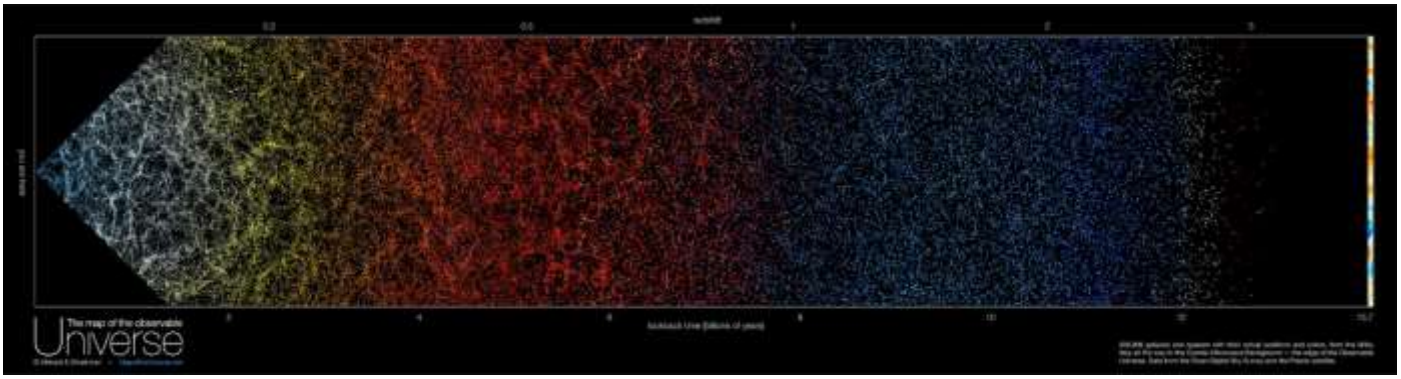
Testing the FLRW model with large-scale structure estimators



Pedro Miguel Silva Fanha
Master in Astronomy and Astrophysics
Department of Physics and Astronomy
Faculty of Sciences of the University of Porto
2025



Testing the FLRW model with large-scale structure estimators

Pedro Miguel Silva Fanha

Dissertation carried out as part of the Master in Astronomy and Astrophysics

Department of Physics and Astronomy

2025

Supervisor
José Fonseca, Researcher, Institute of Astrophysics and Space Sciences (IA)

Co-supervisor
António da Silva, Assistant Professor, Faculty of Sciences of the University of Lisbon (FCUL)

Acknowledgements

Developing this dissertation has not been easy - this work has been the biggest undertaking of my academic life (so far, hopefully). Nonetheless, I realize how incredibly lucky I am to be able to get here and to be able to have finished it. First and foremost, because as I have discovered, not everyone gets the chance to have supervisors who are present, who support you and root for you, and who do everything in their power to help you, guide you and make sure you succeed. I am very grateful to them, and to others who had no obligation to help in any way but still did - in particular, thanks to José Pedro Mimoso, professor and researcher at the University of Lisbon.

Second, I must thank my close family - my mom, my maternal grandparents, my brother, and my godparent - who have supported me unconditionally throughout all my childhood, my bachelor's and my master's, working hard to make sure I always had everything I ever needed without question. I do not think I will ever be able to repay them for this.

Finally, many, many thanks to all of my friends. I will refrain from mentioning them by name for fear of forgetting anyone or of not being able to properly acknowledge how much they mean to me, but I hope, when and if they read this, that they all are and always will be in my heart.

Resumo

O Modelo Cosmológico Padrão assenta no **princípio cosmológico**, o qual afirma que o Universo é homogéneo e isotrópico a grandes escalas. Enquanto que a isotropia já foi testada em certa medida, a homogeneidade é mais difícil de investigar, havendo ainda um número limitado de testes a esta propriedade. Esta questão é particularmente importante de responder tendo em conta as tensões observacionais do momento.

Nesta dissertação, desenvolvo e valido uma metodologia consistente para estimar a **escala angular de transição para a homogeneidade**, θ_H, analisando a dimensão fractal de catálogos de galáxias. Visto que já foi demonstrado que esta escala existe, é essencial agora comparar as escalas previstas por diferentes modelos teóricos com as escalas observadas.

Ao contrário das abordagens existentes na literatura, nem sempre consistentes entre si em relação a definições e em que muitas vezes não consideram todas as fontes possíveis de incerteza nas medições de θ_H, o método proposto combina os perfis de dimensão fractal com estimação de incerteza através de reamostragem do catálogo de dados, obtendo-se assim um procedimento robusto e interpretável.

A metodologia é aplicada ao catálogo de galáxias simulado MICECAT. Comparo diferentes abordagens para a determinação de θ_H e as incertezas associadas, chegando ao método final de reamostragem do catálogo de dados. Os resultados confirmam as expectativas da tendência na relação entre *redshift* z e θ_H, incluindo a grande variabilidade que se obtém nestas medições a baixos *redshifts* ($z \lesssim 0.7$), e as medições estáveis/de baixa variância a altos *redshifts* ($z \gtrsim 0.7$). Mostro ainda que da divisão de um catálogo em várias pequenas subregiões obtêm-se resultados para os θ_H consistentes com os θ_H da análise do catálogo todo como um só, e verifico que a contaminação das medições de θ_H pelas escalas das oscilações acústicas bariónicas (BAO) pode ser mitigada através da escolha ponderada do tamanho do *bin* de *redshift* e da seleção em magnitude dos objetos no catálogo.

De forma geral, este trabalho estabelece uma abordagem consistente, rápida, simples e robusta de testar a homogeneidade cósmica com os catálogos de galáxias de larga-escala já existentes e os previstos para um futuro próximo, tais como os do Euclid e DES, e identifica fatores cruciais a considerar em análises futuras deste tipo.

Palavras-chave: física, astrofísica, cosmologia, estrutura de larga escala, dimensão fractal, modelo cosmológico, escala de homogeneidade angular, reamostragem

Abstract

The Standard Model of Cosmology rests on the **cosmological principle**, which states that the Universe is homogeneous and isotropic at large scales. While isotropy has already been tested to some extent, homogeneity is harder to probe and direct tests of homogeneity remain limited. This question is particularly important to address considering the current observational tensions.

This dissertation develops and validates a consistent methodology for estimating the **angular scale of transition to homogeneity**, θ_H, using fractal dimension analysis of galaxy surveys. Given that such scale has been found to exist, it is essential now to compare the expected scales from different theoretical models to observed scales.

Unlike existing approaches, which are not all consistent with each other on definitions and often do not consider all sources of uncertainty on the measurements of θ_H, the proposed method combines fractal dimension profiles with survey resampling-based error estimation, providing a robust and interpretable framework.

The methodology is applied to the MICECAT mock galaxy catalogue. I compare different approaches for determining θ_H and its associated uncertainties, landing on a final survey resampling-based approach. Results confirm expected redshift-dependent trends in θ_H, including high variability at low redshifts ($z \lesssim 0.7$) and stable, low-variance measurements at higher redshifts ($z \gtrsim 0.7$). I also show that splitting surveys into multiple small subregions yields results for θ_H consistent with full-survey analyses, and verify that contamination by baryon acoustic oscillations (BAO) scales can be mitigated through careful redshift binning and magnitude selection.

Overall, this work establishes a fast, consistent, and reliable approach to testing cosmic homogeneity with existing and upcoming large-scale surveys such as Euclid and DES, and identifies key factors to consider in future analyses.

Keywords: physics, astrophysics, cosmology, large-scale structure, fractal dimension, cosmological model, angular homogeneity scale, resampling

Table of Contents

List of Figures vii
List of Abbreviations ix

1. Introduction 1
2. Theoretical framework 8
   2.1. The Standard Model of Cosmology 8
   2.2. Relaxing the constraint on homogeneity 10
   2.3. Tests of homogeneity 12
      2.3.1. Fractal dimension in three dimensions 12
      2.3.2. Projection to two dimensions 14
      2.3.3. Connection with the angular correlation function 14
      2.3.4. Angular homogeneity scale 15
      2.3.5. Numerical estimators of the fractal dimension 16
3. Methods and improvements 19
   3.1. The state-of-the-art in angular homogeneity scale measurements 19
   3.2. Some considerations on estimators, binning, randoms, and noise 25
      3.2.1. Choice of estimator 25
      3.2.2. Angular separation binning 25
      3.2.3. Number of bins 25
      3.2.4. Random catalogues 26
   3.3. A first proposed method: the covariance matrix-based approach 28
      3.3.1. Estimating uncertainty: covariance matrices 28
      3.3.2. Two different methods for $D_2(\theta)$ 29
      3.3.3. SR as a fitting strategy 30
      3.3.4. Obtaining θ_H 31
      3.3.5. Estimating θ_H uncertainty 32
      3.3.6. Parameter covariance estimation methods 34
   3.4. A second, final proposal: the survey resampling approach 37
   3.5. Additional consideration: BAO scales in the angular correlation function 40
4. Implementation details 42
   4.1. Pair counting codes 42
   4.2. Generating random catalogues 45
      4.2.1. Generating uniform randoms on the spherical surface 46
      4.2.2. Restricting to rectangular boxes 48

4.3. Regularization of covariance matrices 49
4.3.1. Diagonal loading 50
4.3.2. Choice of λ 50
5. Method evaluation 51
5.1. Mock surveys 51
5.1.1. Survey subdivision into regions 52
5.1.2. Uncertainty estimation via subsampling 57
5.2. Symbolic Regression: D_2 covariance matrix regularization 62
5.3. Testing the approaches to θ_H estimation 64
5.4. Comparison of uncertainty estimation methods 70
5.5. Robustness of θ_H estimates across survey area and redshift 71
5.6. Analytic profiles of D_2 74
6. Conclusion 77

Bibliography 80

List of Figures

1.1. View through the NASA/ESA/CSA James Webb Space Telescope's NIRCam (Near-Infrared Camera) instrument of a 50 light-years-wide portion of the Milky Way's dense centre. 2
1.2. Side-by-side comparison of NGC 346 from Hubble vs JWST. 2
1.3. Galaxy distribution from LCRS. 4
1.4. Comparison between homogeneous, non-isotropic configuration and isotropic but non-homogeneous configuration. 5
1.5. Visual proof of global isotropy when isotropy about two points A and B exists. 6

2.1. Example projected matter distribution in an LTB void model. 11
2.2. Examples of different fractal patterns. 12
2.3. An example $D_2(\theta)$ profile. 16

3.1. Flow diagram of the procedure in Gonçalves et al. [44] to estimate the angular homogeneity scale 22
3.2. Flow diagram of the procedure in Avila et al. [46] and [45] to estimate the angular homogeneity scale. 23
3.3. Flow diagram of the procedure in Andrade et al. [43] to estimate the angular homogeneity scale 24
3.4. Effect of log-spaced vs linear-spaced angular separation binning in D_2 curves. 26
3.5. Same as Figure 3.4 but comparing the impact of the number of bins on D_2 profile. 27
3.6. Variability of D_2 curves from using different random catalogues. 27
3.7. Procedure to transform SR expressions into parameterized forms for fine-tuning. 33
3.8. Flow diagram of the SR approach using the **indirect** method to estimate the angular homogeneity scale and uncertainty. 33
3.9. Flow diagram of the SR approach using the **direct** method to estimate the angular homogeneity scale and uncertainty. 34
3.10. Flow diagram of the resampling approach using the **direct** or **indirect** methods to estimate the angular homogeneity scale and uncertainty. 39

4.1. Runtime of different pair counting codes as a function of catalogue size N (number of objects). 45
4.2. Spherical coordinates ϕ and θ as used in description of random catalogue generation. 47
4.3. Example random catalogue generated using algorithm presented in the thesis. 48

5.1. Snapshots from the MICECAT catalogue within the completeness region used in this dissertation. Each plot shows one redshift bin of width $\Delta z = 0.01$. 53
5.2. Distribution on the sky of the 16 regions of 600 deg^2 and of the 16 regions of 1200 deg^2 selected from MICECAT within the completeness region (example for $z = 0.3$). Because of the overlap between regions, many are partially or completely hidden. 56
5.3. Example of region subsampled into 100 roughly equal-area patches by `TreeCorr` for uncertainty estimation. 57
5.4. Comparison of correlation matrix obtained for one $D_2(\theta)$ curve for the 'full area' region of MICECAT using either jackknife resampling or bootstrap resampling. 59
5.5. Same as Figure 5.4 but comparing the correlation matrices obtained for one $D_2(\theta)$ curve for the 'full area' region vs. the '1200 deg^2' region of MICECAT using jackknife resampling. 61

5.6. Eigenvalues of the covariance matrix for $D_2(\theta)$ from one single, random region used in this work. 63
5.7. Percent relative residuals of regularized correlation matrix compared to original correlation matrix. 63
5.8. θ_H measurements using SR fitting approach - indirect method. 65
5.9. θ_H measurements using SR fitting approach - direct method. 65
5.10 θ_H measurements using resampling approach - indirect method. 66
5.11 θ_H measurements using resampling approach - direct method. 67
5.12 D_2 profiles from jackknife resamples along with the corresponding spline fits, for the resampling approach. 68
5.13 Example histogram of θ_H measurements obtained for each jackknife resample shown in Fig. 5.12. 68
5.14 Uncertainties estimated on θ_H measurements for the full area region using different methods - linear approximation and nested sampler (from SR approach, described in subsection 3.3.6) and resample, applying the direct method for θ_H estimation. 70
5.15 Histograms of bootstrap distributions from θ_H measurements plus uncertainties via the resample approach with the direct method over many different regions of area 600 deg^2. 72
5.16 Histograms of bootstrap distributions from θ_H measurements plus uncertainties via the resample approach with the direct method over many different regions of area 1200 deg^2. 73
5.17 Comparison of the bootstrap mean θ_H of many regions of 600 deg^2 and 1200 deg^2 vs the 'full area' region, along with the theoretical expected BAO scale. 74
5.18 Plot of analytic profiles for D_2 obtained for different redshift bins for the full area region of MICECAT 2. 75

List of Abbreviations

2PACF two-point angular correlation function. 16

AIC Akaike Information Criterion. 22

ALFALFA Arecibo Legacy Fast ALFA Survey. 19, 20

BAO Baryon Acoustic Oscillations. v, 24, 40, 41

CDF Cumulative Distribution Function. 47, 49

CMB Cosmic Microwave Background. 3, 12

DES Dark Energy Survey. 20

FLRW Friedmann-Lemaître-Robertson-Walker. 8– 11

GR General Relativity. 8, 9

HAM Halo Abundance Matching. 51

HOD Halo Occupation Distribution. 51

JWST James Webb Space Telescope. vii, 1

LCRS Las Campanas Redshift Survey. vii, 3, 4, 6

LRG Luminous Red Galaxies. 19, 20

LS Landy-Szalay. 22, 23, 25, 29, 31, 37, 38, 75

LTB Lemaître- Tolman-Bondi. 10, 11

MICE-GC Mice Grand Challenge. 51

MICECAT MICE-Grand Challenge Galaxy Catalogue. vii, viii, 26, 52, 53, 56, 59, 61, 75

RMSE root mean squared error. 22

SDSS Sloan Digital Sky Survey. 3, 19, 20

SN supernova. 24

SR Symbolic Regression. v, vii, 30– 34

WMSE weighted mean squared error. 31

1. Introduction

When we observe the night sky with the naked eye, we can identify a variety of objects: the Moon, scintillating stars, and, with some luck, a few planets. We can see constellations, forming apparent patterns that once inspired our ancestors to name and group stars, and under particularly dark skies, parts of our own Milky Way become visible. Under such conditions, we may also detect nearby galaxies and nebulae. When studied in more detail, these objects reveal a rich diversity. We now understand that Earth is just one of several planets orbiting the Sun and, in the same way, the Moon revolves around Earth. The Solar System contains planets of vastly different compositions, atmospheres, surface temperatures, and sizes. The Sun is only one among billions of stars and is far from exceptional. Moreover, our Solar System resides in one of the spiral arms of the Milky Way galaxy, and even this galactic structure is only one type among many: some galaxies exhibit well-defined spiral arms, others are elliptical, while still others have irregular morphologies. Throughout the cosmos, black holes, exploding stars, nebulae - clouds of dust and gas shaped by the stellar processes and fundamental forces that rule our Universe - and so many other objects further enrich this diversity of structures.

Telescopes (like the James Webb Space Telescope (JWST)* or the Hubble Space Telescope†) equipped with sophisticated detectors, sensitive across the electromagnetic spectrum, have enabled us to perceive objects well beyond the reach of the naked eye (Fig. 1.1, Fig. 1.2). Yet, expanding our scope of observation reveals that even the grandest of these structures are small compared to the overall size of the Universe.

For illustration: if the entire Solar System were scaled down to the length of a football field, the Sun would be a few centimeters in diameter, probably less than half the length of any of your fingers. Even at this scale, the Milky Way galaxy would be so large that it would defy intuitive comprehension. At real scales, light itself, the fastest entity in the Universe, takes approximately 8.3 minutes to travel from the Sun to Earth but requires roughly 87,500 years [1] to cross the Milky Way. The nearest major galaxy, Andromeda, is 2.5 million light-years away. On even larger scales, galaxies cluster into groups: the Local Group, which includes the Milky Way, has a diameter of about 10 million light-years.

*The JWST is a space telescope designed to conduct infrared astronomy. As the largest telescope in space, it is equipped with high-resolution and high-sensitivity instruments, allowing it to view objects too old, distant, or faint for the Hubble Space Telescope.

†The Hubble Space Telescope is a space telescope developed by NASA and ESA that was launched into low Earth orbit in 1990 and remains in operation.



Figure 1.1: The full view of the NASA/ESA/CSA James Webb Space Telescope's NIRCam (Near-Infrared Camera) instrument reveals a 50 light-years-wide portion of the Milky Way's dense centre. An estimated 500,000 stars shine in this image of the Sagittarius C (Sgr C) region, along with some as-yet unidentified features. *Credit: NASA, ESA, CSA, STScI, S. Crowe (UVA)*

Image description [*Credit: NASA, ESA, CSA, STScI, S. Crowe (UVA)*]: In a field crowded with stars, a funnel-shaped region of space appears darker than its surroundings with fewer stars. It is wider at the top edge of the image, narrowing towards the bottom. Toward the narrow end of this dark region a small clump of red and white appears to shoot out streamers upward and left. A large, bright cyan-colored area surrounds the lower portion of the funnel-shaped dark area, forming a rough U shape. The cyan-coloured area has needle-like, linear structures and becomes more diffuse in the center of the image. The right side of the image is dominated by clouds of orange and red, with a purple haze.

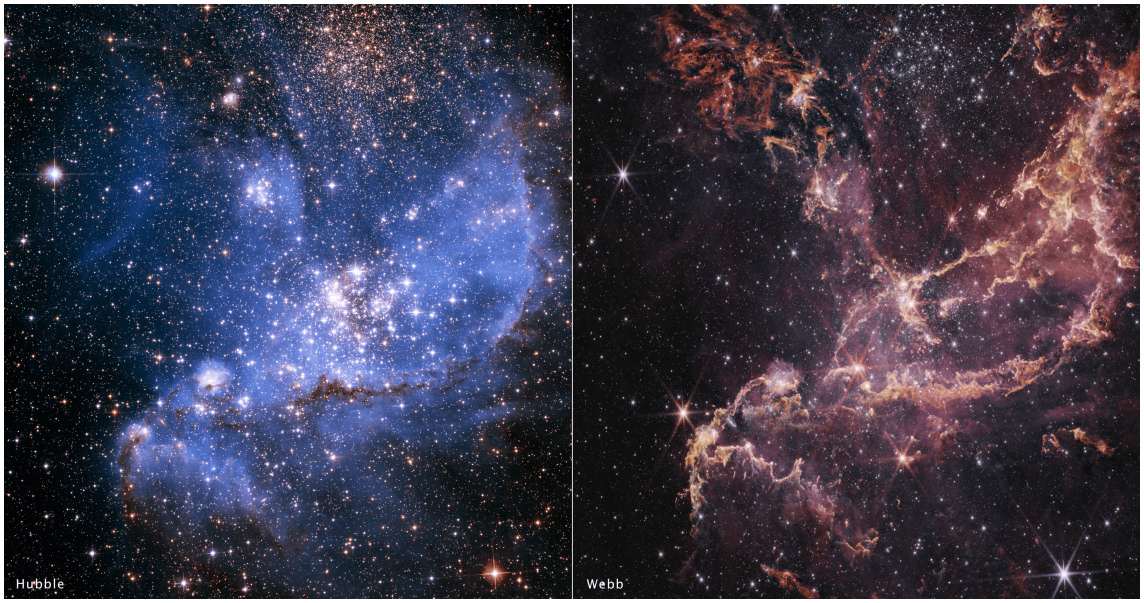


Figure 1.2: This side-by-side comparison shows a Hubble image of the massive star cluster NGC 346 (left) versus a Webb image of the same cluster (right). While the Hubble image shows more nebulosity, the Webb image pierces through those clouds to reveal more of the cluster's structure. NGC 346 has a relative lack of elements heavier than helium and hydrogen, making it a good proxy for stellar environments in the early, distant universe. *Credit: NASA, ESA, CSA, STScI, O. C. Jones (UK ATC), G. De Marchi (ESTEC), M. Meixner (USRA), A. Nota (ESA)*

In contrast, structures like the Huge-LQG (Huge Large Quasar Group), discovered in 2012 [2], extend across 4 billion light-years - nearly 400 times larger than the Local Group. The observable Universe is currently estimated to span 93 billion light-years, though the true Universe may be even larger - and we may never be able to know how big it truly is. These immense scales illustrate both the vastness of space and the persistence of structure wherever we look. In the end, despite the wide variety of objects in the Universe, when observed on sufficiently large scales, the Universe appears remarkably uniform.

One of the first observational efforts to map the large-scale distribution of galaxies was the Las Campanas Redshift Survey (LCRS), which began in 1991 and catalogued 26,418 galaxies up to redshift $z = 0.2$ (*redshift* can be thought of as a measure of distance to an object in the sky and at the same time as a measure of time into the past, with $z = 0$ representing the present time; for comparison, redshift $z_* \approx 1000$ corresponds to the point in time where the Universe cooled down enough for electrons and protons to form hydrogen for the first time [3] *). The LCRS revealed, for the first time, a large-scale galaxy pattern that appeared broadly uniform across the sky. Fig. 1.3 shows the pattern of galaxies in different layers of declination in the sky, at different redshifts. This survey was followed by the Sloan Digital Sky Survey (SDSS), initiated in 2000 and still ongoing, which produced more comprehensive and detailed maps of galaxy distributions at higher redshifts. The SDSS [4] has been invaluable in extending our view further back in time and further into space, providing a wealth of data that strongly supports the notion of uniformity at large scales.

Beyond galaxy surveys, another major discovery further reinforced the idea of large-scale uniformity in the Universe: the Cosmic Microwave Background (CMB). Discovered in 1965, the CMB is relic radiation from the epoch when the expanding Universe had cooled sufficiently for photons to decouple from matter and travel freely across the cosmos, largely unimpeded. Because of this, it preserves valuable information about the conditions of the early Universe. Subsequent analyses [6, 7] of the small temperature fluctuations of the CMB provided strong evidence for an early Universe which at large scales looked uniform in whichever direction we observed it (though there are also studies that challenge these conclusions, see e.g. [8] and [9], the latter of which leads you to the main research that has been done on the topic).

*It is important to note that redshift is not a linear clock into the past. The look-back time spanned by Δz grows rapidly at low redshift and shrinks dramatically at high redshift.

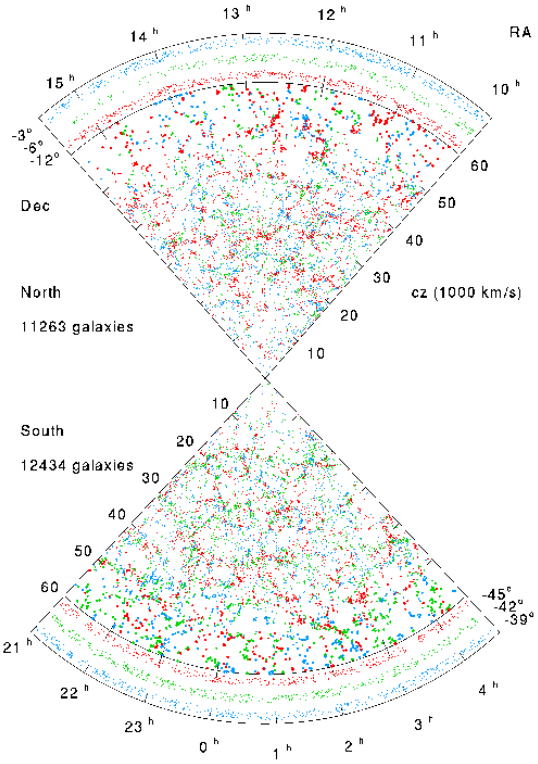


Figure 1.3: A plot of each galaxy from the LCRS (Source: [5]) showing its right ascension (RA), in hours, and recession velocity (cz), as a multiple of 1000 km/s. The color of the dot shows which declination stripe (Dec) the galaxy is in. The dots for more distant galaxies are bigger to allow for the fact that only a small fraction of all galaxies are luminous enough to be seen at great distances.

Although these empirical results are very recent, the philosophical expectation that the Universe should appear similar everywhere has much deeper roots. This perspective reflects an attitude of humility: Humanity and its place in the cosmos are not exceptional. In modern cosmology, this idea is called the Copernican principle, and is embedded in the so-called **cosmological principle**, which can be formally stated as:

> Viewed on a sufficiently large scale, the properties of the Universe are the same for all observers. [10]

From this definition emerge two distinct but related properties: **isotropy** and **homogeneity**.

- *Isotropy* refers to the condition that, from a given location in the Universe, physical observables depend only on the radial distance r from the observer, not on direction. If an observer measures a property at distance r in one direction and repeats the measurement at the same distance in another direction, the results should be identical;

- *Homogeneity*, on the other hand, means that the spatial distribution of matter and physical properties are invariant under translation. Observers at two different locations would record statistically identical conditions.

The two conditions are related, but not equivalent. Homogeneity does not guarantee isotropy, since one could imagine a uniform matter distribution that prefers a given direction (e.g., aligned along the x-axis of a Cartesian frame), in which case rotational symmetry is not satisfied. Similarly, isotropy about a single point does not ensure isotropy everywhere - we could well be at the center of a very large radial outflow of matter, and from elsewhere in such a universe, the sky would not appear isotropic. However, it can be shown that if isotropy holds at two distinct points, then we must necessarily have isotropy everywhere, which implies homogeneity (see Fig. 1.5 for an intuitive, visual proof).

A comparison of the two concepts, homogeneity and isotropy, is illustrated in Fig. 1.4.

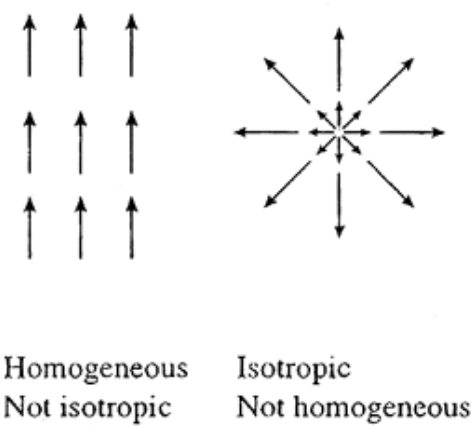


Figure 1.4: Comparison of an example configuration that is homogeneous but not isotropic to one that is isotropic but not homogeneous. *Credit: Unknown source*

The cosmological principle has also received criticism: Karl Popper (philosopher of science, 1902–1994) argued that it turns “our lack of knowledge into a principle of knowing something” [12], meaning that the assumption of uniformity in cosmology was adopted so widely not because it could be proven, but because it could not be falsified. He regarded the cosmological principle as a dogma that should “not have been proposed” [12], emphasizing that scientific theories must be open to falsification, a key idea he developed in *The Logic of Scientific Discovery* [13].

