_{-0.02}$	$1.07^{+0.02}_{-0.04}$	$1.16^{+0.02}_{-0.02}$	$1.23^{+0.02}_{-0.04}$
2+ to 1	$0.25^{+0.01}_{-0.01}$	$0.44^{+0.02}_{-0.02}$	$0.62^{+0.02}_{-0.02}$	$0.76^{+0.02}_{-0.02}$	$0.83^{+0.01}_{-0.01}$	$0.97^{+0.02}_{-0.02}$	$1.08^{+0.01}_{-0.01}$	$1.13^{+0.02}_{-0.02}$	$1.13^{+0.02}_{-0.02}$	$1.20^{+0.02}_{-0.02}$	$1.31^{+0.02}_{-0.02}$
All	$0.27^{+0.01}_{-0.01}$	$0.37^{+0.02}_{-0.02}$	$0.60^{+0.02}_{-0.02}$	$0.75^{+0.01}_{-0.01}$	$0.84^{+0.01}_{-0.02}$	$0.93^{+0.01}_{-0.02}$	$1.02^{+0.01}_{-0.01}$	$1.11^{+0.01}_{-0.02}$	$1.11^{+0.01}_{-0.01}$	$1.17^{+0.02}_{-0.02}$	$1.23^{+0.02}_{-0.02}$

IV. CONCLUSION

With the expected increase in the statistical power of the LSST with respect to previous surveys, the impact of systematic uncertainties will play a larger role. In particular, blending seems like a potential dominant source of uncertainty, which may lead to biases in the inferred shapes, fluxes and number of counts. These biases can affect the inferred cosmological information from LSST. Thus, careful analysis should be performed in order to quantify and potentially mitigate these effects. Blending, in particular, is difficult to quantify in real data, since there is no ground truth information available, and realistic end-to-end simulations such as DC2 offer a unique opportunity to characterize its effects. To this end we used the DC2 data, matched inputs and outputs, computed the correlation function, and obtained the best fit for the galaxy bias and Ω_M.

If the results were to show that the difference in the correlation function between the true, one-to-one correlation and the many-to-one observations resulting from blending affects the calculated parameters, we would need to account for this bias in measurement. However, the difference is largely negligible for both parameters. The variation in the galaxy bias across redshift slices is small enough to be attributed to noise or other error, as shown in table I. The Ω_M values appear distinct when comparing the one-to-one fits to the exclusively many-to-one fits at high z, but this is not significant when comparing the one-to-one fits with all the data (as it will actually be observed). We can see in figure 5 how the corresponding blue and green plots mostly overlap and how the red exclusively many-to-one fit tends to be shifted. But although this is acceptable within the current error, there may be an issue once the error is reduced when using the full LSST footprint. It is also unclear whether the differences in the bias arise from the deblending algorithm used by the LSST Science Pipelines or as a natural effect of the geometric overlap of objects. Further investigation can be done to thus improve our estimates.

V. ACKNOWLEDGEMENTS

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

I thank those working at Fermilab who helped to organize my virtual internship amid the COVID-19 crisis; I really appreciate the opportunity to be able to pursue this project under these unique circumstances. I would especially like to thank Javier Sanchez, who supervised my project and provided attentive technical support throughout my internship at Fermilab.

VI. REFERENCES

Peter Melchior, Fred Moolekamp, et al., "SCARLET: Source separation in multi-band images by Constrained Matrix Factorization", Astronomy & Computing, (2018).

Javier Sánchez, Ismael Mendoza, et al., "Olber's Paradox Revisited- Effects of Overlapping Sources on Cosmic Shear Estimation: Statistical Sensitivity and Pixel-Noise Bias," (unpublished).

William A. Dawson, Michael D. Schneider, et al., "The Ellipticity Distribution of Ambiguously Blended Objects," arXiv:1406.1506v2, (2013).

DESC Collaboration et al., "The LSST DESC DC2 Simulated Sky Survey," (unpublished); "The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document," (2018).

Danila Korytov, Andrew Hearin, et al., "CosmoDC2: A Synthetic Sky Catalog for Dark Energy Science with LSST," Astrophysical Journal Supplement, (2019).

P. J. E. Peebles, "The Galaxy and Mass *N*-Point Correlation Functions: a Blast from the Past," (1980).

Jarvis, Bernstein, & Jain, MNRAS, 352, 338, (2004).

Nora Elisa Chisari, David Alonso, et al., "Core Cosmological Library: Precision Cosmological Prediction for LSST," arXiv:1812.05995v2, (2019).

M. H. Quenouille, "Approximate Tests of Correlation in Time-Series," Journal of the Royal Statistical Society 11 (1), 68-84 (1949).

Daniel Foreman-Mackey, David W. Hogg, et al., "emcee: The MCMC Hammer," Astronomical Society of the Pacific 125, 306–312 (2013).