Testing the validity of the cosmological principle is not straightforward. A robust test of isotropy would require observations from multiple viewpoints, which is fundamentally impossible: we can only observe from Earth. On the other hand, failing to observe isotropy

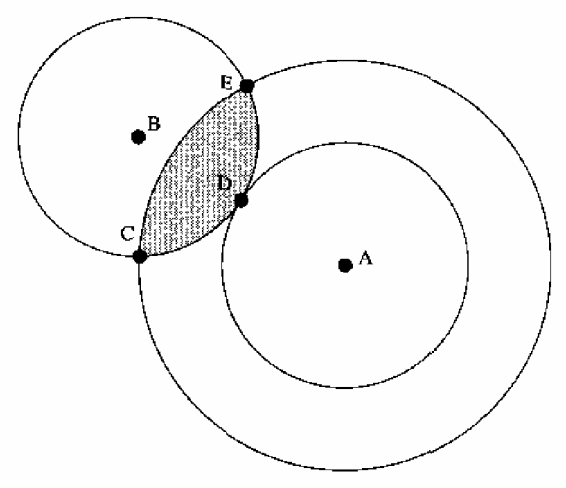


Figure 1.5: An intuitive, visual proof that isotropy about two points A and B guarantees homogeneity. From isotropy about B, the density is the same at each of C, D, E. By constructing spheres of different radii about A, the shaded zone is swept out and shown to be homogeneous. By using large enough shells, this argument extends to the entire universe. *Credit: Peacock [11], Figure 3.1*

from Earth alone would suffice to undermine the principle. At present, analyses of isotropy have yielded mixed results, with no consensus in the scientific community [14–18]. Consequently, one must turn to tests of homogeneity, which can be probed, for example, through large-scale surveys. By developing robust statistical tools to characterize homogeneity, we can both test the fundamental assumptions underpinning cosmology and evaluate competing cosmological models. This issue is particularly important given the current tensions between observational data and the standard cosmological model (e.g., the Hubble tension [19, 20]).

With the advent of increasingly comprehensive galaxy surveys - much larger and more precise than the early LCRS - we now have the unprecedented ability to rigorously test the homogeneity of the matter distribution in the Universe. It is within this context that the present dissertation is situated. The core aim of this work is to explore cosmological model-independent methods for testing large-scale homogeneity that take full advantage of the next generation of galaxy surveys. With that in mind, the dissertation begins in Chapter 2 by describing the standard cosmological model and briefly presenting alternative models of cosmology that arise when the constraint of homogeneity is relaxed, as well as explaining the approach to testing homogeneity that will be explored; in Chapter 3 I review the existing literature using this approach, the methods and data used, and the results obtained, building upon that to develop a methodology to be applied consistently in all future studies; in Chapter 4 I go at length about some non-core issues and questions that had to be addressed or worked on in order to test and develop my work; finally, in

Chapter 5 I present the results of different methodological choices when applied to simulated surveys and discuss the best strategies to use, summarizing all of my work and findings in Chapter 6.

2. Theoretical framework

The previous chapter introduced the cosmological principle and emphasized the importance of testing large-scale homogeneity. In this chapter, I provide the theoretical framework necessary for such tests. I begin by outlining the standard cosmological model, which is based on the Friedmann-Lemaître-Robertson-Walker (FLRW) metric and relies directly on the assumption of homogeneity and isotropy. In addition, I define and describe the cosmological parameters that can be observed or indirectly calculated and that characterize the evolution of our universe. I then discuss alternative cosmological models that arise when the assumption of homogeneity is relaxed, highlighting how these can lead to different predictions for the large-scale structure of the Universe. Finally, I briefly review existing methods used to probe homogeneity and introduce the framework adopted in this work, which is based on the analysis of the fractal dimension in two dimensions (projected sky coordinates or angular coordinates).

2.1. The Standard Model of Cosmology

Under the assumptions of the cosmological principle, a remarkably successful framework has been developed that accurately describes a wide range of cosmological observations: the *Standard Model of Cosmology* [21].

The starting point for deriving any cosmological model is General Relativity (GR), where the geometry of spacetime caused by the distribution of matter is described by a metric. In general form, a metric can be written as

$$ds^2 = c^2dt^2 - R^2(t)\left[f^2(r)dr^2 + g^2(r)d\Omega^2\right] \tag{2.1}$$

where the terms on the right-hand side of the equation correspond, in this order, to the temporal, radial, and transverse or angular components. c is the speed of light constant, t is time, r is the radial distance, and ds^2 corresponds to an interval in GR, a measure of distance in the four dimensions of space-time. The function $R(t)$ represents a time-dependent scale factor, and the specific forms of $f(r)$ and $g(r)$ depend on the symmetries imposed on spacetime.

Specializing to the case of a spatially homogeneous and isotropic Universe leads to the FLRW metric. Written in spherical coordinates, the metric takes the form [22, 23]

$$ds^2 = c^2dt^2 - a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2 d\Omega^2\right], \quad d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2 \tag{2.2}$$

where $a(t)$ is the scale factor describing cosmic expansion and k is a constant that encodes the spatial curvature. While different conventions exist for the units of k and $a(t)$, the key point is that the scale factor relates directly to the observable redshift z through [11]

$$a = \frac{1}{1+z}. \tag{2.3}$$

From this metric, the application of Einstein's equations of GR yields the **Friedmann equations**, which govern the evolution of the Universe as a function of $a(t)$. These can be written as [21, 24]:

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \tag{2.4}$$

$$\frac{\ddot{a}}{a} = \frac{\Lambda c^2}{3} - \frac{4\pi G}{3}(\rho + 3p) \tag{2.5}$$

where $H(t)$ is the Hubble parameter (with H_0 denoting its present-day value), G is Newton's gravitational constant, ρ is the total energy density in matter and radiation and p the corresponding isotropic pressure, and Λ is the cosmological constant, typically associated with dark energy (added in to describe an expanding Universe).

It is often convenient to cast the first of these equations in terms of dimensionless density parameters, normalized to the critical density of a flat Universe, [11]

$$\rho_{\text{crit},0} \equiv \frac{3H_0^2}{8\pi G} \tag{2.6}$$

For example, the present-day baryon density (where the $,0$ subscript denotes "present time") is expressed as:

$$\Omega_{b,0} \equiv \frac{\rho_b(t = \text{present})}{\rho_{\text{crit}}} \tag{2.7}$$

with analogous definitions for other components such as cold dark matter ($\Omega_{c,0}$), radiation ($\Omega_{r,0}$), curvature ($\Omega_{k,0}$), and dark energy ($\Omega_{\Lambda,0}$), such that $\sum_i \Omega_i = 1$ at any instant. Using these definitions, the Friedmann equation can be rewritten as [25]

$$H^2 = H_0^2\left[(\Omega_c + \Omega_b)\,a^{-3} + \Omega_r a^{-4} + \Omega_k a^{-2} + \Omega_\Lambda a^{-3(1+w)}\right] \tag{2.8}$$

where w parametrizes the equation of state of dark energy (more details can be found

in [26]). This expression highlights the role of the fundamental cosmological parameters that are constrained through observational studies.

2.2. Relaxing the constraint on homogeneity

If the condition of homogeneity is relaxed, one can construct a wide variety of alternative cosmological models that remain consistent with isotropy about a given observer. The simplest class of such inhomogeneous yet isotropic solutions is provided by the Lemaître-Tolman-Bondi (LTB) models [22, 27–30]. These are often referred to as void models, since they describe a spherically symmetric inhomogeneous Universe in which the observer is located at or near the center of a matter underdensity.

In the LTB framework, the metric takes the form, [31]

$$ds^2 = c^2dt^2 - \frac{a_{||}^2(t,r)}{1-\kappa(r)r^2}dr^2 + a_{\perp}^2(t,r)r^2d\Omega^2 \tag{2.9}$$

where two distinct scale factors appear: a radial one, $a_{||}(t,r)$, and a transverse (angular) one, $a_{\perp}(t,r)$. They are related through

$$a_{||}(t,r) = \frac{\partial}{\partial r}\left[a_{\perp}(t,r)r\right] \ . \tag{2.10}$$

Unlike in the FLRW case, the spatial curvature term κ is no longer a global constant but a function of the radial coordinate r. The lack of homogeneity introduces a fundamental difference: physical observables may depend on how the radial and transverse (perpendicular) directions evolve in these models. As a result, there are now two distinct Hubble parameters, corresponding to the radial and transverse expansion rates:

$$H_{\perp}(t,r) \equiv \frac{\dot{a}_{\perp}(t,r)}{a_{\perp}(t,r)}, \quad H_{||}(t,r) \equiv \frac{\dot{a}_{||}(t,r)}{a_{||}(t,r)} \ . \tag{2.11}$$

Analogous to the homogeneous case, one can also define dimensionless density parameters, but now allowed to vary with position:

$$\Omega_{\kappa}(r) = -\frac{\kappa(r)}{H_{\perp}^2(t,r)}, \quad \Omega_m(r) = \frac{M}{H_{\perp 0}^2} \tag{2.12}$$

where M is a free function of r and $H_{\perp 0}$ is the present-day transverse expansion rate at radius r. With these definitions, the LTB analogue of the Friedmann equation can be written as

$$\frac{H_{\perp}^2(t,r)}{H_{\perp 0}^2} = \Omega_m(r)a_{\perp}^{-3}(t,r) + \Omega_{\kappa}(r)a_{\perp}^{-2}(t,r) \ . \tag{2.13}$$

This formulation (presented in [31]) clearly illustrates the impact of relaxing homogeneity: the background quantities that are constants in FLRW cosmology become functions of the radial coordinate r. As a consequence, cosmic observables such as the Hubble parameter or matter density can vary depending on the observer's location in the inhomogeneous Universe. In practical terms, this means that the inferred values of cosmological parameters from observations of high-redshift sources depend both on their temporal separation (i.e. look-back time) and on their spatial location relative to our position, thus rendering the standard FLRW framework inaccurate if comparing to results from the local Universe.

An example of the resulting radial density profile in an LTB void model is shown in Fig. 2.1, where the central underdensity smoothly transitions to higher densities at large radii. For an FLRW model, there would be no central void, with the entire rectangle looking everywhere similar to the edges.

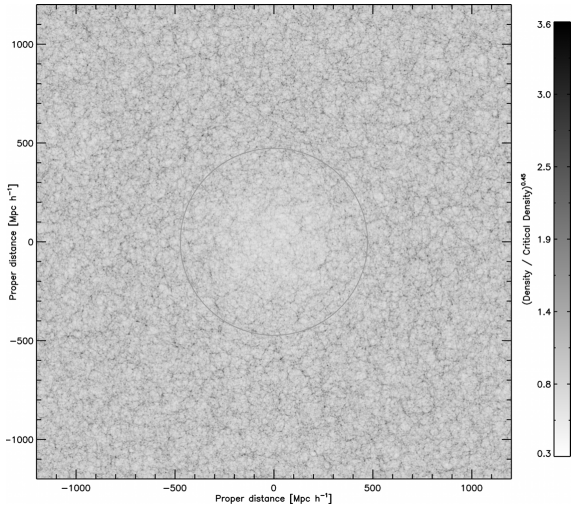


Figure 2.1: The projected matter distribution of the three-dimensional snapshot of a simulation of an LTB void model at $z = 0$, averaged over the axis perpendicular to the figure. Near the center of the void, not only the density is lower, but there is also significantly less structure than outside. *Credit: Alonso et al. [32], figure 1.*

Beyond LTB models, or the more general form of Stephani models, [30], we can have other types of cosmologies if we relax isotropy as well (or only isotropy). If we find the Universe to be homogeneous, one possible alternative class of cosmological models are the Bianchi models, which are spatially homogeneous but anisotropic solutions to Einstein's equations. On the other hand, if we indeed have inhomogeneity but LTB is not sufficient to describe observations, the Szekeres models generalize LTB by dropping the hypothesis of spherical symmetry (isotropy). Even more outlandish cosmologies are possible [18], though discussing or even presenting them is outside the scope of this work.

2.3. Tests of homogeneity

Apart from analyses of distance-redshift relations or studies of the uniformity of primordial density perturbations observed in the CMB [33–37], direct probes of the large-scale structure of the Universe provide an independent, data-driven approach to testing homogeneity.

The foundational motivation of these analyses is the idea that the matter distribution in the Universe can be described as a fractal pattern. A fractal is a statistical structure that exhibits self-similarity: its complexity appears similar when observed at different scales (examples of fractal patterns are shown in Fig. 2.2). A useful statistical measure of such complexity is the **fractal dimension**. In cosmological applications, this dimension characterizes how the number of pairs of objects (e.g., galaxies) increases as a function of scale.

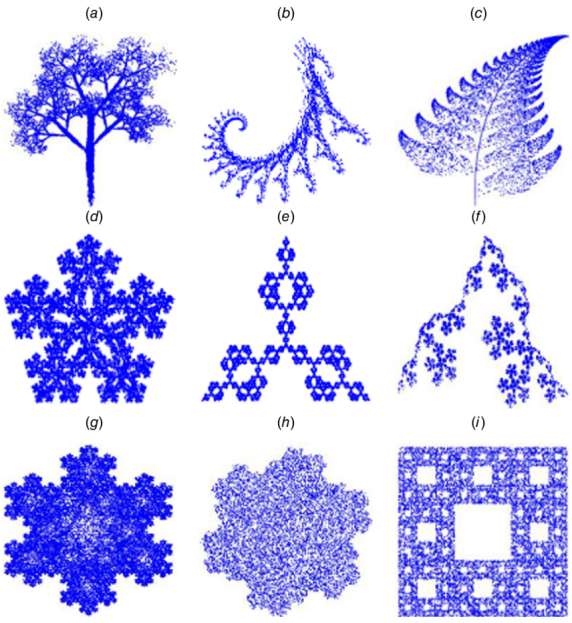


Figure 2.2: Examples of different fractal patterns. *Credit: Cao et al. [38]*

2.3.1 Fractal dimension in three dimensions

The framework is built on *counts-in-spheres*. Consider the number of pairs of objects found within a sphere of radius r, denoted $N(< r)$, compared to the number expected from a homogeneous (random) distribution, $N_R(< r)$. In an homogeneous and isotropic distribution free of survey artifacts, the expected scaling is:

$$N_R(< r) \propto r^3 \, . \tag{2.14}$$

In the real Universe, which is inhomogeneous on small scales, we expect instead:

$$N(< r) \propto r^{D_2} \tag{2.15}$$

where the fractal dimension D_2 quantifies deviations from homogeneity. By definition, $D_2 = 3$ corresponds to a homogeneous three-dimensional distribution.

It is useful to define the *scaled counts-in-spheres*:

$$\mathcal{N}(< r) \equiv \frac{N(< r)}{N_R(< r)} \, . \tag{2.16}$$

This ratio automatically incorporates survey geometry and completeness* effects, since the random catalogue reflects the survey selection function†.

In this case, D_2 is defined as:

$$D_2(r) \equiv \frac{d \, \ln N(< r)}{d \, \ln r} = 3 + \frac{d \, \ln \mathcal{N}(< r)}{d \, \ln r} \, . \tag{2.17}$$

At sufficiently large scales, if the Universe approaches homogeneity, then the change in complexity, given by the second term of the equation above, should approach 0, and so D_2 should asymptotically approach 3:

$$D_2(r) \to 3 \quad \text{as} \quad r \to \infty \tag{2.18}$$

In practice, one typically defines a finite threshold point beyond which we can say that D_2 is *sufficiently* close to 3 - we call this the scale of transition to homogeneity and we refer to it symbolically as R_H.

Previous studies have estimated this homogeneity scale, R_H, to exist and to be $\sim 60 - 80\,h^{-1}$ Mpc [39–41], though the precision of these values has been limited by the scale of the existing galaxy surveys. These results provide motivation for further tests. However, because distances in a three-dimensional distribution cannot be inferred without assuming a cosmological model, any such measurement of R_H is inherently model-dependent.

This sort of analysis is presented, for instance, in [42].

*Survey completeness is the fraction of objects in a given volume that are detected and included in a survey catalogue compared to the true number of objects in the sky, $C = \frac{N_{\text{detected}}}{N_{\text{true}}}$

†The selection function encodes the probability that an object with certain properties is observed and included in the survey.

2.3.2 Projection to two dimensions

An alternative, cosmological model-independent approach is to examine the *projected* distribution of objects in the sky, i.e. in two dimensions only. In this case, the procedure described before becomes a counts-in-spherical-caps analysis.

For a spherical cap of angular radius θ, the number of objects in the observational catalogue is $N(<\theta)$, while for a homogeneous distribution one theoretically expects

$$N_R(<\theta) = 2\pi\bar{\rho}(1-\cos\theta) \tag{2.19}$$

where $\bar{\rho}$ is the mean number density of pairs of objects in the sky. The scaled counts-in-spherical-caps are then:

$$\mathcal{N}(<\theta) \equiv \frac{N(<\theta)}{N_R(<\theta)} \ . \tag{2.20}$$

In analogy with the 3D case, the fractal dimension in two dimensions can be expressed as:

$$D_2(\theta) \equiv \frac{d\ \ln N(<\theta)}{d\ \ln\theta} = \frac{\theta\sin\theta}{1-\cos\theta} + \frac{d\ \ln\mathcal{N}(<\theta)}{d\ \ln\theta} \tag{2.21}$$

For sufficiently small angular scales, $D_2(\theta)$ asymptotically approaches 2 for a homogeneous distribution, whereas in general it will approach the curve

$$D_2^H \equiv \frac{\theta\sin\theta}{1-\cos\theta} \ . \tag{2.22}$$

This profile therefore provides a statistical test of the transverse (angular) homogeneity of the Universe.

2.3.3 Connection with the angular correlation function

The scaled number counts can be linked directly to the two-point angular correlation function $\omega(\theta)$. By definition, the conditional "probability" (or differential number counts) of finding any random object within a solid angle element $\delta\Omega$ at angular separation θ from one particular object is [25]:

$$\delta N = \bar{\rho}\delta\Omega\left[1+\omega(\theta)\right] \tag{2.23}$$

This alone defines a correlation function - it is a quantity that describes the excess probability of finding an object in the conditions described above *relative* to the homogeneous background with density $\bar{\rho}$. It can alternatively be connected to the matter density perturbations throughout the Universe like:

$$\omega(\theta) \equiv \langle\Delta(\hat{\boldsymbol{n}})\Delta(\hat{\boldsymbol{n}}')\rangle \tag{2.24}$$

where Δ is the matter density fraction, in excess of the background density $\bar{\rho}$ in some direction $\hat{\boldsymbol{n}}$,

$$\Delta(\hat{\boldsymbol{n}}) = \frac{\rho(\hat{\boldsymbol{n}}) - \bar{\rho}}{\bar{\rho}} \,. \tag{2.25}$$

Integrating the "probability" δN over a spherical cap yields:

$$N(<\theta) = \bar{\rho} \int d\Omega \, \left[1 + \omega(\theta)\right] \tag{2.26}$$

which is the expected number of pairs within radius θ (note that this formulation implicitly accounts for counts over the entire survey volume, rather than being restricted to pairs within a single spherical cap centered at one particular location). This leads to the expression:

$$\mathcal{N}(<\theta) = 1 + \frac{1}{1-\cos\theta} \int_0^\theta \omega(\theta') \sin\theta' \, d\theta' \tag{2.27}$$

and consequently, the fractal dimension becomes:

$$D_2(\theta) = \frac{\theta \sin\theta}{1-\cos\theta} + \frac{d}{d \, \ln\theta} \ln\left[1 + \frac{1}{1-\cos\theta} \int_0^\theta \omega(\theta') \sin\theta' \, d\theta'\right] \tag{2.28}$$

Because the correlation function quantifies matter clustering, its form depends strongly on the cosmological model and on redshift. $\omega(\theta)$ can also be measured for different tracers of structure, which may be biased with respect to the underlying dark matter distribution. Moreover, the resulting measurement is influenced by survey characteristics, including statistical limitations and the choice of redshift binning (the survey is typically sliced in small redshift intervals to minimize projection effects where objects from different points in time, with different clustering properties, are mixed).

2.3.4 Angular homogeneity scale

The fractal dimension profile $D_2(\theta)$ allows a clear definition of the **angular homogeneity scale**. A commonly used criterion in the literature is the so-called *1% criterion* [40, 41, 43–47], which defines the homogeneity scale θ_H as the angular radius at which $D_2(\theta)$ first reaches 1% from the homogeneous threshold value (a visual description is given in Fig. 2.3):

$$D_2(\theta_H) = 0.99 D_2^H(\theta_H) \tag{2.29}$$

While this criterion is somewhat arbitrary, it has the advantage of being survey-independent and therefore facilitates comparison across different datasets and against theoretical expected values. Alternative definitions have been proposed in the literature [48, 49], though

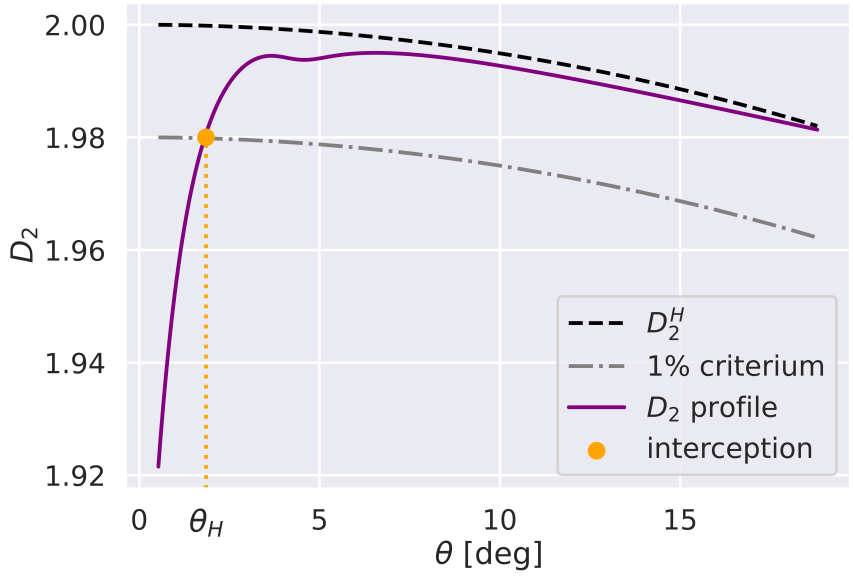


Figure 2.3: An example $D_2(\theta)$ profile is displayed (solid purple curve) for a Universe that is non-homogeneous at the smaller scales, along with the corresponding curve if the Universe were homogeneous at all scales (dashed black curve). The dash-dot-styled grey curve corresponds to the curve 1% away from homogeneity at any θ: the θ value of the first crossing of $D_2(\theta)$ (solid purple curve) from the left with this dash-dot-styled grey curve corresponds to θ_H. Profile obtained from our analysis (see Section 5).

in this work we adopt the 1% criterion and leave it as an open question whether it is the most adequate to use in future work.

2.3.5 Numerical estimators of the fractal dimension

Computing $\omega(\theta)$, $\mathcal{N}(<\theta)$, and $D_2(\theta)$ from observational data requires statistical estimators based on pair counts. For a catalogue containing the right ascension and declination of each object (binned in redshift), one can count the number of object pairs separated by θ, yielding $DD(\theta)$.

To emulate a homogeneous distribution with the same angular footprint, one generates a random catalogue, i.e. an artificial dataset with objects placed randomly across the survey area. Pairs counted in this catalogue yield $RR(\theta)$. One can also consider cross-pairs between the data and the random catalogue, $DR(\theta)$.

From these quantities, several estimators of the two-point angular correlation function (2PACF) have been developed. Examples include:

- Natural estimator:

$$\omega_N(\theta) = \frac{DD(\theta)}{RR(\theta)} - 1 \tag{2.30}$$

This estimator follows intuitively from the definition in Eq. 2.23.

- Hamilton estimator:

$$\omega_H(\theta) = \frac{DD(\theta) \times RR(\theta)}{DR(\theta)^2} \tag{2.31}$$

- Hewett estimator:

$$\omega_{He}(\theta) = \frac{DD(\theta) - DR(\theta)}{RR(\theta)} \tag{2.32}$$

- Davis-Peebles estimator:

$$\omega_{DP}(\theta) = \frac{DD(\theta)}{DR(\theta)} - 1 \tag{2.33}$$

- Landy-Szalay estimator:

$$\omega_{LS}(\theta) = \frac{DD(\theta) - 2DR(\theta) + RR(\theta)}{RR(\theta)} \tag{2.34}$$

Among these, the Landy-Szalay estimator is widely regarded as the most robust [50] and is the choice adopted in this thesis.

In all cases, the pair counts DD, DR, RR are normalized by the total number of pairs in their respective catalogues. This normalization ensures that the resulting correlation function does not depend on the absolute number of objects and that ratios are computed consistently, thereby avoiding biases that would arise from catalogues of differing sizes.

It follows from previous definitions that the scaled counts can be written in terms of $\omega_{LS}(\theta)$ as:

$$\mathcal{N}_{LS}(<\theta) = 1 + \frac{1}{1-\cos\theta}\int_0^\theta \omega_{LS}(\theta')\sin\theta'\, d\theta' \ . \tag{2.35}$$

Alternatively, one may use a direct approach, summing pair counts explicitly:

$$\mathcal{N}_{LS}(<\theta) = 1 + \frac{\sum_{\theta_i=0}^{\theta}\left[DD(\theta_i) - 2DR(\theta_i) + RR(\theta_i)\right]}{\sum_{\theta_i=0}^{\theta} RR(\theta_i)} \ . \tag{2.36}$$

This equation can be derived fairly easily from the Landy-Szalay estimator for the correlation function and assuming $RR(\theta)$ estimates $N_R(\theta)$. Say:

$$1 + \omega(\theta) = \frac{N(\theta)}{N_R(\theta)} \tag{2.37}$$

and:

$$N_R(<\theta) = \sum_{\theta_i=0}^{\theta} RR(\theta_i), \qquad N(<\theta) = \sum_{\theta_i=0}^{\theta} \frac{N(\theta_i)}{N_R(\theta_i)} N_R(\theta_i) \ . \tag{2.38}$$

Substituting in the definition for $\mathcal{N}(<\theta)$:

$$\mathcal{N}_{LS}(<\theta) = \frac{N(<\theta)}{N_R(<\theta)} = \frac{\sum_{\theta_i=0}^{\theta}\left[1+\omega_{LS}(\theta_i)\right]RR(\theta_i)}{\sum_{\theta_i=0}^{\theta}RR(\theta_i)} \ . \tag{2.39}$$

which finally gets us to Eq. 2.36.

This method, while unable to *prove* homogeneity of the Universe, since observations cover only finite survey volumes and cannot probe the entire Universe, provides statistical consistency checks. The fractal dimension profile from the two-dimensional galaxy surveys offers a way to quantify the approach to homogeneity across different scales and across different redshifts, and to compare the predictions of competing cosmological models. Additionally, it is among the most practical approaches, as it relies primarily on angular position data and therefore does not require additional observational information from telescopes that could be difficult to obtain or is biased. This is why my dissertation will focus on developing methods to test homogeneity using the fractal dimension profile.

3. Methods and improvements

Building on the theoretical framework established in the previous chapter, this chapter focuses on the practical aspects of estimating angular homogeneity scales. I begin by reviewing previous relevant work in the literature, summarizing the main methodologies and their limitations. From this, I then develop a consistent methodology focused on the application to upcoming large-scale galaxy surveys: I identify and analyze the various observational and methodological factors that can influence the measurement of homogeneity scales, present different attempts of my own at approaches to determine D_2, θ_H and its uncertainty, and finally come up with one final method.

3.1. The state-of-the-art in angular homogeneity scale measurements

Until now, studies on the angular homogeneity scale have relied on few available, large-scale galaxy surveys:

- The **SDSS [4]:** A global scientific collaboration to collect galactic and extra-galactic optical spectra to map the three-dimensional structure of the Universe. The data from this collaboration has been organized into a number of component surveys, the most relevant to our purposes being the Baryon Oscillation Spectroscopic Survey (BOSS, SDSS-III); the Sloan Extended Quasar, ELG, and Luminous Red Galaxies (LRG) Survey (SEQUELS, SDSS-III/-IV); and the extended-BOSS (eBOSS, which includes SDSS-IV among others), and the data has been distributed through multiple data releases.
- **Arecibo Legacy Fast ALFA Survey (ALFALFA) [51]:** An extragalactic HI* survey designed to conduct a census of the local HI universe out to redshift $z \sim 0.06$.

Though not exhaustive, these are the surveys that have specifically been used in published work on the *angular* fractal dimension and the *angular* scale of homogeneity.

In [44], Gonçalves et al. selected Luminous Red Galaxies from SDSS DR12 to analyze redshift bins of width $\Delta z = 0.01$ in the interval $0.46 < z < 0.62$. This resulted in redshift shells containing between 20,000 and 30,000 galaxies each, over an area of almost 8800 deg^2.

*HI (Hydrogen line or 21 cm line): one can map HI-emitting galaxies by measuring the signal at the corresponding wavelength.

In [46], Avila et al. investigated the scale of transition to homogeneity using the ALFALFA catalogue considering the redshift interval $0 < z < 0.06$. For this analysis, only 13,144 extragalactic HI sources were selected from a survey area of 3284 deg^2.

In [45], the same group replicated these results for the angular scale of homogeneity using a subsample of blue galaxies from SDSS DR12. The selection was designed to match the ALFALFA catalogue in statistics, type of object and redshift distribution.

More recently, in [47], Shao et al. used the LRG subcomponent from SDSS Data Release 16 (DR16) to obtain measurements of angular homogeneity scales across multiple redshift bins of width $\Delta z = 0.01$ within the range $0.46 < z < 0.74$. The DR16 LRG sample contains 174,816 objects in the redshift interval $0.6 < z < 1.0$, covering both the North Galactic Cap and the South Galactic Cap, for a total sky area of 4242 deg^2. However, due to limited sky coverage, the analysis was restricted to the North Galactic Cap, yielding 107,500 objects over 2566 deg^2. This corresponds to between 18,800 and 33,000 galaxies per redshift bin.

While these works have provided valuable first measurements, they are all based on samples with relatively modest numbers of galaxies and incomplete sky coverage. Looking ahead, the situation will be dramatically improved with next-generation large-scale surveys such as Euclid and the Dark Energy Survey (DES). The Euclid Wide Survey is expected to provide a dataset containing potentially a thousand million of galaxies (judging by the statistics of the Euclid Flagship 2 Mock Galaxy Catalogue [52]) out to a redshift $z \sim 3$, over an area of at least $15{,}000$ deg^2 [53], whereas DES has recently released a dataset with 388 million galaxies over an area of 4946 deg^2 [54]. Even if accurate estimates of the number of objects per redshift bin are not exactly known, it is very clear that these upcoming datasets will make possible much more rigorous statistical analyses and that they will significantly improve our capability to determine θ_H and measure the uncertainty on these estimates.

The number of objects in a redshift bin and corresponding areas for each work mentioned are aggregated and presented in Table 3.1.

The methodologies adopted by different groups to determine the angular homogeneity scale are broadly similar, though often implemented with varying levels of complexity. In general, the analysis relies on measuring either the angular correlation function from pair counts (via Eq. 2.34) and using it to obtain the fractal dimension profile $D_2(\theta)$ (via

Table 3.1: Summary of datasets used in angular homogeneity scale studies, as well as expected scale of recent/future surveys.

Work	**Dataset**	**Redshift bin width**	**Objects per redshift bin**	**Sky area covered**
Gonçalves et al. [44]	SDSS DR12 LRG	$\Delta z = 0.01$	20,000–30,000	8800 deg^2
Avila et al. [46]	ALFALFA	$\Delta z = 0.06$	13,144	3284 deg^2
Avila et al. [45]	SDSS DR12, blue galaxies	$\Delta z = 0.06$	13,144	3284 deg^2
Shao et al. [47]	SDSS DR16 LRG	$\Delta z = 0.01$	18,800–33,000	2566 deg^2
(Estimate)	Euclid FS2	$\Delta z = 0.01$	Order of 10^5-10^6	2000 deg^2
(Estimate)	DES Y3 GOLD	$\Delta z = 0.01$	Order of 10^5-10^6	5000 deg^2

Eq. 2.35 and Eq. 2.21) or directly measuring the fractal dimension curve from pair counts (via Eq. 2.36 and Eq. 2.21). The homogeneity scale θ_H is then defined as the intersection of $D_2(\theta)$ with the chosen homogeneity threshold (the 1%-criterium threshold) (as in Eq. 2.29). Subtle differences arise in the choice of estimator, the treatment of random catalogues, the particular way in which the intersection point is obtained, and the approach to deriving uncertainties, and these differences, while often minor in terms of the final value for θ_H, affect the interpretation of the reported error and, more generally, the robustness of the results.

In the following sections, the methods of bootstrap resampling and jackknife resampling are mentioned. For an explanation of these methods, I refer you to [55] or to the documentation of the Python package `TreeCorr`*.

In Gonçalves et al. [44], the authors computed $D_2(\theta)$ directly using one of the estimators described in Subsec. 2.3.5 for scaled number counts and using the definition of D_2 (Eq. 2.21), while approximating the homogeneity reference curve $D_2^H \approxeq 2$. They explicitly acknowledged that this approximation introduces a small error, though negligible up to angular scales of about 20° - nonetheless, it represents an extra systematic uncertainty if one seeks higher precision measurements. Importantly, they compared results using each of the estimators and confirmed consistency across them. For the random catalogues, they used 20 realizations, apparently matched in size to the data catalogues, though the paper is not fully explicit on this point. Their method for deriving θ_H involved taking the 20 D_2 curves (each associated with one random catalogue), then bootstrapping these curves to generate resampled distributions. Each bootstrap sample was fitted with a polynomial, from which θ_H was estimated. The final value of θ_H was taken as the

*`https://rmjarvis.github.io/TreeCorr/_build/html/cov.html`

mean of the bootstrapped estimates, with the uncertainty given by the standard deviation. The exact implementation is not perfectly clear - in principle, they could also have bootstrapped the fitted curves themselves rather than the D_2 data points, or fitted each of the 20 curves separately and resampled the θ_H values instead, where the former options would be preferred - but the main point is that the uncertainty they report only reflects variance due to random catalogues, not genuine cosmic variance. This was likely intentional, given the relatively poor statistics of the data, but it limits the interpretability of their error bars. The general flow of the method used in this work is represented in Fig. 3.1.

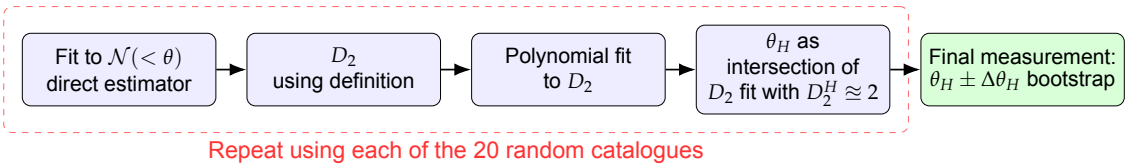


Figure 3.1: Flow diagram of the procedure in Gonçalves et al. [44] to estimate the angular homogeneity scale.

In Avila et al. [46], a similar procedure was adopted: different estimators were tested, with consistent results. I choose to focus on the method using the Landy-Szalay (LS) estimator, given its known minimal variance and bias. As in their later work, they used multiple random catalogues (20 in total) and obtained θ_H values, but yielding the reported mean and standard deviation from a bootstrap resampling procedure of them, whereas later on, in [45] the group abandoned this approach in favor of a simpler procedure.

Specifically, in Avila et al. [45], $\omega(\theta)$ was computed with the LS estimator (Eq. 2.34) using 20 random catalogues, each matched in size to the data. For each resulting correlation function data points, they performed a best-fit (no functional form is specified) and substituted this fit into the analytical expression for scaled number counts (Eq. 2.27). The D_2 values were then calculated numerically, producing a $D_2(\theta)$ curve that was itself fitted with a polynomial chosen by minimizing the root mean squared error (RMSE) under the Akaike Information Criterion (AIC). The intersection of this polynomial fit with the threshold resulted in θ_H. Repeating the process for all 20 randoms provided a distribution of θ_H values, from which the mean and standard deviation were adopted as the measurement and associated error. In essence, this is a more complicated reformulation of the procedure used by Gonçalves et al. [44], with the potential advantage only that a fit to $\omega(\theta)$, as expected by the authors, should reduce the undesirable effect of small-scale oscillations in the distribution of galaxies on the fractal dimension curve and hence on the measurement of the angular homogeneity scale, but with additional intermediate fitting choices

that could propagate systematic differences. The general flow of the method used in this work is represented in Fig. 3.2.

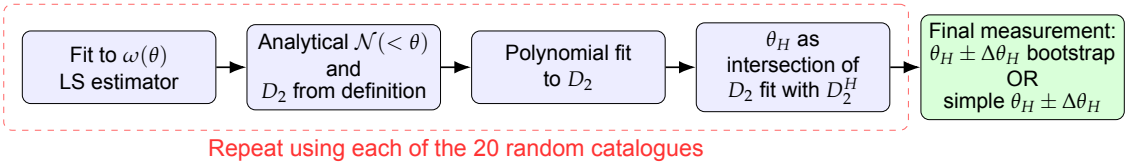


Figure 3.2: Flow diagram of the procedure in Avila et al. [46] and [45] to estimate the angular homogeneity scale.

A somewhat different implementation was presented in Andrade et al. [43]. Here, the LS estimator of the two-point correlation function (Eq. 2.34) was fed directly into the analytical expression for scaled number counts (Eq. 2.35), and $D_2(\theta)$ was then computed numerically (via Eq. 2.21). In constrast to earlier analyses, only a single random catalogue was used, though it was constructed to contain fifty times the number of points in the data - a computationally manageable strategy for current data volumes, but potentially not for the next generation of galaxy surveys, where both data and randoms may contain each hundreds of thousands of objects in a single redshift slice. Uncertainty estimation was more rigorously treated: they used an *internal method** involving bootstrap resampling of the survey footprint, generating 1000 resamples of the original data. Each resampling produced a $D_2(\theta)$ curve and a corresponding θ_H (obtained via interpolation of the D_2 curve around the threshold), resulting in a distribution of 1000 θ_H values. The reported mean and standard deviation of this distribution then served as the final measurement and its error. They confirmed that these bootstrap-based distributions reproduced the expectations from 1000 mock realizations of the original survey data, suggesting the procedure was statistically sound. Still, two concerns remain relevant for future surveys: first, the computational burden of repeating pair count calculations at Euclid-like scales may prove prohibitive (though in practice it may be possible to work around this); and second, the bootstrap technique in its simplest implementation, as described in [55], by its nature, discards some correlation information between pairs of galaxies, which I cannot confidently say would amount to a negligible effect in uncertainty estimation when better statistics are available (this will be treated in some more detail further on). The general flow of the method used in this work is represented in Fig. 3.3.

*An *internal* method of estimating uncertainty relies on the data only, e.g., by survey subsampling and bootstrapping. An *external* method refers to using mock catalogues from theoretical models or using theoretical calculations directly to determine uncertainties. The latter always requires assuming a cosmological model.

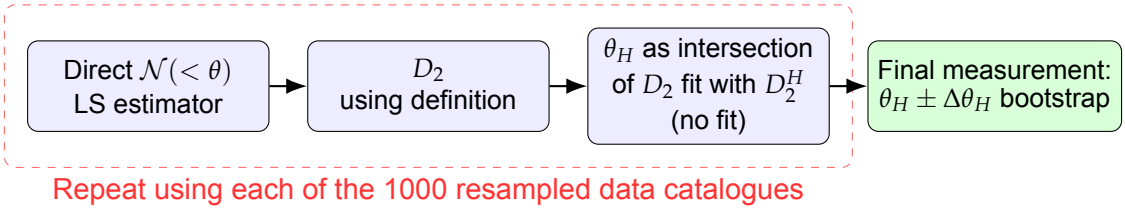


Figure 3.3: Flow diagram of the procedure in Andrade et al. [43] to estimate the angular homogeneity scale.

The most recent work, Shao et al. [47], follows essentially the same conceptual framework. Although the precise details are not presented with much clarity, the mention of 1000 bootstrap realizations for estimating θ_H suggests that, as in earlier works, D_2 curves were generated for different random catalogues, bootstrap-resampled, and then combined to estimate a final θ_H. At its core, their procedure does not, if this is indeed the case, differ significantly from previous analyses.

Taken together, these studies demonstrate that while the definition of θ_H is, in principle, consistent across the literature, the implementations vary in important ways: some rely on bootstrap resampling of random catalogues, others of the data itself; some adopt direct estimators of D_2, while others introduce intermediate fits or approximations. Critically, most reported uncertainties do not reflect true cosmic variance or even in-sample variance, but rather the technical variance associated with random catalogues. Only one work to date has made a systematic attempt to estimate observational uncertainty from the data distribution itself. On the other hand, accurate modelling of fractal dimension profiles can allow us to distinguish between cosmological models. Not only that, but it has been shown [47] that θ_H measurements, when combined with Baryon Acoustic Oscillations (BAO) and Type Ia supernova (SN) data, can help constrain the matter density parameter Ω_{m0} and the Hubble constant H_0. Therefore, a central component of the present work will be to propose and adopt a single consistent methodology that aims to produce uncertainties with a clearer relationship to cosmic variance, and to test its validity, while taking into account performance concerns of applying the methodology to a very large dataset. This will be particularly important as upcoming surveys such as Euclid deliver vastly improved statistics, where precision will be limited not by the number of objects but by the robustness of our methodological choices.

3.2. Some considerations on estimators, binning, randoms, and noise

3.2.1 Choice of estimator

I focus on the Landy-Szalay (LS) estimator for the two-point correlation function, which has been consistently shown to yield the least bias and variance under conditions of vanishing correlation at large scales and simple survey geometry [56]. It is also the most widely adopted estimator in cosmological analyses, with a proven record.

While LS involves computing all three pair counts (DD, DR, RR), making it computationally heavier than some alternatives, the performance cost is justified given its statistical advantages. Note also that more recent alternatives exist - see e.g. [57], which provide promising improvements in certain contexts - but in this work, I adopt the LS estimator as the baseline throughout.

3.2.2 Angular separation binning

An important practical choice is the binning of angular separations. Whichever estimator is used, the angular correlation function estimate obtained is binned in angular separation: and the bins at which we choose to estimate $\omega(\theta)$, i.e. the angular separations at which we calculate pair counts, can bias $\omega(\theta)$. Both linear-space and log-space binning are possible, but log-space binning is clearly advantageous:

- $\omega(\theta)$ decays approximately exponentially, such that variation is fastest at small θ. Finer bins in log-space reflect this behavior naturally.
- At larger scales, log-binning suppresses small-scale oscillations by grouping low-count pairs into wider bins, improving stability without losing overall shape.
- The same reasoning applies to $D_2(\theta)$, which is polynomial/exponential-like.

This binning choice reduces noise in the original function estimates. Fig. 3.4 shows the impact of the different choices on an example correlation function $\omega(\theta)$ and on an example $D_2(\theta)$ profile.

3.2.3 Number of bins

The number of bins is another relevant parameter. Too few bins fail to capture rapid variations at small angles, while too many bins produce noisy functions due to undersampling

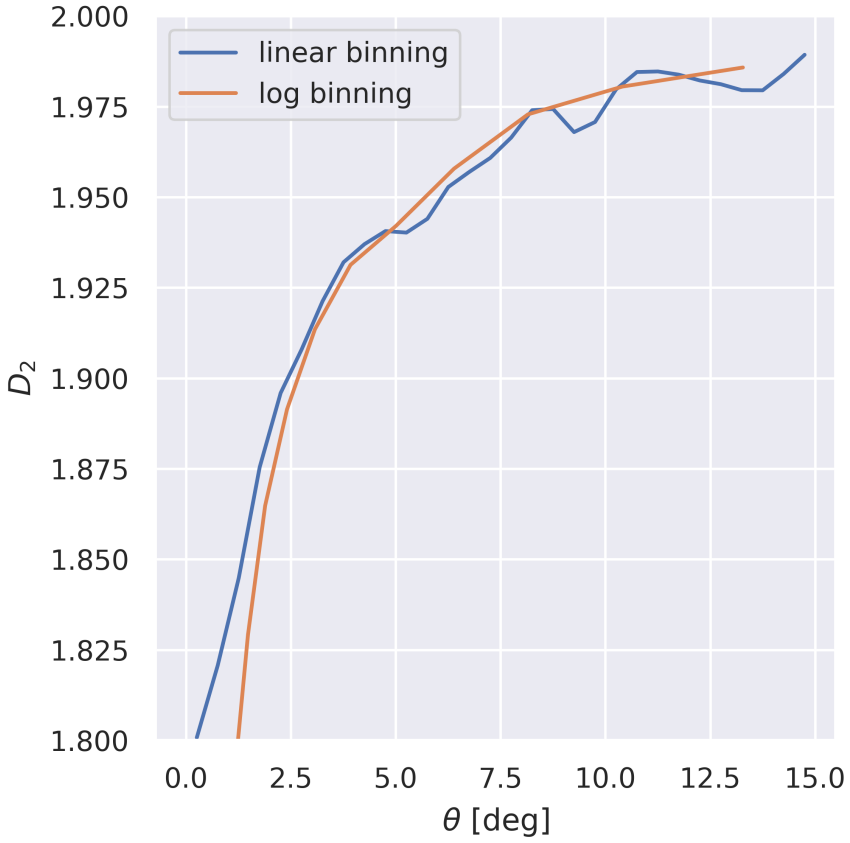


Figure 3.4: Example D_2 curve obtained using MICE-Grand Challenge Galaxy Catalogue (MICECAT) – see section 5.1 – for the same region in the sky but where pair counts were put, in one case, into log-spaced bins, and in another case, into linear-spaced bins.

of the pair counts (see Fig. 3.5 for an example). Although I did not perform a systematic optimization, these two factors should be taken into consideration when selecting the number of bins to use.

3.2.4 Random catalogues

Another practical decision concerns the use of random catalogues. Previous works used either many randoms (e.g. 20) or a single very large one. For upcoming large surveys, the variance introduced by random catalogues will be negligible compared to sample variance in the data. As I demonstrate in Fig.3.6, the variance of $\omega(\theta)$ and $D_2(\theta)$ across multiple random catalogues is very small compared to what was obtained in the works described above, e.g. in [44].

Thus, either approach can be used:

- a single sufficiently large random, or

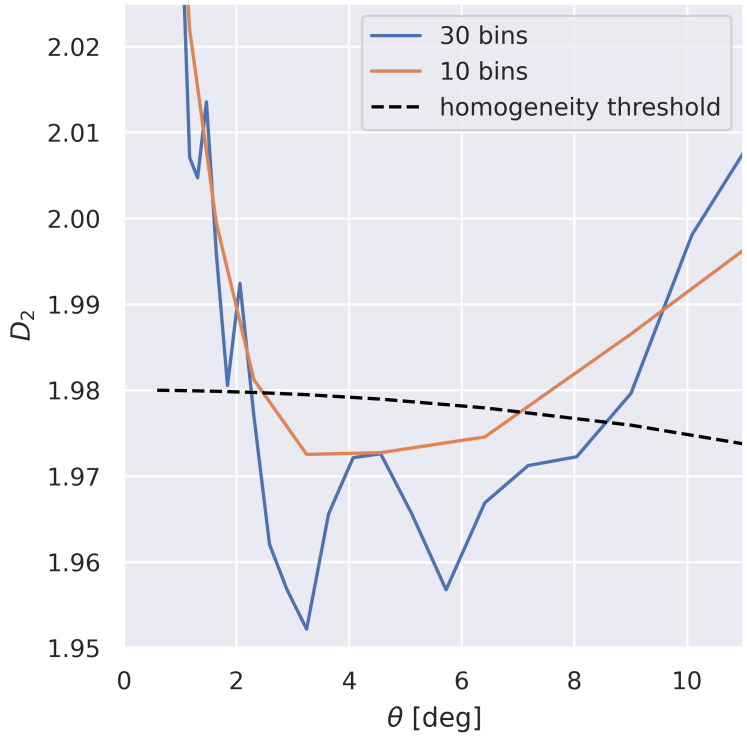


Figure 3.5: Same as Figure 3.4 but comparing the impact of the number of bins on D_2 profile.

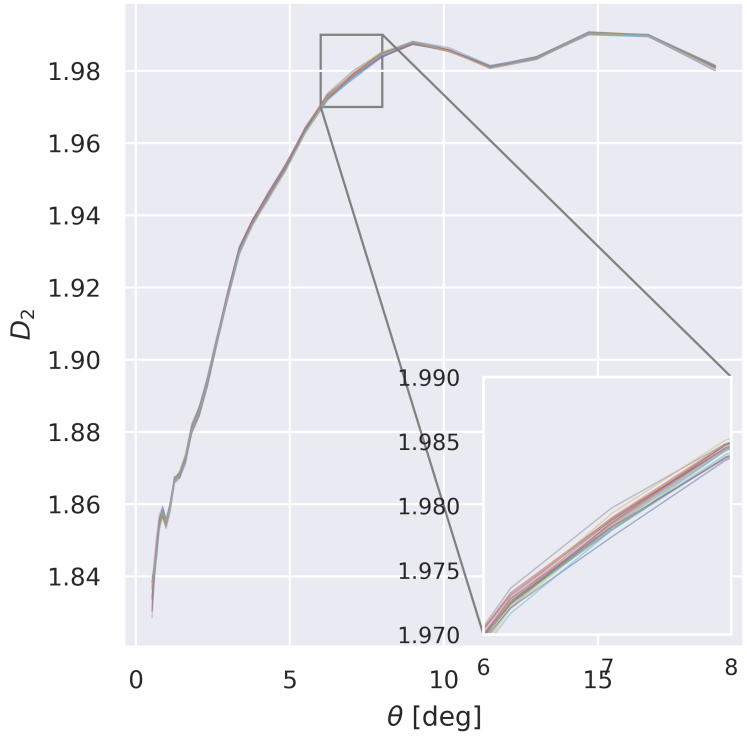


Figure 3.6: D_2 curves obtained from the same galaxy mock catalogue as in Figure 3.4, using one of 20 different random catalogues (generated according to the algorithm in Subsec. 4.2.1) with $N \sim 200,000$ objects. Each line of a different color corresponds to a different curve.

- several smaller randoms whose curves are averaged.

The main consideration is computational performance, especially as data size increases.

3.3. A first proposed method: the covariance matrix-based approach

3.3.1 Estimating uncertainty: covariance matrices

As an alternative to the approach of *1000 bootstrap resamplings* of the survey area, I attempted a covariance matrix-based approach, estimated directly from survey data by subdivision into spatial patches. It suffers from the same limitations, in that:

- It captures variance in the *measured* data, but not necessarily global cosmic variance, which cannot be probed without the full volume of the Universe.
- For small areas, this variance estimate degrades, but within wide surveys like Euclid, I expect patching is sufficient (tests of this are presented further on).

Additionally, as opposed to the literature method, I am assuming the variance in the data can be fully described by a mean and covariance matrix, i.e. variance follows a Gaussian distribution. However, this approach may make it possible to develop an alternative, more efficient, potentially equally reliable method to estimate the final uncertainty in measurements of θ_H, my end goal.

I follow the weighting schemes proposed by [58] for the jackknife and bootstrap resampling methods, which improve covariance estimation by down-weighting but not discarding information from pairs of galaxies crossing patch boundaries into unselected patches (patches that were deleted during one particular iteration of the method). These schemes were shown to reproduce true covariance matrices from mock data within 10%, the minimum error possible with the sample size used in the study, a large improvement on traditional jackknife/bootstrap.

Through my analysis of these methods with different number of patches, I show that bootstrap and jackknife produce similar results if the number of bootstrap resamplings is large enough for bootstrap to converge.

Additionally, both methods expect patches to be statistically independent, so the minimum or typical angular separation in a patch must be greater than the correlation length (as can be inferred from $\omega(\theta)$).

In order to determine the ideal number of patches for a given survey and a given survey area, one can compare the structure of the covariance matrix for different number of patches for a given survey area and determine the point at which we have converged sufficiently to a stable approximation of the underlying covariance matrix.

Tests of these methods will be done and presented in Chapter 5.

3.3.2 Two different methods for $D_2(\theta)$

I distinguish between two methods for obtaining the fractal dimension profile, both previously considered in the literature:

- **Direct method:**
  - Start from the scaled number counts ($\mathcal{N}$) LS estimator (Eq. 2.36),
  - Compute $D_2(\theta)$ directly from its definition (Eq. 2.21),
  - Fit $D_2(\theta)$ to suppress small-scale oscillations and determine the homogeneity threshold from the interpolation.
- **Indirect method:**
  - Start from $\omega(\theta)$ with the LS estimator (Eq. 2.34),
  - Fit $\omega(\theta)$ to suppress small-scale oscillations,
  - Insert it into the analytical expression for scaled number counts $\mathcal{N}$ (Eq. 2.27) - integration can be done numerically or, when possible, analytically using symbolic computation,
  - Use the definition of D_2 (Eq. 2.21) to derive $D_2(\theta)$, from which, once again, you can obtain θ_H.

The indirect method was motivated by the possibility of smoothing out small-scale oscillations in $\omega(\theta)$ in a more robust manner, as suggested in [46], before these oscillations in the correlation function are propagated to D_2, where they will be amplified. However, due to the better statistics of the large-scale galaxy surveys I intend to work with, it is possible that this alternative method does not improve the result.

3.3.3 Symbolic Regression (SR) as a fitting strategy

Previous studies have considered only fixed functional-form fitting methods, such as with polynomials. I initially experimented with such methods, testing forms like simple polynomials of degree n, exponentials of the type:

$$f(\theta) = A \exp\left(-B\theta^C\right) + D \tag{3.1}$$

and power-law forms such as:

$$f(\theta) = A\theta^B + C \tag{3.2}$$

where θ denotes the angular separation. These choices capture some of the broad qualitative behavior of $\omega(\theta)$ and $D_2(\theta)$, and they worked reasonably well in certain cases. However, they did not consistently provide good fits across all redshift bins: depending on the degree of clustering and the noise level in the data, the fits often failed to reproduce the proper shape of the function. Binning choice may have introduced an additional complication. My early implementations used real-space binning of angular separations, which likely contributed to unstable fits. Indeed, even when considering log-space binning, I found that these fixed functional forms would not reliably capture the data trends, with one significant issue in polynomial fits (the most flexible of these choices): high-degree polynomial fits are prone to developing spurious oscillatory behavior between data points, which leads to noise in the integral of Eq. 2.35.

Nevertheless, using these fixed forms, I was able to replicate to some degree the θ_H estimates reported in [44], applying their same data. Although my results did not match perfectly - likely because my interpretation of their algorithm deviated in minor but important details - they were approximately consistent with their published values. This exercise confirmed that fixed functional forms can approximate the homogeneity scale in the presence of noisy data, but also highlighted the intrinsic difficulty of robustly fitting such forms across multiple redshift bins with heterogeneous clustering properties, as well as the ambiguity of choosing any one particular functional form as a basis for the $\omega(\theta)$ and $D_2(\theta)$ profiles.

A key improvement I propose is replacing fixed functional form-fitting with SR. Symbolic regression is a non-supervised machine learning fitting method that searches over mathematical expressions, rather than sticking with parametrized, predefined forms, to identify the best fit to data. Specifically, given a user-defined list of operators and functions (the building blocks of the mathematical expressions), SR evolves populations of candidate

expressions using natural-selection-inspired rules, optimizing both fit accuracy and expression simplicity. Some of the advantages over traditional fitting methods are:

1. Does not require preset functional forms;
2. Produces interpretable closed-form expressions with no free parameters at the end;
3. Can guarantee fits of uniform quality across redshift bins by using predefined quality thresholds;
4. Can provide unbiased, data-driven analytical formulas that can be compared to theoretical expectations.

The one disadvantage of this method is that it is, in general, slower, sometimes by many orders of magnitude. However, I expected that the advantages will outweigh this disadvantage.

3.3.4 Obtaining θ_H

- **Direct method:**
  - SR fitting is applied directly to $D_2(\theta)$ LS data points;
  - Building blocks for SR: limited to the basic operators $+$, $-$, $\times$, $\div$, exponentiation, and to the exponential function e. This is flexible enough to describe the shape of the D_2 curves;
  - Maximum complexity (or number of terms - constants and variables) set to 40, but any reasonably high number will do - there is no valid reason to limit complexity and overfitting is not problematic (if artificial oscillations are minimized via binning choices, at least);
  - Fitness/Loss function defined as weighted mean squared error (WMSE), with weights equal to the inverse covariance matrix obtained as in sub-section 3.3.1;
  - Stopping criterion: WMSE ≤ 1, i.e. residuals at or below D_2 measurement noise.* The justification for this criterion is that, given two fit functions with residuals on average under the D_2 measurement noise, then the choice of fitting function is arbitrary because the noise level does not allow you to distinguish between the two;

*This is the generalization of the result $\sum_i^N (y_i - f(x_i))^2 \leq N\sigma^2$ to non-constant and non-diagonal weights.

  - θ_H obtained as the intersection of the SR-fit $D_2(\theta)$ with the homogeneity threshold curve.

- **Indirect method:**

  - SR is used to fit $\omega(\theta)$. The same building blocks, maximum complexity, fitness function and stopping criterion are applied as for the direct method;
  - Using symbolic computation (or, for a more complicated implementation, by numerical calculation), this fit expression is inserted into the scaled number counts $\mathcal{N}$ integral (Eq. 2.35), and the full $D_2(\theta)$ expression is derived analytically;
  - θ_H is obtained again by root-finding against the threshold value.

Here, symbolic computation refers to the algorithmic manipulation of mathematical expressions (e.g., differentiation, integration) akin to how *Wolfram Alpha*, *Mathematica* or *Sage* can manipulate expressions in an analytic, automated way. With SR-fit symbolic expressions, this allows analytic propagation through the integral definitions to produce equally analytic D_2 expressions.

3.3.5 Estimating θ_H uncertainty

Multiple routes exist to estimate uncertainty on θ_H.

- **Ideal case:** Resample $\omega(\theta)$ or $D_2(\theta)$ data points from the resamples of the survey data (skipping entirely the step of actually generating a covariance matrix, which is unnecessary in this case), fit each resample with SR and follow the methods as explained in subsection 3.3.4, and build a distribution of θ_H. The mean and dispersion give θ_H and its uncertainty. This is similar to the *1000 bootstrap resampling* method in the literature, except it uses a different weighting scheme in the jackknife/bootstrap resampling procedure.

  - Advantage: statistically robust, fully data-driven;
  - Disadvantage: very computationally heavy (many SR fits).

- **Approximate case:** Perform a single SR fit, then parametrize constants as variables/parameters (see Diagram 3.7 for an example of what this means). Estimate parameter covariance (via methods described in 3.3.6) and propagate to θ_H.

- Advantage: (potentially) much faster?;
- Disadvantage: less rigorous; assumes fit profile remains valid under resampling of the data points.

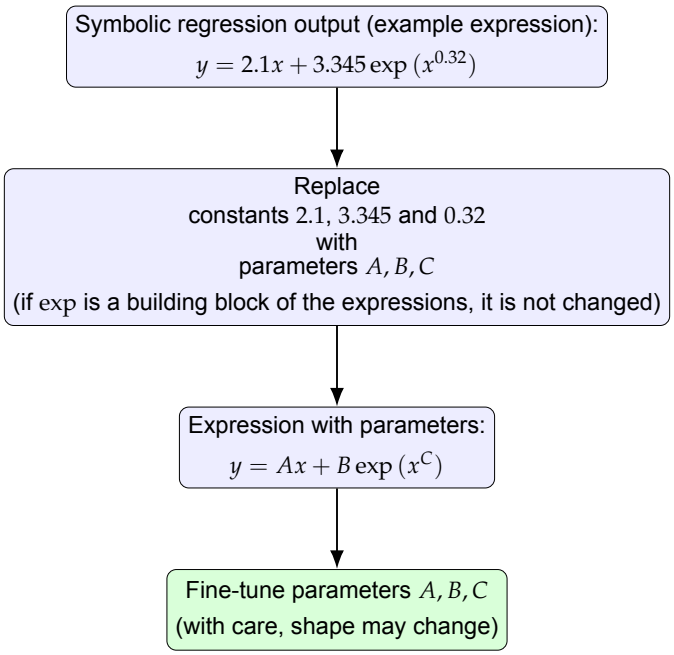


Figure 3.7: Procedure to transform SR expressions into parameterized forms for fine-tuning.

I study in detail the approximate case, analysing its validity and robustness, to test whether the approximation performs well enough in practice.

Finally, I provide schematic diagrams (Fig. 3.8 and Fig. 3.9) illustrating the full flow for obtaining θ_H and uncertainty for each method, direct and indirect.

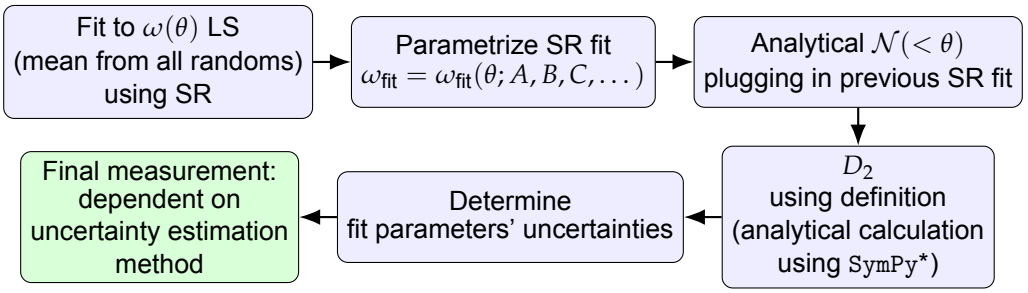


Figure 3.8: Flow diagram of the SR approach using the **indirect** method to estimate the angular homogeneity scale and uncertainty.

*https://www.sympy.org/

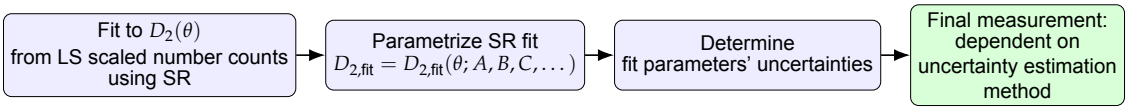


Figure 3.9: Flow diagram of the SR approach using the **direct** method to estimate the angular homogeneity scale and uncertainty.

3.3.6 Parameter covariance estimation methods

In this section I describe two approaches to estimating parameter covariance matrices for fitted functions of the form

$$f(x; A, B, C, ...)$$

where $A, B, C, ...$ denote model parameters obtained from fitting, based on the covariance information C_d from the data.

The first method is a local approximation based on linearization, implemented also in Python, in the method `curve_fit` of the package `scipy`, and detailed in [59]. The second is a Bayesian approach based on nested sampling.

Linear approximation method

The first method relies on the Gauss-Newton approximation within non-linear least-squares fitting.

Suppose the data vector is y and the model prediction is $y_d = f(x; \boldsymbol{\beta})$ for parameters $\boldsymbol{\beta} = (A, B, C, ...)$. The residuals are:

$$r = y - y_d \tag{3.3}$$

and we can write the mean squared errors as

$$\chi^2 = r^T r \tag{3.4}$$

To second order, the expansion of χ^2 around the best-fit parameter set $\hat{\boldsymbol{\beta}}$ (where we assume that $\Delta_{\boldsymbol{\beta}} \chi^2 \Big|_{\hat{\boldsymbol{\beta}}} \approx 0$) becomes

$$\chi^2(\boldsymbol{\beta}) = \chi^2(\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \, (J^T J) \, (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathcal{O}(\boldsymbol{\beta}^3) \tag{3.5}$$

where J is the Jacobian of the residuals, $J = \frac{\partial r}{\partial \boldsymbol{\beta}}$. With this approximation, the Gaussian likelihood can then be written as

$$\mathcal{L}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \, (J^T J) \, (\hat{\boldsymbol{\beta}}) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right) \tag{3.6}$$

from whence we identify $C^{-1} = J^T J$, and therefore, the covariance matrix of the parameters is simply

$$\text{Cov}(\boldsymbol{\beta}) \approx (J^T J)^{-1}. \tag{3.7}$$

This approximation is convenient due to its speed, but it relies critically on the assumption that the model is well-approximated as linear in the neighborhood of the optimal parameters, and that residuals are Gaussian-distributed and small.

Additionally, including uncertainties in the data points is straightforward. if the data covariance is C_d, we can ”whiten” the residuals to incorporate these correlations without changing the algorithm described above. Writing $C_d = LL^T$ as a Cholesky decomposition,

$$\chi^2 = r^T C_d^{-1} r = ||L^{-1} r||^2 \tag{3.8}$$

and we define the whitened residuals as

$$r' = L^{-1} r \ . \tag{3.9}$$

It is then just a matter of replacing r in the algorithm with the whitened residuals r'. Note that χ^2 remains structurally identical:

$$\chi^2 = r'^T r' \ . \tag{3.10}$$

The covariance matrix of the parameters $\text{Cov}(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = (A, B, C, ...)$ for a fit function $f(\theta; \boldsymbol{\beta})$ can be used to estimate an uncertainty on θ_H by applying the linear error propagation formula and the implicit function theorem:

$$\sigma^2(\theta_H) = \left(\frac{\partial \theta_H}{\partial A}\sigma_A\right)^2 + \left(\frac{\partial \theta_H}{\partial B}\sigma_B\right)^2 + 2\frac{\partial \theta_H}{\partial A}\sigma_A \frac{\partial \theta_H}{\partial B}\sigma_B + ... \tag{3.11}$$

where the partial derivatives of θ_H are calculated as

$$\frac{\partial \theta_H}{\partial A} = \left.\frac{\partial f / \partial A}{\partial f / \partial \theta}\right|_{\theta = \theta_H}$$

and so on.

Nested sampling method

The second approach is Bayesian and uses nested sampling to generate posterior distributions for the parameters. Unlike the linear method, this does not assume local linearity, but instead directly explores the structure of the likelihood across parameter space.

The algorithm begins with N points (called ”live points”) drawn from the prior - a user-defined probability distribution for the points we should sample - these points correspond to pairs of parameters of the model fitting function. At each step, the point with the lowest likelihood is replaced by drawing a new point that satisfies the constraint of exceeding this minimum likelihood. Repeating this process iteratively both provides an estimate of the Bayesian evidence and samples the posterior as a by-product. Termination occurs when certain criteria are met: either a number of iterations has been reached or some other convergence criterium is satisfied.

Each sample point is associated with an importance weight proportional to the posterior probability of point $\boldsymbol{\beta}$ and the total volume in parameter space encompassed by the points with the same posterior probability:

$$w_i \propto P(\boldsymbol{\beta}|\text{data})dV(\boldsymbol{\beta})$$

and which essentially describe how likely it is for this set of parameters to properly model my data.

For my implementation, I used the `dynesty` package with a dynamic nested sampling mode, which adaptively generates more live points in undersampled regions and can provide a better final number of samples without wasting additional computational resources. My likelihood is the Gaussian log-likelihood:

$$\ln \mathcal{L} = -\frac{1}{2}\left[r^T C^{-1} r + N \ln(2\pi) + \ln\det C\right] \tag{3.12}$$

where r are residuals, C the data covariance, and N the number of data points fitted.

Priors were taken as normal distributions centered on the optimal parameter values $\hat{\boldsymbol{\beta}}$ with $\sigma = 1$, effectively unbounded. Sampling steps used a random walk method to minimize the risk of infinite loops, a decision that was made after some tests. Convergence was validated by inspection of trace plots showing parameter values as a function of posterior mass.

A posterior covariance matrix can be estimated as the weighted variance of the nested samples. Alternatively, one can bypass covariance entirely by directly propagating each sampled parameter set through the symbolic regression fit to obtain θ_H, thus constructing a distribution of θ_H values weighted by sample importance. The mean and standard deviation of this distribution then provide the final measurement and uncertainty. I adopted the latter method.

Comparison of the two methods

Nested sampling is generally slower than the linear approximation, though not prohibitively so. It yields more reliable posterior distributions, especially in non-linear or high-dimensional parameter spaces, and it naturally respects parameter constraints and boundaries. Comparison of the posterior parameter distributions with the Gaussian expectation or comparison of the covariance matrices of this method with the one obtained from the linear approximation method can shed some light on the reliability of the latter method. Additionally, unlike MCMC methods, nested sampling is guaranteed to converge to a maximum of likelihood, even if not a *global* one.

It is important to note, however, that literature ([59]) finds the linear approximation typically underestimates confidence regions relative to sampling-based methods. Not only that, but the linear approximation cannot take into account parameter value constraints, potentially describing covariance structure that allows parameters that are not valid for the model function at all.

Regardless, I implemented and tested both methods to assess the reliability of the linear approximation method in my particular scenario.

As noted in the same work, there are few case studies of these methods in high-dimensional parameter spaces. Since my symbolic regression fits can involve many parameters, this work provides a useful, albeit small, empirical test of both methods in such contexts.

3.4. A second, final proposal: the survey resampling approach

The second and final method that I propose builds on the ideas discussed previously but simplifies the implementation, while simultaneously achieving higher accuracy and robustness. It is inspired by the strategy of the *1000 bootstrap resamplings* mentioned previously in the literature, but applied here with two modifications:

1. the use of the jackknife method, which I have already identified as the preferred resampling procedure;
2. the adoption of the improved weighting scheme discussed in the first proposal.

In addition, for each resample of the survey, I obtain measurements of $\omega(\theta)$ with the Landy-Szalay (LS) estimator, as well as measurements of $\mathcal{N}(<\theta)$ and $D_2(\theta)$ using the LS estimator, for each different random catalogue. In this way, the resampling strategy already takes into account variance due to randoms.

In practice, resampling the original survey and recomputing pair counts would be slow - hence why this strategy was not initially considered. However, the method of weighting pairs based on the patch they are in, and whether the patch is *selected* or *unselected* in a resample*, as `TreeCorr` does, allows producing the $\omega(\theta)$ and $D_2(\theta)$ functions for the resampled surveys while not having to actually recompute all pair counts for the resampled survey.

This procedure still distinguishes between two possible routes to obtain $D_2(\theta)$ and, ultimately, θ_H:

- **Direct method:**
  - Start from the scaled number counts LS estimator $\mathcal{N}(<\theta)$ (Eq. 2.36).
  - Compute $D_2(\theta)$ directly from its definition (Eq. 2.21).
  - Fit the resulting profile with a cubic spline with or without smoothing (whether smoothing is useful is to be tested).
  - From this spline representation, interpolate to determine the first crossing of $D_2(\theta)$ with the 1%-criterium homogeneity threshold curve $0.99 \times D_2^H(\theta)$, requiring that the slope at the crossing is positive/upward. This defines θ_H.
  - Repeat this procedure across all resamples, obtaining a distribution of θ_H values. The median of this distribution is adopted as the final measurement, with the 16th and 84th percentiles defining the 1σ confidence interval (this definition is up for debate).
- **Indirect method:**
  - Start from $\omega(\theta)$ obtained with the LS estimator (Eq. 2.34), fitting with a cubic spline with or without smoothing, exactly as done for the direct method.
  - Compute $D_2(\theta)$ by numerically integrating the analytical relations between $\omega(\theta)$, scaled number counts, and $D_2(\theta)$ (Eq. 2.35, Eq. 2.21).
  - Proceed as in the direct method: fit once more the resulting $D_2(\theta)$ with a cubic spline to determine the crossing point with the 1%-criterium homogeneity threshold curve $0.99 \times D_2^H(\theta)$ and compile the distribution of θ_H values across all resamples.

*A *selected* patch is one in which it has not been removed from a resample. Remember that in jackknife, each resample is obtained by removing one different patch from the original survey.

In practice, the difference between the two approaches lies in their baseline assumptions. The direct method estimates $N_R(\theta)$ (number of objects in a random distribution separated by an angular distance θ) empirically from the $RR(\theta)$ counts, so both data and random catalogues are consistently handled through pair counts. The indirect method, by contrast, replaces this with the theoretical expectation (Eq. 2.19).

In both methods, one may have noticed that implicitly or explicitly the fractal dimension profile involves an integration from an initial angular separation $\theta = 0$. Not doing this (e.g. by counting pairs starting at some initial $\theta > 0$) introduces a potential source of error. Ideally, to compare the two methods fairly, both should adopt the same lower limit of integration, though one may prove more sensitive to this choice than the other. This remains a detail requiring further investigation.

Unlike the fixed functional forms or symbolic regression fits explored in the first proposed method, spline fitting is both highly flexible and computationally efficient. It eliminates the need to assume any particular fitting function, and because it is applied consistently to all resampled data, small-scale oscillations should no longer pose a problem: these oscillations should cancel out when one considers the full ensemble of jackknife-resampled curves, ensuring that the final θ_H distribution is not unduly influenced by them.

This alternative proposed method thus corresponds exactly to the *ideal case* outlined in the first proposal: uncertainty in θ_H is derived directly from the distribution of values across resamples, without additional assumptions about Gaussianity or covariance structure. This survey resampling approach therefore provides the simplest, fastest, and potentially the most robust estimation of θ_H and its associated uncertainty among all the methods considered. Its only limitation is that it does not produce closed-form analytic expressions for $D_2(\theta)$, as symbolic regression does. However, given its speed, interpretability, and adherence to the data-driven philosophy, it is the best candidate for the final and preferred approach in this work.

This approach is illustrated as a flow diagram in Fig. 3.10.

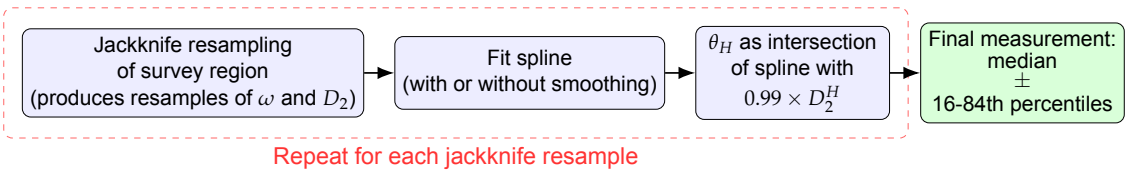


Figure 3.10: Flow diagram of the resampling approach using the **direct** or **indirect** methods to estimate the angular homogeneity scale and uncertainty.

3.5. Additional consideration: BAO scales in the angular correlation function

In the very early Universe, shortly after the Big Bang, space was filled with a hot, dense plasma of charged particles (electrons and protons) strongly coupled to photons. This photon-baryon fluid oscillated with characteristic frequencies as density fluctuations evolved under the competing effects of radiation pressure and gravitation. As the Universe expanded and cooled, the photons decoupled from the plasma around the time we call *recombination*, leaving a permanent imprint of these oscillations in the clustering of matter.

This imprint, known as BAO, manifests itself in the matter power spectrum* as a series of peaks. In the three-dimensional correlation function $\xi(r)$, these peaks correspond to a single broad peak at the scale of the sound horizon at recombination (~ 100 Mpc/h), the typical comoving scale of those early density fluctuations.

While this feature is clearly visible in $\xi(r)$, its imprint in the angular correlation function $\omega(\theta)$ is less straightforward. Because $\omega(\theta)$ describes a *projected* distribution of objects, integrated over a redshift distribution, projection effects can both *dillute* and *shift* the BAO peak. The degree of such distortion depends on two key factors:

- Redshift z;
- Redshift bin width Δz.

A systematic investigation of these effects was carried out in [60], where the authors analyzed BAO structures in $\omega(\theta)$ across different cosmological models. They found that the position of the BAO peak observed in $\omega(\theta)$ shifts more significantly at low redshift if the redshift bin width is large (see figure 3 of the work cited). However, for sufficiently thin bins ($\Delta z \sim 0.01$), the observed peak closely matches the theoretical expectations in position, regardless of redshift. On the other hand, increasing the bin width can strongly dilute the BAO feature in the angular correlation function, until it becomes practically undetectable (see figure 1 of the same work).

In my work I validate that the BAO peaks do not coincide with the measured θ_H, as that would make such measurements unreliable, but I have not performed a thorough analysis of BAO features or developed algorithms to isolate or correct for their effects. Nevertheless, this is an important consideration for future applications of my methodology

*The matter power spectrum is the Fourier transform of the correlation function. It is a decomposition of the correlated modes of oscillation in the matter density field in terms of their frequencies.

to real survey data. A natural avenue for further work is to investigate whether the indirect method of determining θ_H (based on fitting $\omega(\theta)$ first) may be especially well-suited to mitigate or model out BAO-related distortions (and how that would be done) or whether it would be preferable (and more importantly, viable) to fine-tune survey bins in terms of redshift bin width or density of objects to shift the positions of the θ_H value and BAO peak away from each other. Alternatively, one may also consider whether using different tracers for pair counts could keep us away from the BAO scales.

4. Implementation details

In this chapter I discuss several additional aspects of the analysis that, while not the central methodological improvements proposed in the previous chapter, are nonetheless crucial for obtaining accurate and reliable results. Specifically, I focus on (i) pair counting and correlation function codes, (ii) generation of random catalogues, and (iii) regularization of covariance matrices. Each of these falls under what might be called the ”back end” of the analysis: choices that determine performance, stability and the robustness of the methods and therefore of the derived angular homogeneity scale θ_H.

4.1. Pair counting codes

In the introduction I already defined the Landy-Szalay estimator and the corresponding quantities DD, DR, and RR, including their interpretation as pair counts of objects in the data (D) or randoms (R).

The simplest algorithm for computing pair counts can be written in a form equivalent to the pseudocode in Lst. 4.1. This direct approach simply loops through all pairs in the catalogue(s), testing if their separations fall within a given bin. While conceptually straightforward, it scales as $\mathcal{O}(N^2)$ for N objects, rendering it unusable for modern surveys with millions of galaxies.

```
for(int i=0;i<N1;i++){
    for(int j=0;j<N2;j++){
        double dist = distance_metric(i, j); // calculate distance between objects i and j
        if (dist < min_dist || dist >= maxdist){
            continue; // consider only pairs separated by distances in [min_dist, max_dist[
        }

        int ibin = dist_to_bin_index(dist); // pair counts are binned according to distance
        numpairs[ibin]++;
        weight[ibin] += weight_func(i, j); // pairs can have weight information added
    }
}
```

Listing 4.1: Pseudocode for the simplest algorithm to compute pair counts. Source: Sinha, Manodeep, and Lehman Garrison. ”CORRFUNC: Blazing Fast Correlation Functions with AVX512FSIMD Intrinsics.” Workshop on Software Challenges to Exascale Computing. Springer, Singapore, 2018.

Therefore, one must rely on optimized pair counting libraries rather than reinventing the wheel. I tested the performance of the three major publicly available codes currently in use: `Corrfunc`*, `TreeCorr`†, and `GUNDAM`‡. To do so, I constructed mock surveys of galaxy clusters containing various numbers of objects (up to about 1.6×10^4) and paired it with one random catalogue of the same size. All codes were run in multithreaded mode (12 threads), and I carefully specified consistent configurations for each package. The precise Python calls and parameter settings used for each library are listed in 4.2.

```
# Corrfunc
DD_counts = corrfunc.DDtheta_mocks(autocorr=1,
                                   nthreads,
                                   theta_bins,
                                   data_RA,
                                   data_DEC,
                                   output_thetaavg=False,
                                   verbose=False,
                                   copy_particles=False,
                                   ra_refine_factor=5,
                                   dec_refine_factor=5)

DR_counts = corrfunc.DDtheta_mocks(autocorr=0,
                                   nthreads,
                                   theta_bins,
                                   data_RA,
                                   data_DEC,
                                   RA2=rand_RA,
                                   DEC2=rand_DEC,
                                   output_thetaavg=False,
                                   verbose=False,
                                   copy_particles=False,
                                   ra_refine_factor=5,
                                   dec_refine_factor=5)

RR_counts = corrfunc.DDtheta_mocks(autocorr=1,
                                   nthreads,
                                   theta_bins,
                                   rand_RA,
                                   rand_DEC,
                                   output_thetaavg=False,
                                   verbose=False,
                                   copy_particles=False,
                                   ra_refine_factor=5,
                                   dec_refine_factor=5)
```

*https://corrfunc.readthedocs.io/
†https://rmjarvis.github.io/TreeCorr
‡https://gundam.readthedocs.io/

```
cf = corrfunc.convert_3d_counts_to_cf(ND1=len(data_RA),
                                      ND2=len(data_RA),
                                      NR1=len(rand_RA),
                                      NR2=len(rand_RA),
                                      D1D2=DD_counts,
                                      D1R2=DR_counts,
                                      D2R1=DR_counts,
                                      R1R2=RR_counts)

# TreeCorr
data_cat = treecorr.Catalog(ra=data_RA, dec=data_DEC,
                            ra_units='deg', dec_units='deg',
                            npatch=1)

rand_cat = treecorr.Catalog(ra=rand_RA, dec=rand_DEC,
                            ra_units='deg', dec_units='deg',
                            patch_centers=data_cat.patch_centers)

config = {'nbins': len(theta_bins) - 1,
          'min_sep': theta_bins[0],
          'max_sep': theta_bins[-1],
          'num_threads': nthreads,
          'bin_slop': 0.,
          'sep_units': 'deg',
          'bin_type': 'Linear',
          'metric': chosen_metric} # chosen_metric = 'Euclidean' or 'Arc'

nn = treecorr.NNCorrelation(config=config)
dr = treecorr.NNCorrelation(config=config)
rr = treecorr.NNCorrelation(config=config)

nn.process(data_cat)
dr.process(data_cat, rand_cat)
rr.process(rand_cat)
nn.calculateXi(dr=dr, rr=rr)

# GUNDAM
data_gun = Table([data_RA, data_DEC])
data_gun['wei'] = 1.
rand_gun = Table([rand_RA, rand_DEC])
rand_gun['wei'] = 1.

par = gun.packpars(kind='acf',
                   nsept=len(theta_bins) - 1,
                   logsept=False,
                   septmin=theta_bins[0],
                   dsept = theta_bins[1] - theta_bins[0],
                   estimator='LS',
                   doboot=False)
```

```
par.mxh3 = 0

cnt = gun.acf(data_gun, rand_gun, par, nthreads=nthreads, write=False)
```

Listing 4.2: Pseudocode comparison of the test implementation and settings using each package: Corrfunc, TreeCorr and GUNDAM.

Runtime as a function of catalogue size N is shown in figure 4.1, where we can see that Corrfunc follows an almost $\mathcal{O}(N^2)$ time complexity for large enough N, whereas `TreeCorr`, while not the most efficient code for these low values of N, becomes closer to the other codes for larger N.

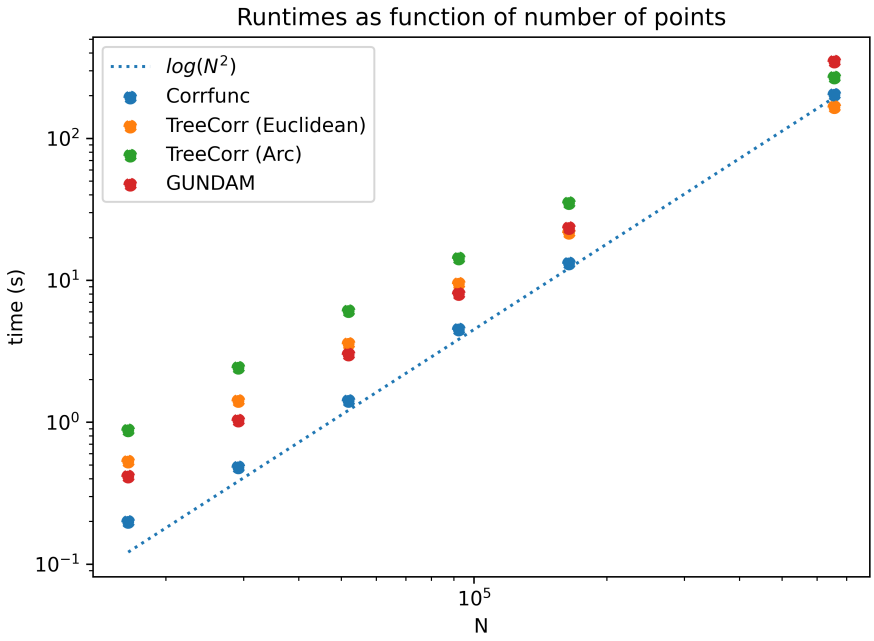


Figure 4.1: Runtime as a function of catalogue size N (with the same number of objects in one single random catalogue) using different Python packages. 'Euclidean' and 'Arc' refer to the distance formula used: 'Euclidean' is a straight line approximation whereas 'Arc' calculates the distance over the spherical surface.

Because of this and because it also offers built-in patch-based resampling and covariance matrix estimation, I adopted `TreeCorr` for pair counts, calculating correlation functions and D_2, subsampling surveys, and covariance estimation.

4.2. Generating random catalogues

The Landy-Szalay estimator requires normalizing data pair counts against those from random catalogues. Conceptually, randoms are catalogues of uniformly distributed points that replicate the survey's geometry, so that survey incompleteness is not misinterpreted as actual fluctuations in the density of objects in the sky.

In practice, how one constructs randoms depends on the survey:

- **Mock surveys:** generally have no geometric defects (holes or irregular footprints), so random catalogues can be simple uniform point samples across the sky region;
- **Real surveys:** true survey footprints almost always contain holes and irregularities, which do not correspond to genuine cosmic underdensities. These are usually caused by bright foreground objects that obfuscate tiny regions in the sky. It is therefore essential that the randoms reproduce these holes by computing a survey footprint mask and applying it to them. However, constructing footprint masks is outside the scope of my work, since I will only be using simulations.

4.2.1 Generating uniform randoms on the spherical surface

Because object coordinates in the survey (and randoms) are given in (RA, Dec), a simple Cartesian sampling will not yield uniform density on the sky, and will only be a good approximation for small spatial patches. Instead, random points must be generated on the unit sphere according to the correct angular probability distribution.

We first transform coordinates:

- $\phi =$ RA (in radians);
- $\theta = \pi/2 -$ Dec (in radians).

so that (ϕ, θ) are standard spherical polar angles.

Figure 4.2 shows coordinates ϕ and θ on a unit sphere.

The probability that a point lies in an infinitesimal cone on a sphere is:

$$P(\Omega)d\Omega$$

which must satisfy the relation:

$$\int_S P(\Omega)d\Omega = 1 \tag{4.1}$$

For a uniform distribution of points on the unit sphere with surface area 4π, then,

$$P(\Omega) = \frac{1}{4\pi}\,. \tag{4.2}$$

On the other hand,

$$P(\Omega)d\Omega = P(\phi,\theta)d\phi d\theta \tag{4.3}$$

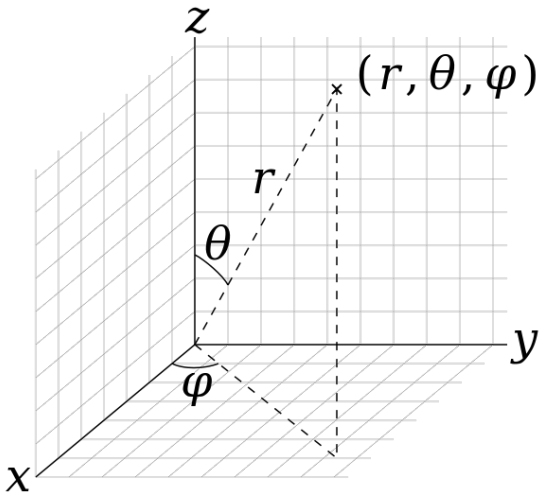


Figure 4.2: Spherical coordinates ϕ and θ as used in this description. (source: Wikimedia Commons [61])

and since the element of solid angle is:

$$d\Omega = \sin\theta d\phi d\theta \tag{4.4}$$

then we can say:

$$\frac{1}{4\pi}\sin\theta d\phi d\theta = P(\phi,\theta)d\phi d\theta \tag{4.5}$$

or

$$P(\phi,\theta) = \frac{\sin\theta}{4\pi}\ . \tag{4.6}$$

As you can see, this is the kind of relation we would expect: the probability of finding a point when moving in the ϕ-direction, the direction parallel to the equator, does not change if θ remains fixed. However, the probability decreases as you go further to the poles, into lower θ.

Marginalizing, we have:

$$P(\theta) = \frac{\sin\theta}{2}, \quad P(\phi) = \frac{1}{2\pi}\ . \tag{4.7}$$

The corresponding Cumulative Distribution Function (CDF), $F_X(x) = P(X \leq x)$, are then:

$$F_\theta(\theta) = \frac{1-\cos\theta}{2}, \quad F_\phi(\phi) = \frac{\phi}{2\pi} \tag{4.8}$$

Inverse transform sampling states that, given a cumulative distribution F_X and a uniform variable $U \in [0,1]$, the random variable $X = F_X^{-1}(U)$ has the distribution F_X. This means that if we sample uniformly in $[0,1]$ (the codomain of a CDF) and calculate the inverse of the CDF for each coordinate ϕ, θ, then we can obtain a distribution of points uniformly

spaced on the spherical surface. Specifically, if u and v are independent, uniform random numbers in $[0, 1]$, then the angular coordinates (ϕ, θ) for a point distributed correctly are:

$$\phi = 2\pi u, \quad \theta = \cos^{-1}(2v - 1) \tag{4.9}$$

A final transformation can then return us back to (RA, Dec) coordinates.

Fig. 4.3 shows an example random catalogue obtained with this method and plotted on a sphere.

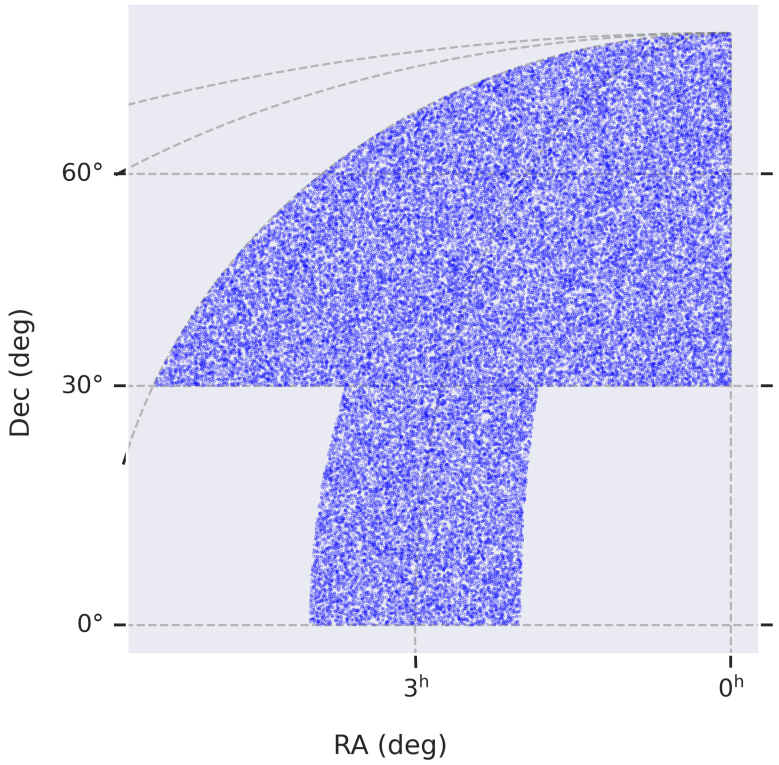


Figure 4.3: Example random catalogue with $N = 300,000$ points and downsampled to 50,000 for representation here, generated over an almost entire octanct of the spherical surface using the method described in this subsection.

This method is described in [62], where one may find additional references to back up this algorithm.

4.2.2 Restricting to rectangular boxes

The method described above was then modified by me: If we wish to generate points only within a subregion (RA_{min}, RA_{max}, Dec_{min}, Dec_{max}), we must restrict the uniform variables

to corresponding bounds in the CDF space:

$$u' = u_{\text{min}} + (u_{\text{max}} - u_{\text{min}}) \times \text{Uniform}(0,1) \tag{4.10}$$

$$v' = v_{\text{min}} + (v_{\text{max}} - v_{\text{min}}) \times \text{Uniform}(0,1) \tag{4.11}$$

where $u_{\text{min}} = F_\phi(\phi_{\text{min}})$, $u_{\text{max}} = F_\phi(\phi_{\text{max}})$, and similarly for v using F_θ.

More generally, for the simplest of irregular surveys - one that can be decomposed into many such rectangular boxes -, if a total of N points is desired for the random catalogue throughout the entire area, one can apply the algorithm above to each of the partial regions and perform the assignment of points per region i scaled according to its fractional area A_i in (u,v) space:

$$N_i = N\frac{A_i}{\sum_j A_j} \tag{4.12}$$

thus guaranteeing a uniform distribution of N points throughout the entire survey.

4.3. Regularization of covariance matrices

A final technical issue concerns the stability of covariance matrices, particularly those associated with the D_2 estimator.

The angular correlation function $\omega(\theta)$ has relatively simple covariance properties, but scaled number counts $\mathcal{N}(<\theta)$, from which $D_2(\theta)$ is derived, are strongly self-correlated: each value of $\mathcal{N}(<\theta)$ accumulates contributions from all smaller angles. As a result, the profile of $D_2(\theta)$ is highly sensitive to changes in the correlation structure. Estimating the covariance of D_2 thus requires more bootstrap/jackknife resamplings to converge than for $\omega(\theta)$, and even then, the resulting matrices exhibit large condition numbers, indicating numerical instability in inversion.

For a covariance matrix, the condition number κ is the ratio of its largest to smallest eigenvalue (diagonal variance):

$$\kappa(C) = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$$

A large κ means that the matrix inversion (needed to compute χ^2 or weighted errors) amplifies tiny numerical errors in the covariance matrix entries and assigns unreliable weights. In practice, the extreme eigenvalue spread in D_2 covariance matrices makes this inevitable. My assumption - which deserves further testing - is that this arises intrinsically from the cumulative nature of $\mathcal{N}(<\theta)$.

4.3.1 Diagonal loading

To mitigate this, I applied regularization by diagonal loading, adding a small number λ to each diagonal element:

$$C' = C + \lambda I \tag{4.13}$$

where I is the identity matrix. This procedure increases small variances (corresponding to the smallest eigenvalues) much more than large variances, thereby reducing the condition number without substantially altering the dominant modes.

4.3.2 Choice of λ

I adopted two criteria for setting λ:

1. **Condition number constraint:** require $\kappa(C') \leq \kappa_{\text{target}}$. For target κ, this corresponds to
$$\lambda = \frac{\lambda_{\text{max}} - \kappa_{\text{target}}\lambda_{\text{min}}}{\kappa_{\text{target}} - 1} \tag{4.14}$$
which comes simply from saying that
$$\kappa_{\text{target}} = \frac{\lambda_{\text{max}} + \lambda}{\lambda_{\text{min}} + \lambda} \, . \tag{4.15}$$

2. **Variance preservation constraint:** require that the added diagonal variance not exceed a fraction f of the mean variance:
$$\lambda = f\langle \text{diag}(C) \rangle \, . \tag{4.16}$$

I then selected the maximum of these two prescriptions, with the requirement $\lambda \geq 0$. In essence, this allows me to shrink the condition number while setting an upper bound on the increase in variance. Throughout my work, all covariance matrices for $D_2(\theta)$ using the direct method are regularized with $\kappa_{\text{target}} = 10^4$ and $f = 0.01$. In Chapter 5, I show that these choices reduce condition numbers to manageable levels without significantly distorting the covariance structure.

5. Method evaluation

In this chapter, I use mock surveys to investigate methodological choices in redshift binning, survey region area selection, survey subsampling strategies, random catalogues, and different approaches to estimating θ_H and its uncertainties. I also present analytical profiles for the fractal dimension profiles (D_2 curves). Finally, I synthesize these tests to identify the most reliable strategy for θ_H estimation and draw conclusions on best practices.

5.1. Mock surveys

I use ready-made galaxy mock simulations derived from the Mice Grand Challenge (MICE-GC) N-body simulation, designed to model the large-scale structure of the Universe assuming large-scale homogeneity and isotropy (FLRW cosmology). These simulations provide estimates of galaxy luminosities tailored for wide-field surveys such as the Dark Energy Survey (DES) [63] and the Euclid mission [64]. The mock galaxy catalogue is constructed using a hybrid approach that combines a Halo Occupation Distribution (HOD) (see e.g. [65–67]) model with Halo Abundance Matching (HAM) (see e.g. [68–70]) to populate Friends-of-Friends (FoF; see e.g. [71–73]) dark matter halos identified in the MICE-GC simulation. The cosmological parameters adopted for the simulation are $\Omega_m = 0.25$ (matter density fraction), $\sigma_8 = 0.8$ (matter density fluctuations averaged over a scale of $8\ h^{-1}$ Mpc), $n_s = 0.95$ (scalar spectral index, tells us how the amplitude of the primordial density perturbations depend on scale), $\Omega_b = 0.044$ (baryonic matter density fraction), $\Omega_\Lambda = 0.75$ (dark energy density fraction), and $h = 0.7$ (the dimensionless Hubble parameter).

The full version 2 of the MICE mock galaxy catalogue [74–76] comprises 499,609,997 galaxies, making it one of the largest currently in existence, and is publicly available through the CosmoHub portal* [77]. For my analysis, I use a smaller subset of the catalogue, covering roughly 3,438 square degrees from a sky region where DES i-band magnitudes are complete down to $i = 24$, and Euclid H-band magnitudes are complete down to H ~ 23.0, up to $z \sim 1.4$, though I select only galaxies up to $i = 23.5$. I bin the galaxy catalogue in redshift and select bins with left edges at $z = 0.3, 0.5, 0.7, 0.9,$ and 1.1,

*https://cosmohub.pic.es

each with a width of $\Delta z = 0.01$ *. These particular redshifts were chosen in order to probe θ_H and its uncertainty estimates at epochs with varying degrees of galaxy clustering, and to allow a direct comparison with the expected redshift evolution of θ_H. A bin width of $\Delta z = 0.01$ is adopted because it is commonly used in the literature to minimize projection effects within redshift slices, see e.g.[60].

Fig. 5.1 shows the subset of MICECAT that I use in my work: specifically, it shows the galaxy distribution at each redshift bin over the entire survey area considered, whereas Table 5.1 presents the average redshift and number of objects in each redshift bin.

Redshift bin	Average redshift	Number of objects
0.3–0.31	0.305	1,543,290
0.5–0.51	0.505	1,895,610
0.7–0.71	0.705	1,938,871
0.9–0.91	0.905	1,652,408
1.1–1.11	1.105	1,105,749

Table 5.1: Average redshift and number of objects in each redshift bin for the MICECAT catalogue within the entire completeness region used in this dissertation (as plotted in Fig. 5.1).

For all the analyses presented, I consider 30 log-spaced angular separation bins from 0.5 to 20.0 deg. For any one survey region, I generate 20 random catalogues with the same number of objects as in the survey itself (the *data* catalogue).

5.1.1 Survey subdivision into regions

From the survey presented above, I subpartition the catalogue into multiple smaller regions of varying areas. This allows for a detailed analysis of how survey area affects estimates of θ_H, as well as the variability of θ_H across different regions. In addition, this approach enables a comparison of the uncertainty estimates between individual regions, the uncertainties from regions of different areas, and their reliability as beyond-sample variance estimates (which, in practice, are likely to be limited).

For this study, I select:

- 16 regions of 600 deg^2;
- 16 regions of 1200 deg^2;
- one region covering the full survey area.

*Although the width of the bin should affect the statistics when looking at different redshifts, I chose to maintain one single bin width. Tests I made suggested that at the order of magnitude of the number of objects is sufficiently high for the difference between bins to not be relevant.

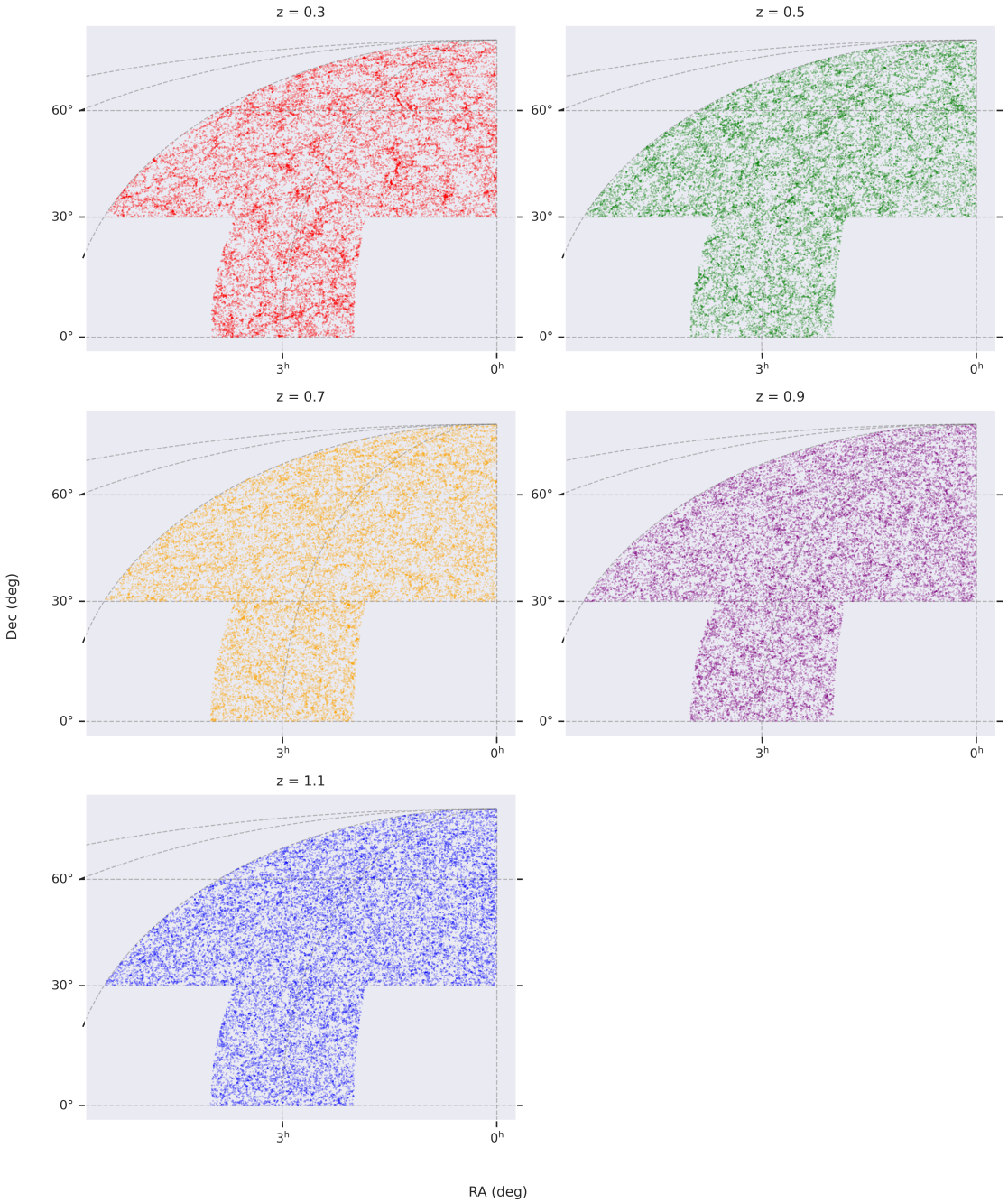


Figure 5.1: Snapshots from the MICECAT catalogue within the completeness region used in this dissertation. Each plot shows one redshift bin of width $\Delta z = 0.01$.

All regions are contained within the completeness region mentioned, as illustrated in Fig. 5.1 from the previous section.

All regions are rectangular in equatorial coordinates (RA, Dec) and contain no holes nor irregularities. Because regions of the same area contain roughly the same number of objects, the statistics derived from any two regions of identical area are comparable. While differences in region shape could, in principle, affect the estimation of θ_H, preliminary tests suggested that such effects are negligible; however, these results are not formally presented and should therefore be treated cautiously.

Regions of the same area are selected using a simple rejection sampling algorithm applied to the edge points ($\mathrm{RA_{min}}$, $\mathrm{RA_{max}}$, $\mathrm{Dec_{min}}$, $\mathrm{Dec_{max}}$). Each candidate region is required to satisfy the area formula for a spherical surface with unit radius:

$$A = \left(\phi_{\mathrm{max}} - \phi_{\mathrm{min}}\right) \left|\cos\theta_{\mathrm{min}} - \cos\theta_{\mathrm{max}}\right| \,, \tag{5.1}$$

where A is in steradians and ϕ_{min}, ϕ_{max}, θ_{min}, θ_{max} are expressed in radians. Here, ϕ and θ correspond to the transformed spherical coordinates equivalent to RA and Dec, as described in Section 4.2.1.

Although these subpartitions overlap considerably, reasonably large survey areas are necessary to mimic expected near-future survey sizes, and the survey footprint imposes practical limits. In practice, the overlap does not pose a significant issue; however, care is required when interpreting variations between regions, as the small number of regions may underestimate true variability.

Table 5.2 summarizes the areas, edge coordinates, and total number of objects in each region, while Table 5.3 presents the average number of objects per region area and per redshift bin. The spatial distribution of the 32 regions on the sky is shown in Fig. 5.2, illustrated for $z = 0.3$, although the region boundaries are identical for all other redshifts. Because of the overlap, not all regions are visible in the figure.

Table 5.2: Summary of regions: areas, coordinate ranges, and number of objects.

Region	Area	RA min–max	DEC min–max	Number of objects
1	600	6.00–85.56	57.00–76.00	1,473,704
2	600	37.00–48.92	2.00–66.00	1,449,338

Continued on next page

Region	Area (deg^2)	RA min–max (deg)	DEC min–max (deg)	Number of objects
3	600	38.00–65.38	36.00–76.00	1,433,679
4	600	48.00–80.01	32.00–59.00	1,389,278
5	600	2.00–84.77	60.00–83.00	1,462,325
6	600	55.00–84.99	37.00–72.00	1,408,837
7	600	21.00–83.23	52.00–73.00	1,446,106
8	600	34.00–69.37	43.00–78.00	1,447,100
9	600	44.00–57.41	1.00–53.00	1,407,193
10	600	5.00–68.57	32.00–44.00	1,383,418
11	600	26.00–60.32	44.00–88.99	1,449,765
12	600	59.00–87.04	36.00–74.00	1,410,366
13	600	17.00–57.24	36.00–58.00	1,404,773
14	600	47.00–83.93	31.00–53.00	1,384,303
15	600	16.00–54.63	41.00–68.00	1,421,101
16	600	20.00–75.95	54.00–85.00	1,459,862
1	1200	30.00–87.78	38.00–78.00	2,857,692
2	1200	19.00–64.28	32.00–83.00	2,840,059
3	1200	30.00–76.39	33.00–85.00	2,826,258
4	1200	6.00–53.13	32.00–77.00	2,832,542
5	1200	37.00–80.69	31.00–84.00	2,854,724
6	1200	25.00–84.33	32.00–62.00	2,808,385
7	1200	14.00–81.78	42.00–78.00	2,856,379
8	1200	0.00–58.86	40.00–87.00	2,844,197
9	1200	33.00–86.92	37.00–82.00	2,850,527
10	1200	35.00–57.19	3.00–85.00	2,865,241
11	1200	15.00–60.00	30.00–75.00	2,837,056
12	1200	30.00–75.00	30.00–75.00	2,829,154
13	1200	45.00–90.00	30.00–75.00	2,839,836
14	1200	0.00–45.00	30.00–75.00	2,804,259
15	1200	15.00–60.00	30.00–75.00	2,837,056
16	1200	5.00–50.00	30.00–75.00	2,814,555
–	full area	0.00–90.00	0.00–89.95	8,135,924

Table 5.3: Average number of objects per region area and redshift bin.

Region area (deg^2)	Redshift bin z	Average number of objects
600	0.3	266,788
	0.5	336,740
	0.7	341,200
	0.9	287,230
	1.1	194,986
1200	0.3	527,368
	0.5	669,082
	0.7	675,683
	0.9	575,954
	1.1	389,281
full area	0.3	1,543,288
	0.5	1,895,609
	0.7	1,938,870
	0.9	1,652,408
	1.1	1,105,749

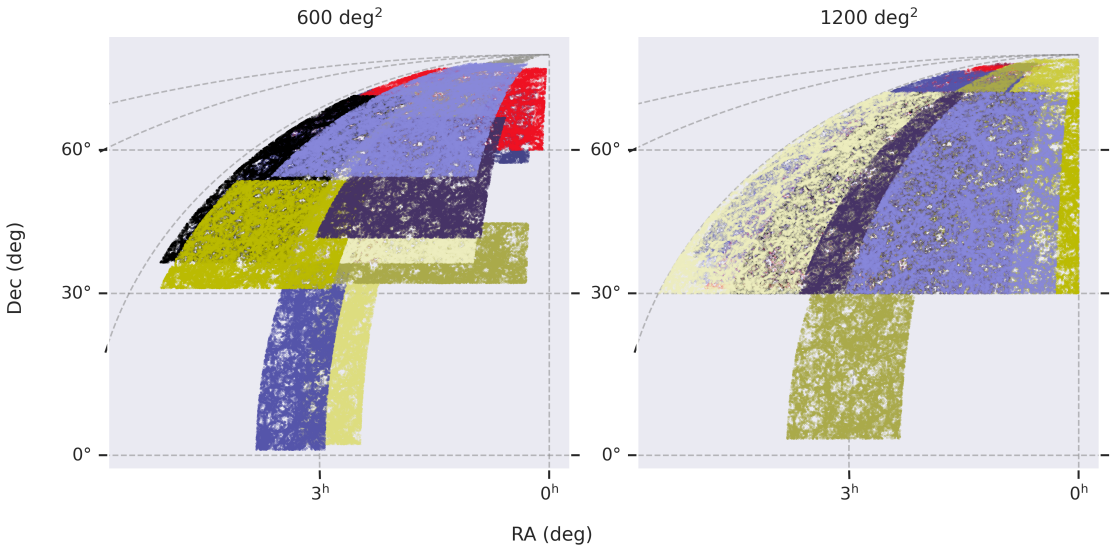


Figure 5.2: Distribution on the sky of the 16 regions of 600 deg^2 and of the 16 regions of 1200 deg^2 selected from MICECAT within the completeness region (example for $z = 0.3$). Because of the overlap between regions, many are partially or completely hidden.

5.1.2 Uncertainty estimation via subsampling

To estimate uncertainties on θ_H, each survey region is further subdivided into patches. For this purpose, I use the built-in functions in `TreeCorr`, which partitions a region using a K-means clustering algorithm. This algorithm produces roughly equal-area patches, using information from its pair-counting procedure to determine initial patch centers. Fig. 5.3 illustrates one example region subdivided into 100 patches, with each color representing a distinct patch.

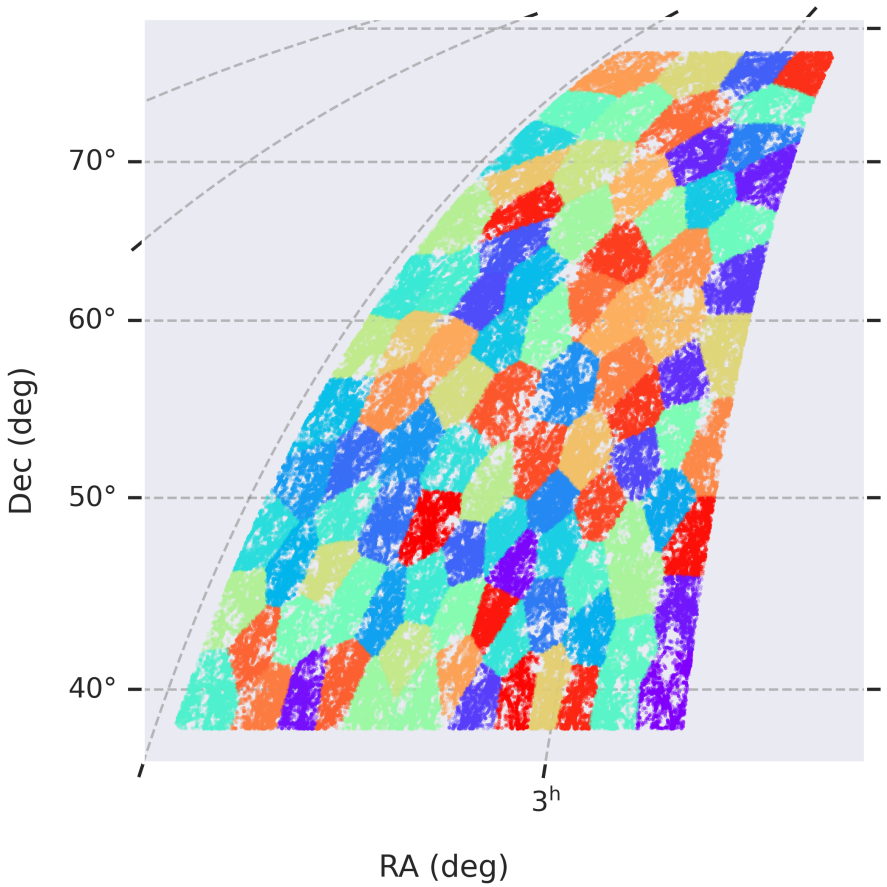


Figure 5.3: Example of region subsampled into 100 roughly equal-area patches by `TreeCorr` for uncertainty estimation.

The choice of resampling method, bootstrap or jackknife, must be tested empirically. Previous studies, e.g. [58], demonstrated that appropriate weighting schemes allow both methods to recover the correct covariance matrix structure for $\omega(\theta)$ up to survey limits. However, validation is required for the fractal dimension $D_2(\theta)$, which exhibits a highly auto-correlated structure (in Eq. 2.36, one may see that the value at each angular separation bin θ depends on the sum of pair counts beginning at an initial bin $\theta_i = 0$ up until the desired final bin, θ_f). In addition, the number of bootstrap resamplings and the number of patches may influence the performance and suitability of each method.

To analyze this, I examine the covariance matrices obtained from both jackknife and bootstrap under different configurations, using the $D_2(\theta)$ profile. The fractal dimension is chosen because its strong auto-correlation makes it most sensitive to resampling artifacts; methods validated for $D_2(\theta)$ are likely to perform well for $\omega(\theta)$ as well. The covariance matrix is analyzed in two parts: the diagonals (eigenvalues/variances) and the correlation matrix, which normalizes the covariance matrix by the diagonals to reveal the relative structure. This normalization is important because the raw covariance matrix entries can differ by orders of magnitude, obscuring the correlation patterns.

As explained in Chapter 3, there exists an upper limit on the number of patches: each patch must be large enough that its typical scale exceeds the correlation scale of the region, ensuring approximate statistical independence (a necessary prerequisite of both jackknife and bootstrap resampling techniques). Conversely, maximizing the number of patches improves the resolution of within-region variability. For 30 angular separation bins, I adopt a minimum of 30 patches so that jackknife resampling yields 30 independent estimates of $\omega(\theta)$ and $D_2(\theta)$. The number of patches is varied up to 200, which remains below the upper limit for the full-area region.

Fig. 5.4 compares the covariance matrices obtained from jackknife and bootstrap (with $N_{\text{bootstrap}} = 500$), with each row corresponding to a fixed number of patches. The figure shows that both methods recover a similar overall structure. Nonetheless, there are, in theory, limitations to the bootstrap procedure:

1. For jackknife, the number of resamplings equals the number of patches, so increasing patches naturally increases the number of independent resamples. For bootstrap, the number of resamplings is independent of the number of patches and remains fixed at 500;

2. Bootstrap resamples can include patches multiple times while omitting others, potentially producing unrealistic configurations. Previous studies [43, 58] suggest this is not problematic, but the current statistics may reveal limitations.

Given these considerations, jackknife resampling likely provides the most accurate representation of uncertainties in θ_H, and it is computationally faster. Therefore, I adopt jackknife as the preferred method for all subsequent analyses.

From Fig. 5.4, the covariance structure stabilizes above $N_{\text{patch}} = 100$ for the full-area region. To ensure robustness, I adopt $N_{\text{patch}} = 200$ for this region (still below the upper

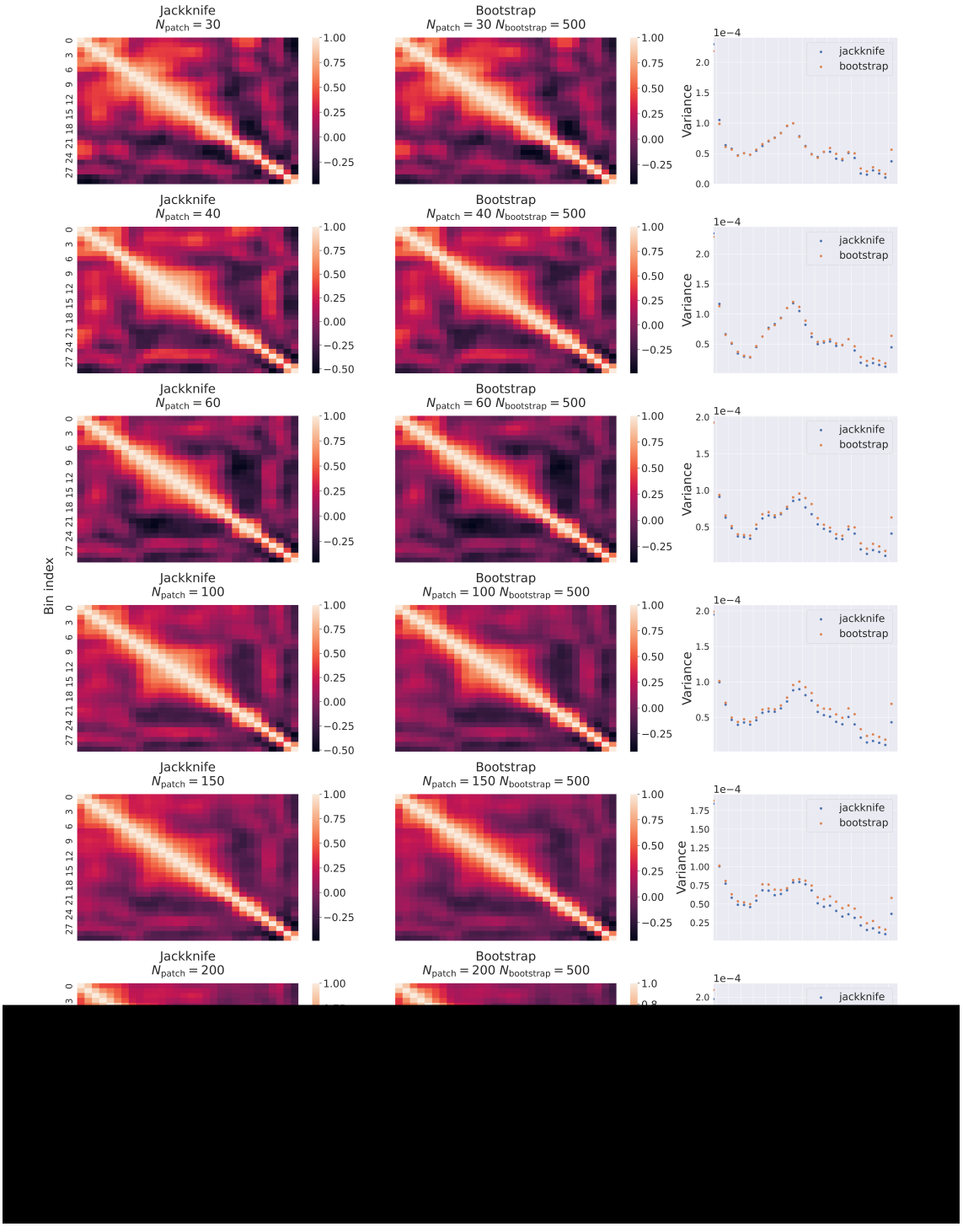


Figure 5.4: Comparison of correlation matrix obtained for one $D_2(\theta)$ curve for the 'full area' region of MICECAT using either jackknife resampling or bootstrap resampling. On the right hand side, the plots show the variance (diagonals) of the covariance matrices for jackknife (*left*) compared to bootstrap (*centre*), given the configuration in the corresponding row. One determines that $N_{\text{bootstrap}} = 500$ is probably enough resamplings to converge to a stable representation of the underlying covariance structure, with both resampling methods leading to matching results, though the variances between the two methods do diverge slightly as you increase number of patches.

limit). To study how the optimal number of patches depends on survey area, I analyze jackknife covariance matrices for different region sizes in Fig. 5.5. For both the full area and a 1200 deg^2 region, the covariance structure stabilizes at roughly 100-150 patches. Smaller regions, such as 600 deg^2, may require more patches, but increasing patches risks introducing correlations between patches. Bootstrap could potentially be used in such cases, although its limitations remain. For the 600 deg^2 regions I have selected $N_{\text{patch}} = 50$, but this may have been a mistake. It is important to test whether my analysis of the covariance matrix obtained from jackknife and from bootstrap is necessary for the resampling approach, where I do not actually make use of such a matrix, but of the resampled curves themselves - it may be that this latter method may be more lenient.

All analyses in this section are performed for $z = 0.3$, where clustering is strongest (clustering becomes stronger as the Universe evolves due to the effect of gravity). Decisions made at this redshift are expected to be valid for higher redshift bins as well.

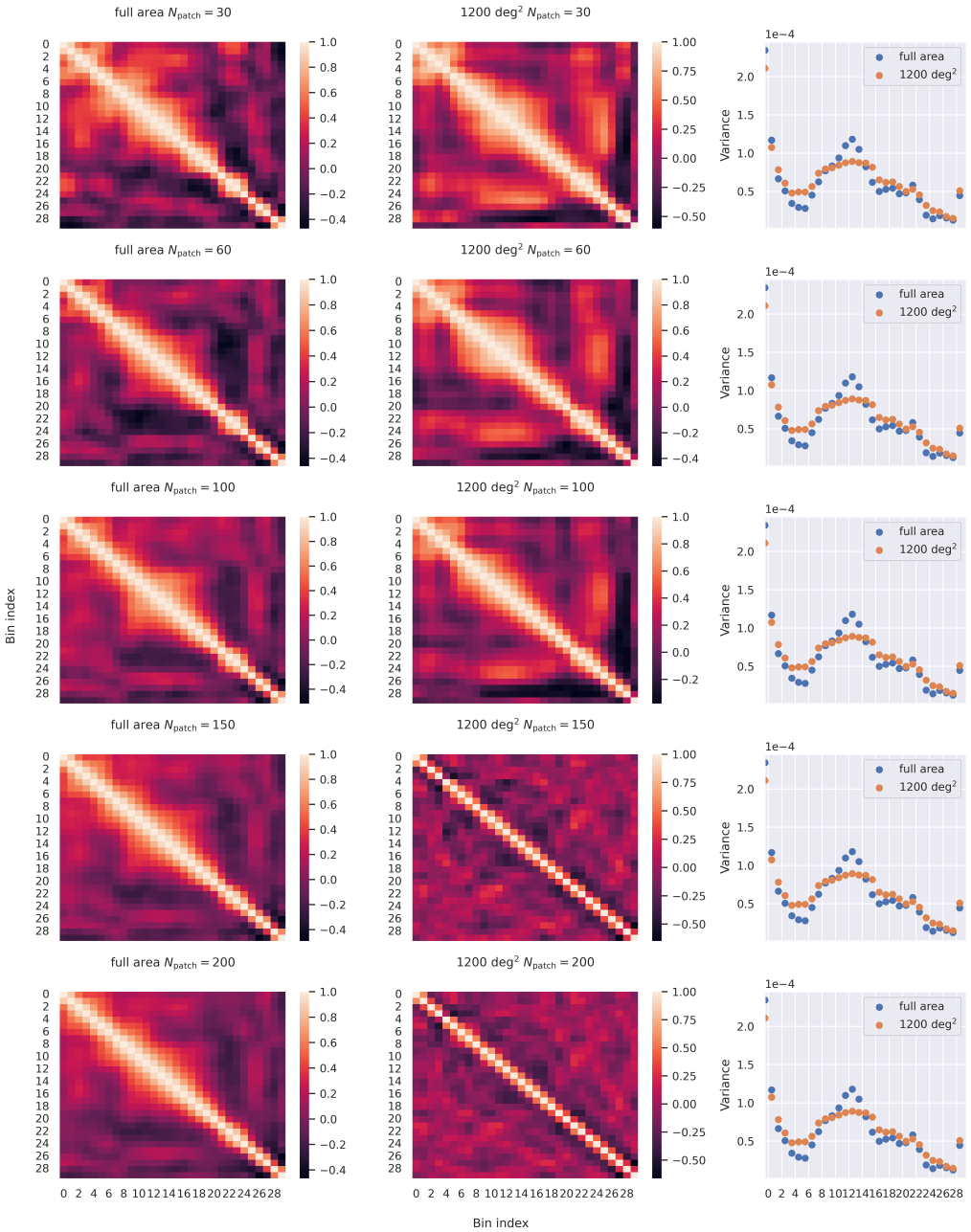


Figure 5.5: Same as Figure 5.4 but comparing the correlation matrices obtained for one $D_2(\theta)$ curve for the 'full area' region vs. the '1200 deg^2' region of MICECAT using jackknife resampling.

5.2. Symbolic Regression: D_2 covariance matrix regularization

In the case of the symbolic regression (SR) approach, uncertainty estimation relies on numerically stable covariance matrices. An analysis of the covariance matrices for $\omega(\theta)$ and $D_2(\theta)$ through their condition numbers κ shows that, while the covariance matrix for $\omega(\theta)$ is well-behaved and can be inverted directly (condition numbers in the range $10^2 - 10^4$), the covariance matrix for $D_2(\theta)$ consistently exhibits very high condition numbers ($10^{15} - 10^{18}$), regardless of the configuration tested (large number of patches, large number of bootstrap resamplings). This indicates that the issue is intrinsic to the structure of D_2, rather than a problem in the estimation procedure. As such, following the strategy outlined in Chapter 3, I applied a regularization procedure to the D_2 covariance matrices to make them numerically stable for inversion, and in addition used the pseudo-inverse* instead of the true inverse.

Fig. 5.6 shows, in logarithmic scale, the eigenvalue spectrum of one example D_2 covariance matrix before and after regularization. One sees that very small eigenvalues (corresponding to directions of low variance) are significantly affected, although they remain much smaller than the largest eigenvalues. In contrast, the largest eigenvalues are barely changed. For uncertainty estimation in the SR approach, where these eigenvalues enter as weights of the form $1/\text{variance}$, this implies that data points originally associated with small variances (thus high weights) are now attributed larger variances (smaller weights). In effect, this amounts to discarding some information for the purposes of SR fitting.

In practice, examining the percent residual of the regularized correlation matrix relative to the original one (Fig. 5.7), one finds that the regularization only impacts the correlations at very high bins (corresponding to the largest angular separations, far beyond the angular homogeneity scales of interest). Even there, the maximum percent difference is only about 4-5%. While this alone does not fully quantify the effect on the uncertainty estimation of θ_H, it indicates that the technique modifies the covariance matrices only mildly, while still enabling stable application and yielding results for θ_H measurements.

*The pseudo-inverse A^+ of a matrix A generalizes the inverse A^{-1}: it coincides with the usual inverse when the latter exists, and otherwise provides the closest possible generalization. Unlike the inverse, it is defined for any rectangular or singular matrix.

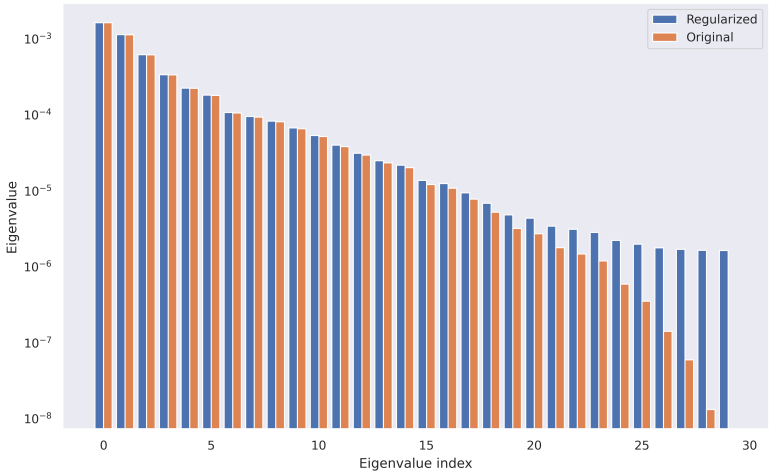


Figure 5.6: Eigenvalues of the covariance matrix for $D_2(\theta)$ from one single, random region used in this work. The orange bars correspond to the original eigenvalues, pre-regularization, and the blue bars to the post-regularization eigenvalues. The vertical axis is in logarithmic scale. One sees that very small eigenvalues, corresponding to directions of very low variance, are greatly affected, yet still much smaller than the largest components, whereas the largest eigenvalues are barely affected at all.

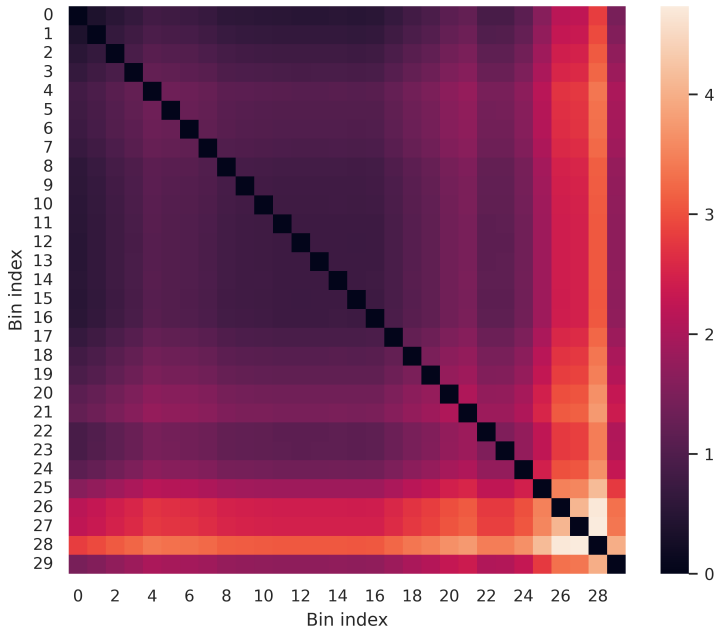


Figure 5.7: Percent relative residuals of regularized correlation matrix compared to original correlation matrix. I find that this sort of regularization only affects high bin indices more significantly, with a maximum percentual difference of around 4-5%, corresponding to the largest angular separations probed, much higher than the angular homogeneity scales expected.

5.3. Testing the approaches to θ_H estimation

SR fitting approach

I begin by comparing the direct and indirect methods using the symbolic regression (SR) approach. As explained earlier, this method fits the measured $\omega(\theta)$ or $D_2(\theta)$ profiles with symbolic regression, and uncertainties are propagated using an estimated covariance matrix. The assumption here is that the uncertainty of the fitted function from varying its parameters adequately reflects the underlying statistical uncertainty of the profiles in the survey. For this comparison I use the 1200 deg^2 regions, which balance having sufficient patches for subsampling while also providing multiple independent estimates of θ_H, and I show the results obtained using the linear approximation method, which takes θ_H as the value obtained from the intersection of the single SR fit with the homogeneity threshold.

Figure 5.8 presents θ_H measurements for the 16 regions across all redshift bins, together with the bootstrap distribution of these values, for the **indirect** method. At low redshift, measurements are widely dispersed, while at higher redshifts the results start to overlap between bins. This behavior does not reflect a feature of the survey, but rather limitations of the method:

1. **Integration from $\theta = 0$:** The indirect method requires integrating from angular separation $\theta = 0$. Since my pair counts begin at $\theta = 0.5$ deg, extrapolating the SR fit to smaller scales introduces instability. The SR fit is not constrained below this scale and may produce nonphysical behavior (including complex values), leading to $D_2(\theta)$ profiles without the expected positive slope towards homogeneity;

2. **Inadequate SR fits:** Despite its flexibility, SR does not always capture the large-scale behavior of $\omega(\theta)$. With the stopping criterion WMSE ≤ 1, some profiles are underfit. Tightening the criterion could mitigate this but could also substantially increase computational time;

3. **Dependence on the random distribution:** As discussed in Chapter 2, $\mathcal{N}(< \theta)$ can be calculated numerically by comparison of pair counts to a theoretically expected random distribution (see Eq. 2.35 and the surrounding description). A mismatch between the theoretical expectation and the random distribution estimator RR could create discrepancies and create issues with this method.

In contrast, Fig. 5.9 shows the results for the **direct** method. Here, dispersion and overlap are somewhat reduced, although not eliminated. Some overlap between redshift bins

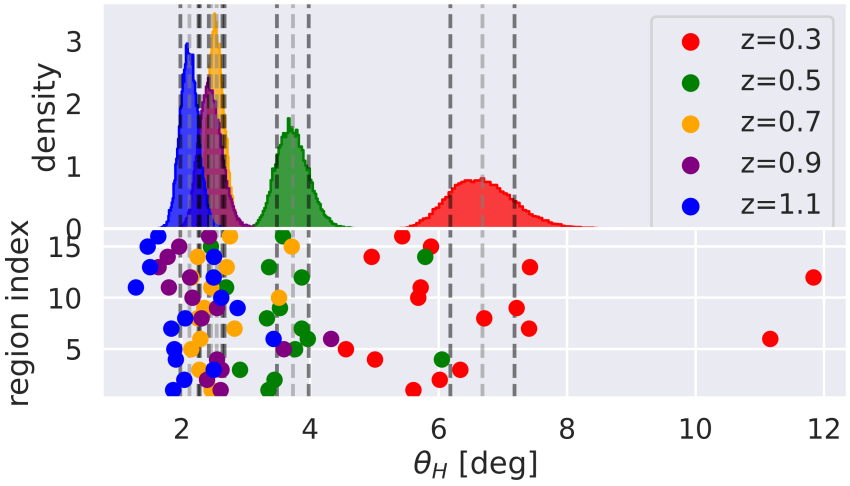


Figure 5.8: **SR fitting approach - indirect method:** θ_H measurements for many different regions and for each redshift bin. We can see the expected trend in θ_H and higher variability of the measurements at low redshift. The vertical dashed lines correspond to the 16/84% percentiles of each distribution.

persists. As will be made clear when analysing the results for the resampling approach, these problems are not driven by the survey itself, nor only by artifacts of the SR fitting procedure.

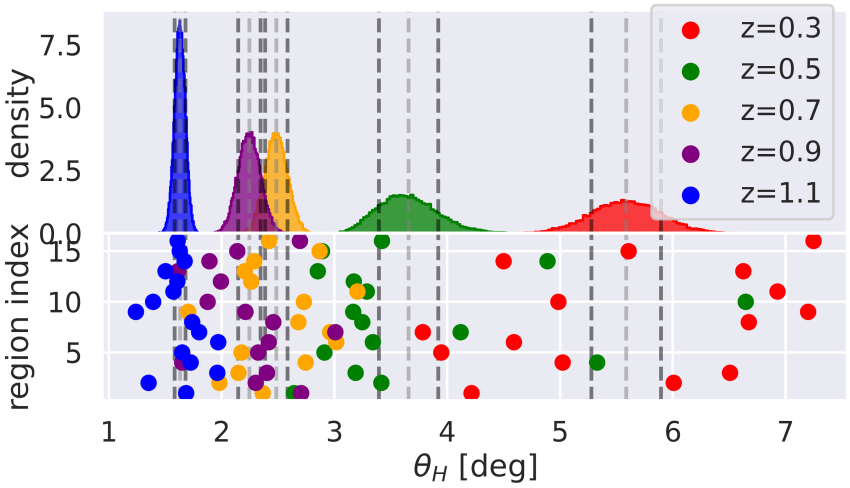


Figure 5.9: **SR fitting approach - direct method:** θ_H measurements for many different regions and for each redshift bin. We do not see as large an overlap in the bootstrap distributions between different redshift bins as we do with the indirect method. The vertical dashed lines correspond to the 16/84% percentiles of each distribution.

Overall, the SR fitting approach provides some advantages - flexibility, analytic forms for the profiles, and the possibility of comparison to theory - but in practice, difficulties in fitting broad trends, along with the problematic indirect method, limits the effectiveness of this

approach for robust θ_H method (though, at the very least, the indirect method may yet be improved).

Resampling approach

I next test the resampling approach, in which θ_H is estimated directly from jackknife resamples of the survey. I find that applying smoothing to the spline improves dispersion of results compared to a spline with no smoothing*. Fig. 5.10 shows the results for the **indirect** method. As with the SR approach, this method struggles, particularly at $z = 1.1$. The simplest and most likely solution to fix this would be to recompute pair counts starting from smaller separations, which would stabilize the integral. A useful question for future work is whether, once corrected, jackknife resampling of $\omega(\theta)$ could perform better than jackknife resampling of $D_2(\theta)$ (the latter might require increasing the number of patches to achieve stability of the covariance structure).

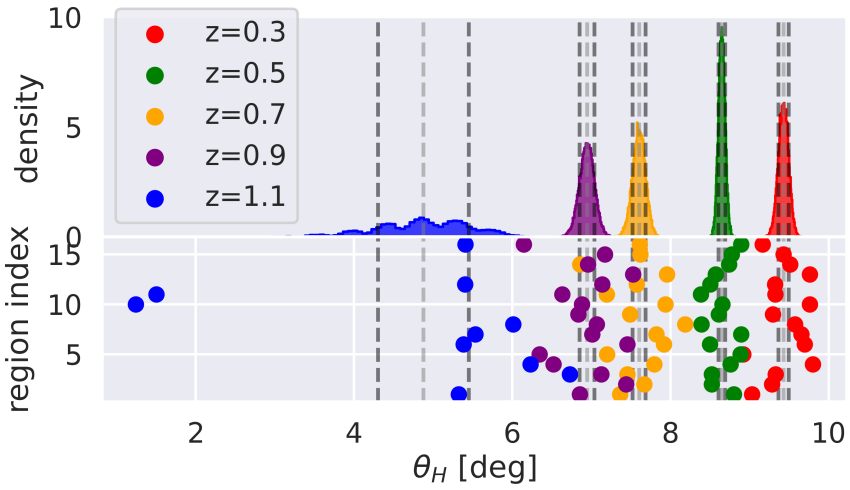


Figure 5.10: **Resampling approach - indirect method:** θ_H measurements for many different regions and for each redshift bin. This method, similarly to what happens with SR, suffers from the problem with the integral for the scaled number counts, leading to bad fits or bad D_2 profiles from $\omega(\theta)$. The vertical dashed lines correspond to the 16/84% percentiles of each distribution.

By contrast, Fig. 5.11 presents the results for the **direct** method (still with a smoothed spline). Here, dispersion is small and distributions of θ_H for different redshifts are well separated, consistent with theoretical expectations. At low redshifts, where clustering is stronger, variability between regions is higher, whereas at high redshift measurements

*The smoothing factor s is related to the number of knots of the spline: the number of knots is increased until the smoothing condition is satisfied:

$$\sum_i (y_i - \text{spl}(x_i))^2 \leq s$$

, where (x_i, y_i) are the data points that the spline fits

are very precise. The expected evolutionary trend of θ_H (studied, for instance, in [42, 47]) is clearly visible: the homogeneity scale evolves rapidly at low redshift (where a small change in z corresponds to a larger cosmic time interval) but changes little at higher redshifts.

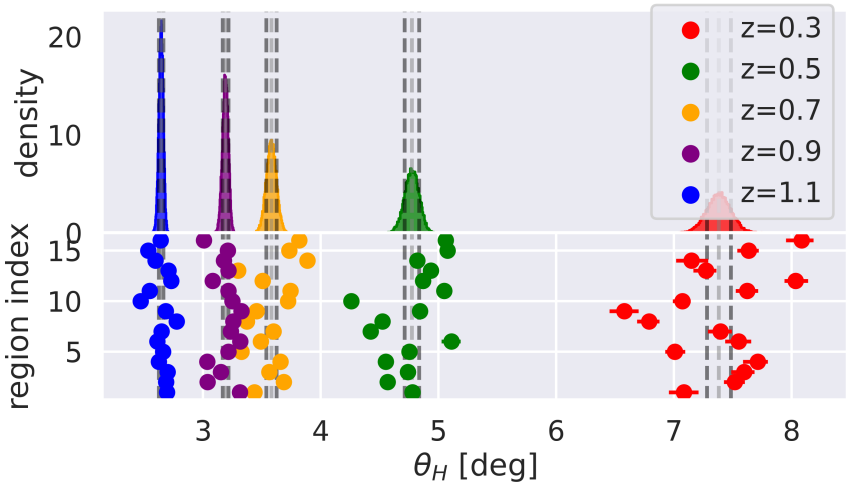


Figure 5.11: **Resampling approach - direct method:** θ_H measurements for many different regions and for each redshift bin. This method produces deterministic θ_H measurements with lower variability than obtained with SR, though still showing higher variability in measurements at low redshift, and no overlap in the bootstrap distributions of θ_H at different redshift slices. The vertical dashed lines correspond to the 16/84% percentiles of each distribution.

The resampling approach also has the practical advantage of being transparent and easy to interpret. Fig. 5.12 and 5.13 illustrate this clearly: the first shows all jackknife resampled $D_2(\theta)$ curves with their spline fits and final θ_H, while the second shows the distribution of θ_H values from resamples. The median is adopted as the central estimate – given the typically skewed distributions, the median is a more reliable central estimate than the mean –, with the 16th-84th percentiles defining the uncertainty interval, though one may leave it up to discussion whether a 2σ interval (2.5-97.5th percentiles) would perhaps be more appropriate.

Choice of approach and method

In light of these results, I adopt the **resampling direct method** as the standard procedure for the remainder of this work. This method is accurate, robust, computationally efficient, free of the stochasticity inherent in SR fitting, and straightforward to interpret. Although SR fitting initially appeared promising, requiring only a single fit, naturally discarding small-scale oscillations, and producing analytic functional forms, its instability, computational

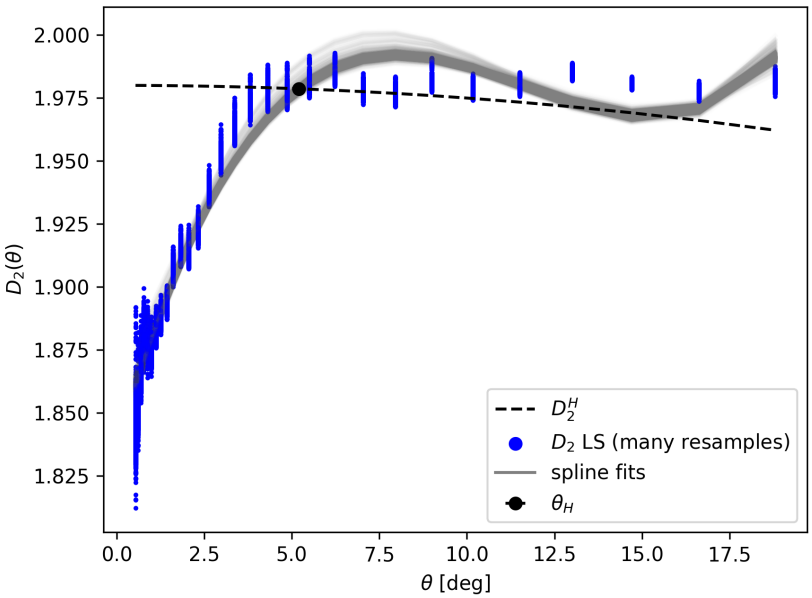


Figure 5.12: D_2 profiles from jackknife resamples (in blue) along with the corresponding spline fits (in gray). The black point is the final, measured θ_H.

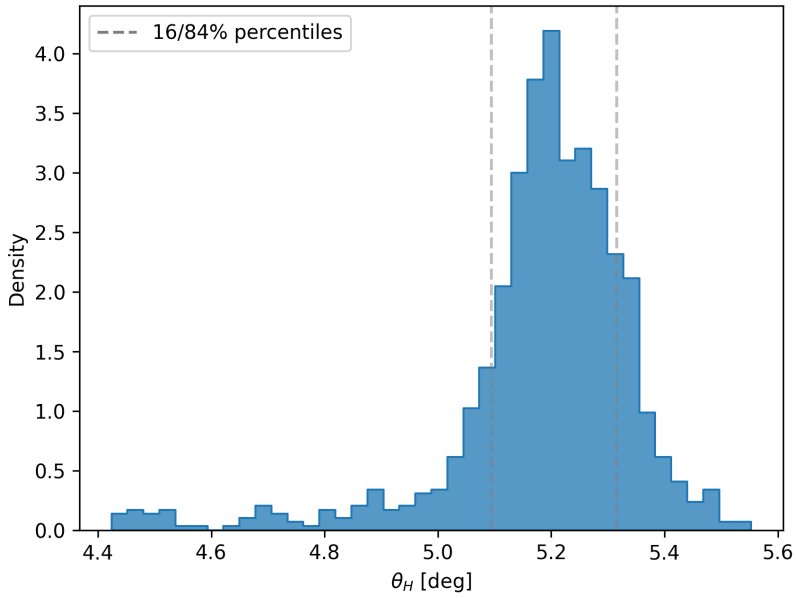


Figure 5.13: Example histogram of θ_H measurements obtained for each jackknife resample shown in Fig. 5.12.

cost, and difficulties in generating reliable uncertainty estimates ultimately made it less practical than resampling.

5.4. Comparison of uncertainty estimation methods

Although the SR fitting approach has already been shown to be problematic for θ_H estimation, it is still useful to compare the two SR-based uncertainty estimation methods (linear approximation and nested sampler) against the resampling approach. To do this, I focus on the **direct method** applied to the full-area region, which avoids the issues identified with the indirect method. Figure 5.14 presents the uncertainties obtained from SR (linear approximation, nested sampler) and from resampling. Since the uncertainties are not always symmetric, I display both the negative and positive error components. Although only the full-area region is shown, the results are broadly similar for smaller regions as well.

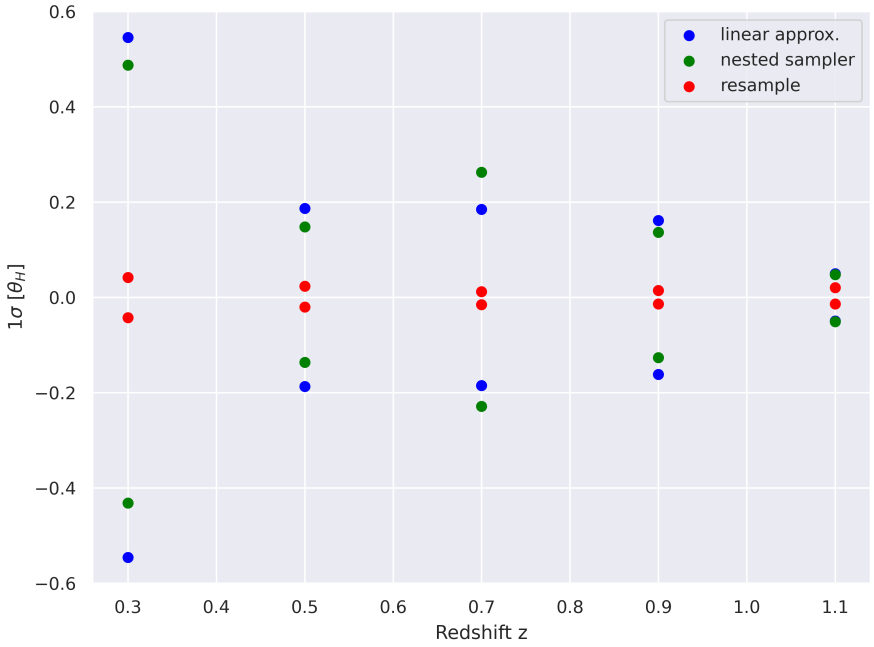


Figure 5.14: Uncertainties estimated on θ_H measurements for the full area region using different methods - linear approximation and nested sampler (from SR approach, described in subsection 3.3.6) and resample, applying the direct method for θ_H estimation.

Several points are apparent from this comparison:

1. **Consistency of SR methods:** The nested sampler and the linear approximation methods yield remarkably similar uncertainty estimates. This is somewhat surprising, as the original paper [59] introducing the linear approximation method suggested that it would significantly underestimate uncertainties relative to a more rigorous sampling approach;

2. **Small uncertainties from resampling:** The resampling method produces considerably smaller uncertainties. These are consistent with the distribution of $D_2(\theta)$ curves obtained from jackknife resamples, though the agreement may be somewhat suspiciously tight;

3. **Comparison to bootstrap:** Bootstrap resampling of patches produces more variable $D_2(\theta)$ curves and correspondingly larger uncertainty estimates, closer to those obtained from the SR-based methods. However, as discussed before, bootstrap may not be reliable. A definitive resolution would require mock catalogues constructed under a standard FLRW cosmology to establish the expected level of uncertainty in θ_H given the sample sizes I have.

An important advantage of the resampling approach is its **interpretability**: uncertainties are obtained directly from the empirical distribution of resamples and are thus transparent and easy to communicate. By contrast, the SR-based methods rely on approximations and optimization criteria that act as ”black boxes,” making their results less straightforward to validate.

Finally, I note that uncertainties estimated using random catalogues are still much smaller than those obtained from resampling. This reinforces the idea that random-based estimates, while computationally cheap, severely underestimate the true uncertainty and should not be used as a substitute for resampling-based approaches. On the other hand, SR fitting-based methods such as the linear approximation and the nested sampler appear to overestimate uncertainties (if we assume jackknife resampling provides the most reliable estimate, which so far I have no reason to doubt). While overestimation is preferable to underestimation - since it errs on the side of caution - the ultimate goal remains to obtain realistic uncertainty estimates that are as close as possible to the true variability.

5.5. Robustness of θ_H estimates across survey area and redshift

In this section, I study the robustness of the θ_H estimates with respect to survey area, their variability across redshift bins, and whether increasing the survey area improves measurement stability. A further goal is to identify which redshift bins are most suitable for determining reliable θ_H estimates. As seen in previous sections, the resampling direct method successfully reproduces some of the expected trends of θ_H with redshift. In addition, I also verify here that the characteristic BAO scales do not coincide with the values of θ_H obtained.

Fig. 5.15 and Fig. 5.16 present the results obtained for the 600 deg^2 and 1200 deg^2 regions, respectively, using the resampling direct method. From these figures, several conclusions can be drawn:

- As expected, the dispersion in θ_H for each redshift bin is larger for the smaller 600 deg^2 regions;
- There is less overlap between θ_H measurements in different redshift bins for larger-area regions compared to smaller-area regions;
- The bootstrap standard deviation is slightly larger than the individual uncertainties of each region, implying that individual uncertainty estimates tend to underestimate the variability across regions (which is expected, since internal estimates are limited to the volume enclosed by the region);
- If θ_H measurements are to be used to constrain cosmological parameters, higher-redshift bins ($z \gtrsim 0.7$) are preferable. At high redshifts, uncertainties are smaller and variability between regions is reduced, making these measurements comparatively more robust than those at low redshift.

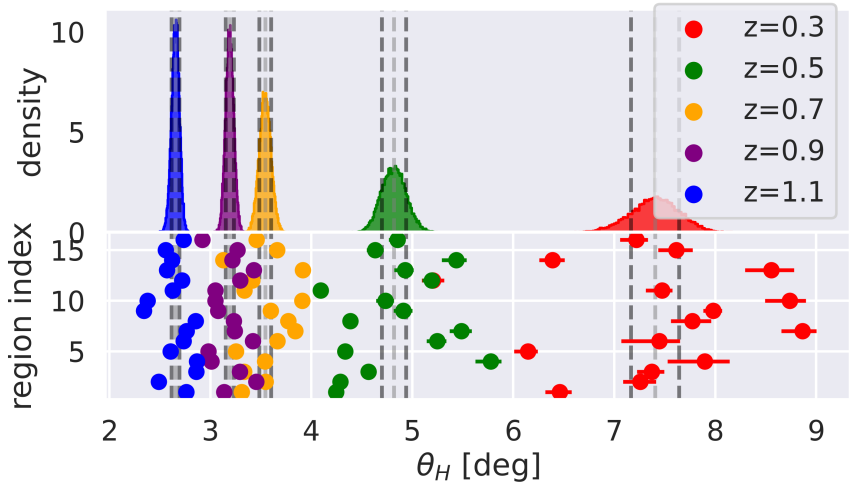


Figure 5.15: Histograms of bootstrap distributions from θ_H measurements plus uncertainties via the resample approach with the direct method over many different regions of area 600 deg^2.

A second question concerns the dependence of θ_H estimates on survey area: how do the measurements evolve when comparing regions of 600 deg^2, 1200 deg^2, and the full-area survey? To address this, Fig. 5.17 plots the bootstrap mean with its associated standard deviation for the 600 deg^2 and 1200 deg^2 regions, alongside the single measurement (with

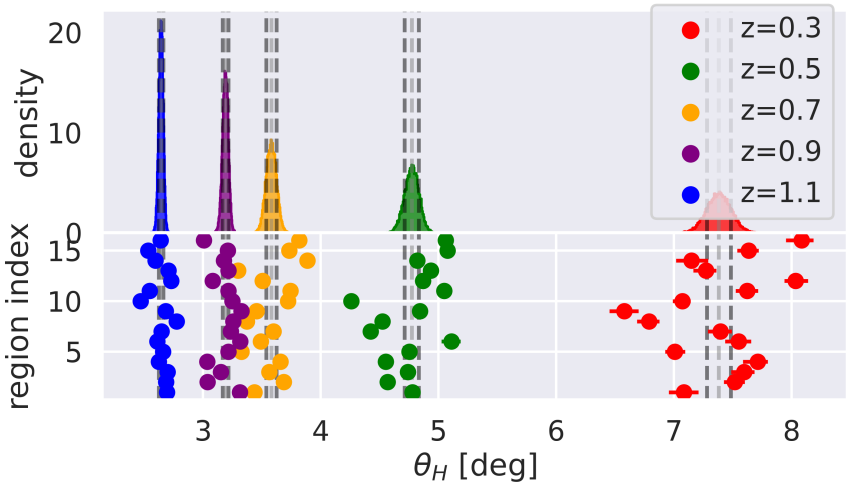


Figure 5.16: Histograms of bootstrap distributions from θ_H measurements plus uncertainties via the resample approach with the direct method over many different regions of area 1200 deg^2.

uncertainty) from the full-area survey. For comparison, the BAO scales are also shown. The following conclusions can be drawn:

- Except at $z = 0.3$, the bootstrap means (and their dispersions) for smaller-area regions match closely the full-area single measurement. This indicates that measuring θ_H across many smaller regions (even overlapping ones) or using one large contiguous area produces consistent results for $z \gtrsim 0.5$;
- Given the small variability at high redshift, the BAO scale does not appear to be significantly close to θ_H. Nonetheless, it may be advisable to fine-tune binning and magnitude cuts to ensure that the two scales remain well separated.

Finally, across all redshift bins and survey areas, the same non-linear, decreasing trend of θ_H with redshift is recovered, consistent with expectations.

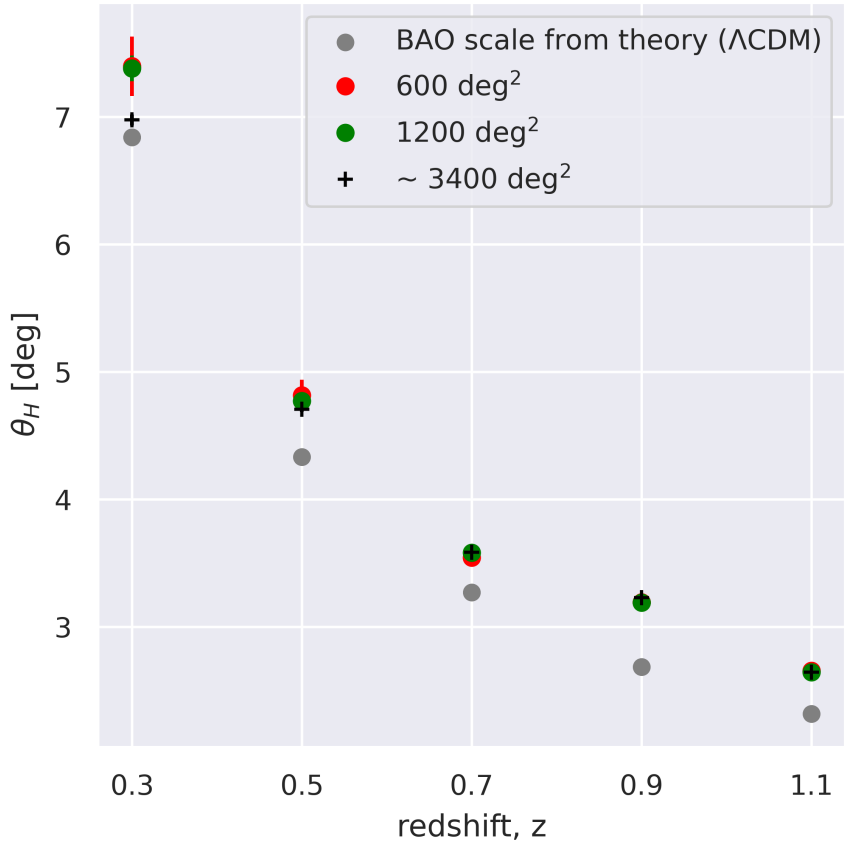


Figure 5.17: Comparison of the bootstrap mean θ_H of many regions of 600 deg^2 and 1200 deg^2 vs the 'full area' region, along with the theoretical expected BAO scale computed according to Eq. 5 of [78] and the relation $\theta_{BAO}(z) = r_s / \left((1+z)D_A(z)\right)$, where r_s is the sound horizon scale at decoupling and D_A is the angular diameter distance at redshift z, see e.g. [21]. I conclude that, except for low redshifts ($z = 0.3$), the bootstrap mean from 600 deg^2 (and 1200 too) matches the expected result obtained from the 'full area' region, although individual measurements of θ_H were previously shown to have high variability at the lower redshifts. I also notice that the BAO scale for low redshifts is too close to the 'full area' region estimate of θ_H, which might be problematic.

5.6. Analytic profiles of D_2

In this section, I present analytic profiles for the fractal dimension D_2 at different redshift bins of width $\Delta z = 0.01$, the same used in the previous analysis. These profiles were obtained from the full area region also discussed before by taking the D_2 curve obtained from the direct method (using the LS estimator for scaled number counts) and applying symbolic regression to this curve.

The symbolic regression was configured as follows:

- **Maximum complexity:** 50;
- **Binary operators:** $+$, $-$, $\times$, $/$, exponentiation;

- **Unary operators:** `exp`, `sinh`, `cosh`, `tanh`, `log`;
- **Constraints:** the exponent operator may only include a single variable x to avoid nested exponentials;
- **Selection criteria:** the candidate expression with the lowest score* amongst those with 1.5 times the lowest mean squared error.

Fig. 5.18 shows the analytic D_2 profiles obtained for each redshift bin, and Table 5.4 lists the mean squared errors for each profile along with the corresponding analytic expressions. These fits were obtained using data points in the angular separation range $[0.5, 17.5]$ degrees.

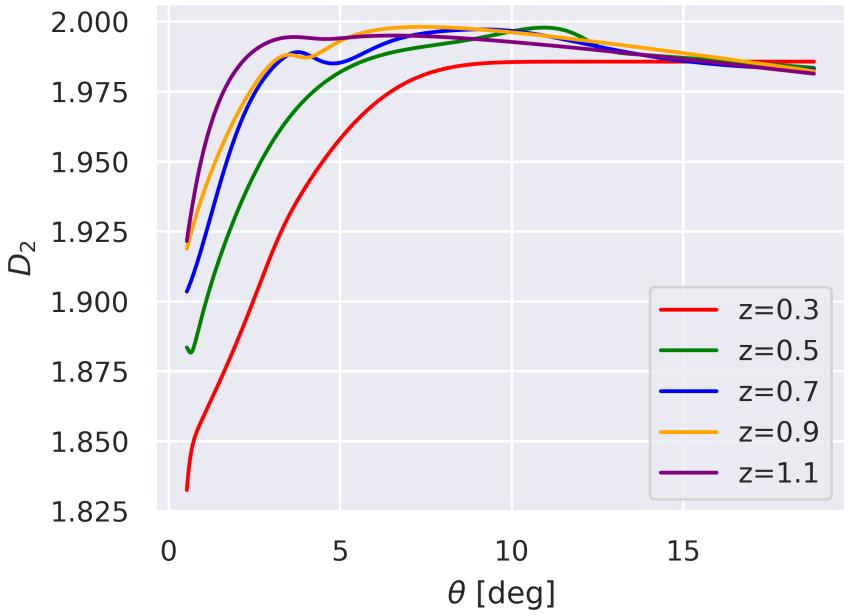


Figure 5.18: Plot of analytic profiles for D_2 obtained for different redshift bins for the full area region of MICECAT 2.

These analytic profiles provide an initial, parametric representation of D_2, whose benefit is in bypassing the computational effort required to calculate pair counts and generate jackknife resamples, offering a stepping stone for future analyses. Nevertheless, the obtained profiles are far from the optimal representations. It is clear from Fig. 5.18 that they overfit to some oscillations in the curves, though it is not clear whether these are actual features of the structure of the Universe in general or features of the survey area considered, or perhaps due to methodological choices in estimating the D_2 LS data points.

*Score is defined as the negated derivative of the log-loss with respect to complexity - if an expression has a much better loss at a slightly higher complexity, it is preferred. In this case, loss is the mean squared error.

Redshift bin	MSE	Complexity	Mathematical expression
$z = 0.3$	5.97×10^{-6}	18	$\cosh\left(a_0 + \tanh\left(a_1^{\theta}\right) \tanh\left(\cosh\left(a_2 e^{a_3\theta + a_4 \cosh\left(\tanh(a_5 \cosh(\theta))\right)}\right)\right)\right)$ $a_0 = 0.3086, \quad a_1 = 6.2824,$ $a_2 = 0.8902, \quad a_3 = 0.0821,$ $a_4 = 0.0821, \quad a_5 = 0.1300$
$z = 0.5$	1.62×10^{-6}	29	$\log\left(a_0 - (a_1\theta + a_2(a_3^{\theta})^{\theta}) \tanh\left(a_4\theta - \left(a_5 + a_6 a_7^{\theta} + \sinh\left(a_8 e^{a_9\theta}\right)\right) \tanh(\theta)\right) + \tanh(a_{10}\theta)\right)$ $a_0 = 6.3980, \quad a_1 = -0.006916,$ $a_2 = -0.081974, \quad a_3 = 0.151788,$ $a_4 = 0.704932, \quad a_5 = 6.3980,$ $a_6 = -33.4298, \quad a_7 = 0.050746,$ $a_8 = 0.047306, \quad a_9 = 0.299275,$ $a_{10} = 0.287567$
$z = 0.7$	3.04×10^{-6}	27	$a_0 + \tanh\left(\cosh\left(\left(a_1 + \theta\left(a_2 + \cosh\left(\theta \cosh\left(\cosh\left(a_3 + a_6\left((a_7\theta^3 + 1)^{a_8} - \tanh(a_4 + \cosh(a_5 + \theta))\right)\right)\right)^{a_9}\right)\right)\right)^{a_{10}}\right)\right)$ $a_0 = 0.996612, \quad a_1 = 0.29386455,$ $a_2 = -0.16337353, \quad a_3 = 1.0374497,$ $a_4 = -0.6110885, \quad a_5 = -5.1695876,$ $a_6 = 1.6063857671933437, \quad a_7 = 0.06294208950113359,$ $a_8 = 0.17139031, \quad a_9 = -0.10387308,$ $a_{10} = 0.16680695$
$z = 0.9$	1.85×10^{-6}	18	$\exp\left(-\tanh\left(\tanh\left(a_0 \tanh(a_1 + a_2\theta) + \frac{a_3\theta}{\cosh(a_4 + \theta \tanh(\frac{a_5}{\cosh(a_6\theta)}))}\right)\right)\right)$ $a_0 = -1.3151599, \quad a_1 = 0.95153487,$ $a_2 = 0.24869674, \quad a_3 = 0.009320084,$ $a_4 = -1.1134975, \quad a_5 = 92.5352,$ $a_6 = 1.5576785$
$z = 1.1$	1.60×10^{-6}	20	$a_0 \exp\left(\tanh\left(\tanh\left(\tanh\left(\exp\left(a_1 e^{a_2\theta} + e^{a_3 e^{\tanh(a_4 e^{\theta})}} + a_5\theta\right)\right)\right)\right)\right)$ $a_0 = 1.060847, \quad a_1 = -0.953825,$ $a_2 = -0.535361, \quad a_3 = -0.082685,$ $a_4 = 0.018811, \quad a_5 = -0.022149$

Table 5.4: Mean squared error and symbolic regression expressions for D_2 in different redshift bins.

Further investigation is warranted, and a more scientifically sound selection criterion for the symbolic regression best fit functions should be validated and used. For these reasons, I am not able to conclude whether the D_2 curve at all redshifts follow the same simple and general profile, or if the mathematical structure can in fact only be described by different equations.

6. Conclusion

The goal of this work was to develop a consistent and robust methodology to test the cosmological principle through tests of cosmic homogeneity. Given current observational tensions and the fact that the cosmological principle, that underpins the standard model of cosmology, has not been thoroughly tested yet, it is essential to design reliable tools for homogeneity analyses. With the advent of future large-scale galaxy surveys, fractal dimension analysis has become a promising avenue, enabling the determination of both the fractal dimension profile and the angular scale of transition to homogeneity, θ_H.

To this end, I examined and systematized existing approaches for estimating θ_H and their associated uncertainties. Previous work in the literature has lacked consistency, with different authors adopting different definitions of θ_H, and with uncertainties often estimated only through random catalogues, an approach insufficient for the precision required in forthcoming surveys. In contrast, I investigated two methodologies to measure θ_H:

- **Indirect method:** based on $\omega(\theta)$ (obtained via the Landy-Szalay estimator), followed by estimating the scaled number counts $\mathcal{N}(<\theta)$ and then obtaining $D_2(\theta)$;
- **Direct method:** based on a numerical Landy-Szalay estimator of the scaled number counts $\mathcal{N}(<\theta)$ to obtain $D_2(\theta)$ "directly".

I then implemented two distinct approaches to determine θ_H from these profiles: **symbolic regression (SR) fitting** and **resampling**.

The SR fitting approach initially appeared promising: it made no assumptions about profile shape, offered flexibility, allowed small-scale oscillations (irrelevant to a measure of *large-scale* homogeneity) to be discarded, and could provide analytic functional forms for $D_2(\theta)$. However, it proved impractical in this context: the method was occasionally slow, uncertainty propagation required approximations, and SR in some cases failed to capture broad trends for certain redshift bins and survey regions.

A simpler but ultimately more successful approach was the **resampling method**, in which jackknife or bootstrap resampling produced many realizations of $D_2(\theta)$ from the same survey region, from which θ_H could be robustly extracted. Of the two, jackknife resampling emerged as the fastest and most reliable choice. Bootstrap results produced consistently higher uncertainties and require further study to determine their validity.

These methods were tested on the MICECAT mock galaxy survey, divided into regions of varying area. Several important results followed:

- The SR approach was found to be unreliable, while the resampling method was robust, fast, and easily interpretable;
- The indirect method was hindered by its reliance on pair counts at $\theta < 0.5^\circ$, and showed no clear advantage over the direct method;
- Among uncertainty estimation strategies, the linear approximaton and nested sampler methods gave results consistent with one another, in contrast to previous claims in the literature [43], but both substantially overestimated uncertainties relative to the resampling method;
- Survey area impacts θ_H estimates to some extent, and variability across regions is high at low redshift but significantly reduced at higher redshifts;
- High-redshift bins ($z \gtrsim 0.7$) provide the most stable and reliable θ_H measurements for cosmological applications, with uncertainties that more closely reflect cosmic variance;
- Averaging θ_H across many small subregions yields results consistent with measurements from the full survey region (except at $z = 0.3$; one possible explanation is that at lower redshift clustering is stronger and the small subregions are too small to estimate properly an overall angular homogeneity scale);
- The obtained θ_H values appear not to be contaminated by BAO scales (or at least, not significantly so), though this requires further validation.

The main takeaway of this dissertation is the development of a single, fast, and easily interpretable methodology to estimate θ_H robustly, along with its uncertainties, suitable for application to future large-scale surveys. I have also identified the principal factors that must be taken into account when using θ_H for cosmological analyses, and understood how choices of redshift binning and magnitude cuts can be tuned to separate BAO scales from θ_H, ensuring robustness of the measurement.

Limitations

This analysis has several limitations.

- The comparison between jackknife and bootstrap resampling remains unresolved: bootstrap consistently yielded larger uncertainties, and the reasons for this are not yet understood;
- It is not completely clear whether the analysis of the ideal number of patches, resampling technique, or other parameters, based on the covariance matrix, applies directly to using the data resamples to estimate θ_H;
- The indirect method requires pair counts at smaller angular separations, which were not available here, and this may have limited its performance;
- The SR fitting approach, while not adopted, could be improved in robustness and efficiency, but was not pursued further;
- The results have not yet been validated against FLRW mock catalogues based on N-body simulations, which would allow a direct comparison of the resampling technique to expected θ_H uncertainties;
- Finally, possible contamination from BAO scales, although unlikely to be significant, was not definitively ruled out.

Future work

Several avenues for future research arise naturally from these limitations:

- Validation of the methodology against realistic FLRW mocks based on N-body simulations, to determine whether jackknife uncertainties indeed provide the correct estimate of variability;
- Further study of bootstrap resampling, including its dependence on the number of patches and resamplings, to clarify the origin of its discrepancy with jackknife;
- Evaluation of the ideal confidence interval to consider from the θ_H distribution as the uncertainty measurement obtained with the resampling approach;
- Testing of how survey masks with irregular shapes or holes affect θ_H measurements;
- Investigation of the impact of photometric redshift uncertainties on binning choices;
- Obtaining better quality $D_2(\theta)$ analytic profiles for the different redshift bins, along with confidence regions.;

- Testing my methodology on mock galaxy surveys obtained with different cosmological models;
- Application of the methodology to real survey data, once available, to provide the first large-scale empirical constraints on θ_H at different epochs;
- Exploration of the use of θ_H measurements for cosmological parameter constraints, following approaches such as those of [47].

In summary, this dissertation established a consistent framework for mesuring the angular homogeneity scale θ_H, demonstrated its robustness under a variety of tests, and clarified the methodological choices required for its reliable estimation. These results provide a solid foundation for applying fractal dimension analysis to upcoming galaxy surveys, enabling new tests of the cosmological principle and, ultimately, new constraints on the standard model of cosmology.

References

[1] S. P. Goodwin, J. Gribbin, and M. A. Hendry, “The relative size of the Milky Way,” *The Observatory*, vol. 118, pp. 201–208, Aug. 1998. [Cited on page 1.]

[2] R. G. Clowes, K. A. Harris, S. Raghunathan, L. E. Campusano, I. K. Söchting, and M. J. Graham, “A structure in the early universe at z 1.3 that exceeds the homogeneity scale of the r-w concordance cosmology,” *Monthly Notices of the Royal Astronomical Society*, vol. 429, no. 4, pp. 2910–2916, 01 2013. [Online]. Available: https://doi.org/10.1093/mnras/sts497 [Cited on page 3.]

[3] W. Hu and S. Dodelson, “Cosmic microwave background anisotropies,” *Annual Review of Astronomy and Astrophysics*, vol. 40, no. 1, p. 171–216, Sep. 2002. [Online]. Available: http://dx.doi.org/10.1146/annurev.astro.40.060401.093926 [Cited on page 3.]

[4] D. G. York, J. Adelman, J. E. Anderson, Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. Fan, G. R. Federwitz, L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C.-h. Huang, C. Hull, D. Husby, S.-I. Ichikawa, T. Ichikawa, Z. Ivezic, S. Kent, R. S. J. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider,

M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, S.-i. Wang, M. Watanabe, D. H. Weinberg, B. Yanny, and N. Yasuda, "The sloan digital sky survey: Technical summary," *The Astronomical Journal*, vol. 120, no. 3, p. 1579–1587, Sep. 2000. [Online]. Available: http://dx.doi.org/10.1086/301513 [Cited on pages 3 and 19.]

[5] E. L. Wright. (2005) Las campanas redshift survey. [Online]. Available: https://www.astro.ucla.edu/~wright/lcrs.html [Cited on page 4.]

[6] Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, K. Benabed, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J.-F. Cardoso, B. Casaponsa, H. C. Chiang, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, P. de Bernardis, G. de Zotti, J. Delabrouille, J.-M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, R. Fernandez-Cobos, F. Finelli, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J.-M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, F. Levrier, M. Liguori, P. B. Lilje, V. Lindholm, M. López-Caniego, Y.-Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Mennella, M. Migliaccio, M.-A. Miville-Deschênes, D. Molinari, A. Moneti, L. Montier, G. Morgante, A. Moss, P. Natoli, L. Pagano, D. Paoletti, B. Partridge, F. Perrotta, V. Pettorino, F. Piacentini, G. Polenta, J.-L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, R. Sunyaev, A.-S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, A. Zacchei,

J. P. Zibin, and A. Zonca, “Planck2018 results: Vii. isotropy and statistics of the cmb,” *Astronomy and Astrophysics*, vol. 641, p. A7, Sep. 2020. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201935201 [Cited on page 3.]

[7] R. A. Oliveira, T. S. Pereira, and M. Quartin, “Cmb statistical isotropy confirmation at all scales using multipole vectors,” *Physics of the Dark Universe*, vol. 30, p. 100608, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212686420302533 [Cited on page 3.]

[8] Kester, C. E., Bernui, A., and Hipólito-Ricaldi, W. S., “Probing the statistical isotropy of the universe with planck data of the cosmic microwave background,” *Astronomy and Astrophysics*, vol. 683, p. A176, 2024. [Online]. Available: https://doi.org/10.1051/0004-6361/202348160 [Cited on page 3.]

[9] P. K. Samal, R. Saha, P. Jain, and J. P. Ralston, “Testing isotropy of cosmic microwave background radiation,” *Monthly Notices of the Royal Astronomical Society*, vol. 385, no. 4, pp. 1718–1728, 03 2008. [Online]. Available: https://doi.org/10.1111/j.1365-2966.2008.12960.x [Cited on page 3.]

[10] W. Keel, *The Road to Galaxy Formation*, ser. Springer Praxis Books. Springer Berlin Heidelberg, 2007. [Online]. Available: https://books.google.pt/books?id=BUgJGypUYF0C [Cited on page 4.]

[11] J. A. Peacock, *Cosmological Physics*. Cambridge University Press, 1998. [Cited on pages 6 and 9.]

[12] H. Kragh, ““the most philosophically important of all the sciences”: Karl popper and physical cosmology,” *Perspectives on Science*, vol. 21, no. 3, pp. 325–357, 2013. [Cited on page 5.]

[13] K. R. Popper, *The Logic of Scientific Discovery*. London: Routledge, 2002, originally published in 1959. [Cited on page 5.]

[14] S. Sarkar, B. Pandey, and R. Khatri, “Testing isotropy in the universe using photometric and spectroscopic data from the sdss,” *Monthly Notices of the Royal Astronomical Society*, vol. 483, no. 2, pp. 2453–2464, 11 2018. [Online]. Available: https://doi.org/10.1093/mnras/sty3272 [Cited on page 6.]

[15] S. Yeung and M.-C. Chu, “Directional variations of cosmological parameters from the planck cmb data,” *Phys. Rev. D*, vol. 105, p. 083508, Apr 2022. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.105.083508

[16] X. Wang and Z. Huang, “Testing cosmological principle under copernican principle,” 2025. [Online]. Available: https://arxiv.org/abs/2505.19055

[17] Hu, J. P., Wang, Y. Y., Hu, J., and Wang, F. Y., “Testing the cosmological principle with the pantheon+ sample and the region-fitting method,” *A&A*, vol. 681, p. A88, 2024. [Online]. Available: https://doi.org/10.1051/0004-6361/202347121

[18] P. Fosalba and E. Gaztañaga, “Explaining cosmological anisotropy: evidence for causal horizons from cmb data,” *Monthly Notices of the Royal Astronomical Society*, vol. 504, no. 4, pp. 5840–5862, 04 2021. [Online]. Available: https://doi.org/10.1093/mnras/stab1193 [Cited on pages 6 and 11.]

[19] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J.-P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J.-F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J.-M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J.-M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y.-Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M.-A. Miville-Deschênes, D. Molinari,

L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J.-L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A.-S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca, "Planck 2018 results: Vi. cosmological parameters," *Astronomy and Astrophysics*, vol. 641, p. A6, Sep. 2020. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201833910 [Cited on page 6.]

[20] A. G. Riess, W. Yuan, L. M. Macri, D. Scolnic, D. Brout, S. Casertano, D. O. Jones, Y. Murakami, G. S. Anand, L. Breuval, T. G. Brink, A. V. Filippenko, S. Hoffmann, S. W. Jha, W. D'arcy Kenworthy, J. Mackenty, B. E. Stahl, and W. Zheng, "A comprehensive measurement of the local value of the hubble constant with 1 km s−1 mpc−1 uncertainty from the hubble space telescope and the sh0es team," *The Astrophysical Journal Letters*, vol. 934, no. 1, p. L7, Jul. 2022. [Online]. Available: http://dx.doi.org/10.3847/2041-8213/ac5c5b [Cited on page 6.]

[21] D. Baumann, *Cosmology*. Cambridge University Press, 2022. [Cited on pages 8, 9, and 74.]

[22] R. Wald, *General Relativity*. University of Chicago Press, 1984. [Online]. Available: https://books.google.pt/books?id=ibSdQgAACAAJ [Cited on pages 9 and 10.]

[23] S. Carroll, *Spacetime and Geometry*. Cambridge University Press, 2019. [Online]. Available: https://books.google.pt/books?id=PTGdDwAAQBAJ [Cited on page 9.]

[24] S. Navas, C. Amsler, T. Gutsche, C. Hanhart, J. J. Hernández-Rey, C. Lourenço, A. Masoni, M. Mikhasenko, R. E. Mitchell, C. Patrignani, C. Schwanda, S. Spanier, G. Venanzoni, C. Z. Yuan, K. Agashe, G. Aielli, B. C. Allanach, J. Alvarez-Muñiz, M. Antonelli, E. C. Aschenauer, D. M. Asner, K. Assamagan, H. Baer, S. Banerjee, R. M. Barnett, L. Baudis, C. W. Bauer, J. J. Beatty, J. Beringer, A. Bettini, O. Biebel, K. M. Black, E. Blucher, R. Bonventre, R. A. Briere, A. Buckley, V. D. Burkert, M. A. Bychkov, R. N. Cahn, Z. Cao, M. Carena, G. Casarosa, A. Ceccucci, A. Cerri,

R. S. Chivukula, G. Cowan, K. Cranmer, V. Crede, O. Cremonesi, G. D'Ambrosio, T. Damour, D. de Florian, A. de Gouvêa, T. DeGrand, S. Demers, Z. Demiragli, B. A. Dobrescu, M. D'Onofrio, M. Doser, H. K. Dreiner, P. Eerola, U. Egede, S. Eidelman, A. X. El-Khadra, J. Ellis, S. C. Eno, J. Erler, V. V. Ezhela, A. Fava, W. Fetscher, B. D. Fields, A. Freitas, H. Gallagher, T. Gershon, Y. Gershtein, T. Gherghetta, M. C. Gonzalez-Garcia, M. Goodman, C. Grab, A. V. Gritsan, C. Grojean, D. E. Groom, M. Grünewald, A. Gurtu, H. E. Haber, M. Hamel, S. Hashimoto, Y. Hayato, A. Hebecker, S. Heinemeyer, K. Hikasa, J. Hisano, A. Höcker, J. Holder, L. Hsu, J. Huston, T. Hyodo, A. Ianni, M. Kado, M. Karliner, U. F. Katz, M. Kenzie, V. A. Khoze, S. R. Klein, F. Krauss, M. Kreps, P. Križan, B. Krusche, Y. Kwon, O. Lahav, L. P. Lellouch, J. Lesgourgues, A. R. Liddle, Z. Ligeti, C.-J. Lin, C. Lippmann, T. M. Liss, A. Lister, L. Littenberg, K. S. Lugovsky, S. B. Lugovsky, A. Lusiani, Y. Makida, F. Maltoni, A. V. Manohar, W. J. Marciano, J. Matthews, U.-G. Meißner, I.-A. Melzer-Pellmann, P. Mertsch, D. J. Miller, D. Milstead, K. Mönig, P. Molaro, F. Moortgat, M. Moskovic, N. Nagata, K. Nakamura, M. Narain, P. Nason, A. Nelles, M. Neubert, Y. Nir, H. B. O'Connell, C. A. J. O'Hare, K. A. Olive, J. A. Peacock, E. Pianori, A. Pich, A. Piepke, F. Pietropaolo, A. Pomarol, S. Pordes, S. Profumo, A. Quadt, K. Rabbertz, J. Rademacker, G. Raffelt, M. Ramsey-Musolf, P. Richardson, A. Ringwald, D. J. Robinson, S. Roesler, S. Rolli, A. Romaniouk, L. J. Rosenberg, J. L. Rosner, G. Rybka, M. G. Ryskin, R. A. Ryutin, B. Safdi, Y. Sakai, S. Sarkar, F. Sauli, O. Schneider, S. Schönert, K. Scholberg, A. J. Schwartz, J. Schwiening, D. Scott, F. Sefkow, U. Seljak, V. Sharma, S. R. Sharpe, V. Shiltsev, G. Signorelli, M. Silari, F. Simon, T. Sjöstrand, P. Skands, T. Skwarnicki, G. F. Smoot, A. Soffer, M. S. Sozzi, C. Spiering, A. Stahl, Y. Sumino, F. Takahashi, M. Tanabashi, J. Tanaka, M. Taševský, K. Terao, K. Terashi, J. Terning, U. Thoma, R. S. Thorne, L. Tiator, M. Titov, D. R. Tovey, K. Trabelsi, P. Urquijo, G. Valencia, R. Van de Water, N. Varelas, L. Verde, I. Vivarelli, P. Vogel, W. Vogelsang, V. Vorobyev, S. P. Wakely, W. Walkowiak, C. W. Walter, D. Wands, D. H. Weinberg, E. J. Weinberg, N. Wermes, M. White, L. R. Wiencke, S. Willocq, C. L. Woody, R. L. Workman, W.-M. Yao, M. Yokoyama, R. Yoshida, G. Zanderighi, G. P. Zeller, R.-Y. Zhu, S.-L. Zhu, F. Zimmermann, P. A. Zyla, J. Anderson, M. Kramer, P. Schaffner, and W. Zheng, "Review of particle physics," *Phys. Rev. D*, vol. 110, p. 030001, Aug 2024. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.110.030001 [Cited on page 9.]

[25] P. J. E. Peebles, *Principles of Physical Cosmology*. Princeton University Press, 09 2020. [Online]. Available: https://doi.org/10.23943/princeton/9780691209814.001.0001 [Cited on pages 9 and 14.]

[26] W. Hu, "Frw cosmology," https://background.uchicago.edu/~whu/Courses/Ast242_08/ast242_06.pdf, 2025, accessed October 7, 2025. [Cited on page 10.]

[27] G. Lemaitre, "The expanding universe," *Annales Soc. Sci. Bruxelles A*, vol. 53, pp. 51–85, 1933. [Cited on page 10.]

[28] R. Tolman, "Effect of imhomogeneity on cosmological models," *Proc. Nat. Acad. Sci.*, vol. 20, pp. 169–176, 1934.

[29] H. Bondi, "Spherically symmetrical models in general relativity," *Mon. Not. Roy. Astron. Soc.*, vol. 107, pp. 410–425, 1947.

[30] D. F. C. Torres, "Pushing the boundaries of modern cosmology: physics beyond the copernican principle," Ph.D. dissertation, Universidade Federal do Espírito Santo, Programa de Pós-Graduação em Astrofísica, Cosmologia e Gravitação, 2022. [Cited on pages 10 and 11.]

[31] C. Clarkson, "Establishing homogeneity of the universe in the shadow of dark energy," *Comptes Rendus. Physique*, vol. 13, no. 6–7, p. 682–718, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.crhy.2012.04.005 [Cited on pages 10 and 11.]

[32] D. Alonso, J. García-Bellido, T. Haugbølle, and J. Vicente, "Large scale structure simulations of inhomogeneous lemaître-tolman-bondi void models," *Physical Review D*, vol. 82, no. 12, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1103/PhysRevD.82.123530 [Cited on page 11.]

[33] R. Maartens, "Is the universe homogeneous?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1957, p. 5115–5137, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1098/rsta.2011.0289 [Cited on page 12.]

[34] R. Jimenez, R. Maartens, A. R. Khalifeh, R. R. Caldwell, A. F. Heavens, and L. Verde, "Measuring the homogeneity of the universe using polarization drift," *Journal of Cosmology and Astroparticle Physics*, vol. 2019, no. 05, p. 048, may 2019. [Online]. Available: https://dx.doi.org/10.1088/1475-7516/2019/05/048

[35] S. Räsänen, K. Bolejko, and A. Finoguenov, “New test of the friedmann-lemaître-robertson-walker metric using the distance sum rule,” *Physical Review Letters*, vol. 115, no. 10, Sep. 2015. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.115.101301

[36] S. Cao, J. Qi, Z. Cao, M. Biesiada, J. Li, Y. Pan, and Z.-H. Zhu, “Direct test of the flrw metric from strongly lensed gravitational wave observations,” 2019. [Online]. Available: https://arxiv.org/abs/1910.10365

[37] C. Bengaly, “A null test of the cosmological principle with bao measurements,” *Physics of the Dark Universe*, vol. 35, p. 100966, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212686422000103 [Cited on page 12.]

[38] N. T. N. K. Cao, C. Mayol, A. Salmachi, and A. Labrinidis, “Modelling and 3d printing of a random porous media using iterative function system fractals,” *The APPEA Journal*, vol. 63, pp. 47–55, 2023. [Online]. Available: https://doi.org/10.1071/AJ22144 [Cited on page 12.]

[39] P. Ntelis, J.-C. Hamilton, J.-M. L. Goff, E. Burtin, P. Laurent, J. Rich, N. G. Busca, J. Tinker, E. Aubourg, H. d. M. d. Bourboux, J. Bautista, N. P. Delabrouille, T. Delubac, S. Eftekharzadeh, D. W. Hogg, A. Myers, M. Vargas-Magaña, I. Pâris, P. Petitjean, G. Rossi, D. P. Schneider, R. Tojeiro, and C. Yeche, “Exploring cosmic homogeneity with the boss dr12 galaxy sample,” *Journal of Cosmology and Astroparticle Physics*, vol. 2017, no. 06, p. 019–019, Jun. 2017. [Online]. Available: http://dx.doi.org/10.1088/1475-7516/2017/06/019 [Cited on page 13.]

[40] M. I. Scrimgeour, T. Davis, C. Blake, J. B. James, G. B. Poole, L. Staveley-Smith, S. Brough, M. Colless, C. Contreras, W. Couch, S. Croom, D. Croton, M. J. Drinkwater, K. Forster, D. Gilbank, M. Gladders, K. Glazebrook, B. Jelliffe, R. J. Jurek, I.-h. Li, B. Madore, D. C. Martin, K. Pimbblet, M. Pracy, R. Sharp, E. Wisnioski, D. Woods, T. K. Wyder, and H. K. C. Yee, “The wigglez dark energy survey: the transition to large-scale cosmic homogeneity: Cosmic homogeneity in the wigglez survey,” *Monthly Notices of the Royal Astronomical Society*, vol. 425, no. 1, p. 116–134, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2966.2012.21402.x [Cited on page 15.]

[41] B. L. Dias, F. Avila, and A. Bernui, “Probing cosmic homogeneity in the local universe,” 2023. [Online]. Available: https://arxiv.org/abs/2310.04594 [Cited on pages 13 and 15.]

[42] X. Shao, R. S. Gonçalves, C. A. P. Bengaly, U. Andrade, G. C. Carvalho, and J. Alcaniz, “Can the angular scale of cosmic homogeneity be used as a cosmological test?” 2024. [Online]. Available: https://arxiv.org/abs/2309.07263 [Cited on pages 13 and 67.]

[43] U. Andrade, R. S. Gonçalves, G. C. Carvalho, C. A. P. Bengaly, J. C. Carvalho, and J. Alcaniz, “The angular scale of homogeneity with sdss-iv dr16 luminous red galaxies,” 2022. [Online]. Available: https://arxiv.org/abs/2205.07819 [Cited on pages vii, 15, 23, 24, 58, and 78.]

[44] R. S. Gonçalves, G. C. Carvalho, C. A. P. Bengaly Jr, J. C. Carvalho, A. Bernui, J. S. Alcaniz, and R. Maartens, “Cosmic homogeneity: a spectroscopic and model-independent measurement,” *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 475, no. 1, p. L20–L24, Dec. 2017. [Online]. Available: http://dx.doi.org/10.1093/mnrasl/slx202 [Cited on pages vii, 19, 21, 22, 26, and 30.]

[45] F. Avila, C. P. Novaes, A. Bernui, E. de Carvalho, and J. P. Nogueira-Cavalcante, “The angular scale of homogeneity in the local universe with the sdss blue galaxies,” *Monthly Notices of the Royal Astronomical Society*, vol. 488, no. 1, p. 1481–1487, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1093/mnras/stz1765 [Cited on pages vii, 20, 21, 22, and 23.]

[46] F. Avila, C. P. Novaes, A. Bernui, and E. de Carvalho, “The scale of homogeneity in the local universe with the alfalfa catalogue,” *Journal of Cosmology and Astroparticle Physics*, vol. 2018, no. 12, p. 041–041, Dec. 2018. [Online]. Available: http://dx.doi.org/10.1088/1475-7516/2018/12/041 [Cited on pages vii, 20, 21, 22, 23, and 29.]

[47] X. Shao, C. A. P. Bengaly, R. S. Gonçalves, G. C. Carvalho, and J. Alcaniz, “Cosmological constraints from angular homogeneity scale measurements,” 2025. [Online]. Available: https://arxiv.org/abs/2409.06009 [Cited on pages 15, 20, 21, 24, 67, and 80.]

[48] D. W. Hogg, D. J. Eisenstein, M. R. Blanton, N. A. Bahcall, J. Brinkmann, J. E. Gunn, and D. P. Schneider, “Cosmic homogeneity demonstrated with luminous red

galaxies," *The Astrophysical Journal*, vol. 624, no. 1, p. 54–58, May 2005. [Online]. Available: http://dx.doi.org/10.1086/429084 [Cited on page 15.]

[49] A. Heinesen, "Cosmological homogeneity scale estimates are dressed," *Journal of Cosmology and Astroparticle Physics*, vol. 2020, no. 10, p. 052–052, Oct. 2020. [Online]. Available: http://dx.doi.org/10.1088/1475-7516/2020/10/052 [Cited on page 15.]

[50] C. M. Hirata, "Correlation functions: Generalities," 2009, draft: November 19, 2009, Caltech cosmology notes. [Online]. Available: http://www.tapir.caltech.edu/~chirata/ay211/Hirata_CF.pdf [Cited on page 17.]

[51] R. Giovanelli, M. P. Haynes, B. R. Kent, P. Perillat, A. Saintonge, N. Brosch, B. Catinella, G. L. Hoffman, S. Stierwalt, K. Spekkens, M. S. Lerner, K. L. Masters, E. Momjian, J. L. Rosenberg, C. M. Springob, A. Boselli, V. Charmandaris, J. K. Darling, J. Davies, D. G. Lambas, G. Gavazzi, C. Giovanardi, E. Hardy, L. K. Hunt, A. Iovino, I. D. Karachentsev, V. E. Karachentseva, R. A. Koopmann, C. Marinoni, R. Minchin, E. Muller, M. Putman, C. Pantoja, J. J. Salzer, M. Scodeggio, E. Skillman, J. M. Solanes, C. Valotto, W. van Driel, and L. van Zee, "The arecibo legacy fast alfa survey. i. science goals, survey design, and strategy," *The Astronomical Journal*, vol. 130, no. 6, p. 2598–2612, Dec. 2005. [Online]. Available: http://dx.doi.org/10.1086/497431 [Cited on page 19.]

[52] Euclid Collaboration, F. J. Castander, P. Fosalba, J. Stadel, D. Potter, J. Carretero, P. Tallada-Crespí, L. Pozzetti, M. Bolzonella, G. A. Mamon, L. Blot, K. Hoffmann, M. Huertas-Company, P. Monaco, E. J. Gonzalez, G. De Lucia, C. Scarlata, M. A. Breton, L. Linke, C. Viglione, S. S. Li, Z. Zhai, Z. Baghkhani, K. Pardede, C. Neissner, R. Teyssier, M. Crocce, I. Tutusaus, L. Miller, G. Congedo, A. Biviano, M. Hirschmann, A. Pezzotta, H. Aussel, H. Hoekstra, T. Kitching, W. J. Percival, L. Guzzo, Y. Mellier, P. A. Oesch, R. A. A. Bowler, S. Bruton, V. Allevato, V. Gonzalez-Perez, M. Manera, S. Avila, A. Kovács, N. Aghanim, B. Altieri, A. Amara, L. Amendola, S. Andreon, N. Auricchio, C. Baccigalupi, M. Baldi, A. Balestra, S. Bardelli, R. Bender, F. Bernardeau, C. Bodendorf, D. Bonino, E. Branchini, M. Brescia, J. Brinchmann, S. Camera, V. Capobianco, C. Carbone, S. Casas, M. Castellano, G. Castignani, S. Cavuoti, A. Cimatti, C. Colodro-Conde, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, F. Courbin, H. M. Courtois, A. Da Silva,

H. Degaudenzi, A. M. Di Giorgio, J. Dinis, M. Douspis, F. Dubath, C. A. J. Duncan, X. Dupac, S. Dusini, A. Ealet, M. Farina, S. Farrens, S. Ferriol, S. Fotopoulou, N. Fourmanoit, M. Frailis, E. Franceschi, P. Franzetti, S. Galeotta, W. Gillard, B. Gillis, C. Giocoli, P. Gómez-Alvarez, B. R. Granett, A. Grazian, F. Grupp, S. V. H. Haugan, M. S. Holliman, W. Holmes, I. Hook, F. Hormuth, A. Hornstrup, P. Hudelot, S. Ilić, K. Jahnke, M. Jhabvala, B. Joachimi, E. Keihänen, S. Kermiche, A. Kiessling, M. Kilbinger, R. Kohley, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, O. Lahav, R. Laureijs, D. Le Mignant, P. Liebing, S. Ligori, P. B. Lilje, V. Lindholm, I. Lloro, D. Maino, E. Maiorano, O. Mansutti, S. Marcin, O. Marggraf, K. Markovic, M. Martinelli, N. Martinet, F. Marulli, R. Massey, D. C. Masters, S. Maurogordato, H. J. McCracken, E. Medinaceli, S. Mei, M. Melchior, M. Meneghetti, E. Merlin, G. Meylan, J. J. Mohr, M. Moresco, L. Moscardini, E. Munari, R. Nakajima, R. C. Nichol, S. M. Niemi, C. Padilla, K. Paech, S. Paltani, F. Pasian, J. A. Peacock, K. Pedersen, V. Pettorino, S. Pires, G. Polenta, M. Poncet, L. A. Popa, F. Raison, R. Rebolo, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, C. Rosset, E. Rossetti, B. Rusholme, R. Saglia, Z. Sakr, A. G. Sánchez, D. Sapone, J. A. Schewtschenko, M. Schirmer, P. Schneider, T. Schrabback, M. Scodeggio, A. Secroun, E. Sefusatti, G. Seidel, S. Serrano, C. Sirignano, G. Sirri, L. Stanco, J. L. Starck, J. Steinwagner, A. N. Taylor, and H. I. Teplitz, “Euclid: V. The Flagship galaxy mock catalogue: A comprehensive simulation for the Euclid mission,” *Astronomy and Astrophysics*, vol. 697, p. A5, May 2025. [Cited on page 20.]

[53] R. Laureijs, J. Amiaux, S. Arduini, J. L. Auguères, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, B. Garilli, P. Gondoin, L. Guzzo, J. Hoar, H. Hoekstra, R. Holmes, T. Kitching, T. Maciaszek, Y. Mellier, F. Pasian, W. Percival, J. Rhodes, G. S. Criado, M. Sauvage, R. Scaramella, L. Valenziano, S. Warren, R. Bender, F. Castander, A. Cimatti, O. L. Fèvre, H. Kurki-Suonio, M. Levi, P. Lilje, G. Meylan, R. Nichol, K. Pedersen, V. Popa, R. R. Lopez, H. W. Rix, H. Rottgering, W. Zeilinger, F. Grupp, P. Hudelot, R. Massey, M. Meneghetti, L. Miller, S. Paltani, S. Paulin-Henriksson, S. Pires, C. Saxton, T. Schrabback, G. Seidel, J. Walsh, N. Aghanim, L. Amendola, J. Bartlett, C. Baccigalupi, J. P. Beaulieu, K. Benabed, J. G. Cuby, D. Elbaz, P. Fosalba, G. Gavazzi, A. Helmi, I. Hook, M. Irwin, J. P. Kneib, M. Kunz, F. Mannucci, L. Moscardini, C. Tao, R. Teyssier, J. Weller, G. Zamorani, M. R. Z. Osorio, O. Boulade, J. J. Foumond, A. D. Giorgio, P. Guttridge, A. James, M. Kemp, J. Martignac, A. Spencer,

D. Walton, T. Blümchen, C. Bonoli, F. Bortoletto, C. Cerna, L. Corcione, C. Fabron, K. Jahnke, S. Ligori, F. Madrid, L. Martin, G. Morgante, T. Pamplona, E. Prieto, M. Riva, R. Toledo, M. Trifoglio, F. Zerbi, F. Abdalla, M. Douspis, C. Grenet, S. Borgani, R. Bouwens, F. Courbin, J. M. Delouis, P. Dubath, A. Fontana, M. Frailis, A. Grazian, J. Koppenhöfer, O. Mansutti, M. Melchior, M. Mignoli, J. Mohr, C. Neissner, K. Noddle, M. Poncet, M. Scodeggio, S. Serrano, N. Shane, J. L. Starck, C. Surace, A. Taylor, G. Verdoes-Kleijn, C. Vuerli, O. R. Williams, A. Zacchei, B. Altieri, I. E. Sanz, R. Kohley, T. Oosterbroek, P. Astier, D. Bacon, S. Bardelli, C. Baugh, F. Bellagamba, C. Benoist, D. Bianchi, A. Biviano, E. Branchini, C. Carbone, V. Cardone, D. Clements, S. Colombi, C. Conselice, G. Cresci, N. Deacon, J. Dunlop, C. Fedeli, F. Fontanot, P. Franzetti, C. Giocoli, J. Garcia-Bellido, J. Gow, A. Heavens, P. Hewett, C. Heymans, A. Holland, Z. Huang, O. Ilbert, B. Joachimi, E. Jennins, E. Kerins, A. Kiessling, D. Kirk, R. Kotak, O. Krause, O. Lahav, F. van Leeuwen, J. Lesgourgues, M. Lombardi, M. Magliocchetti, K. Maguire, E. Majerotto, R. Maoli, F. Marulli, S. Maurogordato, H. McCracken, R. McLure, A. Melchiorri, A. Merson, M. Moresco, M. Nonino, P. Norberg, J. Peacock, R. Pello, M. Penny, V. Pettorino, C. D. Porto, L. Pozzetti, C. Quercellini, M. Radovich, A. Rassat, N. Roche, S. Ronayette, E. Rossetti, B. Sartoris, P. Schneider, E. Semboloni, S. Serjeant, F. Simpson, C. Skordis, G. Smadja, S. Smartt, P. Spano, S. Spiro, M. Sullivan, A. Tilquin, R. Trotta, L. Verde, Y. Wang, G. Williger, G. Zhao, J. Zoubian, and E. Zucca, “Euclid definition study report,” 2011. [Online]. Available: https://arxiv.org/abs/1110.3193 [Cited on page 20.]

[54] I. Sevilla-Noarbe, K. Bechtol, M. Carrasco Kind, A. Carnero Rosell, M. R. Becker, A. Drlica-Wagner, R. A. Gruendl, E. S. Rykoff, E. Sheldon, B. Yanny, A. Alarcon, S. Allam, A. Amon, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. L. Burke, J. Carretero, A. Choi, H. T. Diehl, S. Everett, B. Flaugher, E. Gaztanaga, J. Gschwend, I. Harrison, W. G. Hartley, B. Hoyle, M. Jarvis, M. D. Johnson, R. Kessler, R. Kron, N. Kuropatkin, B. Leistedt, T. S. Li, F. Menanteau, E. Morganson, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, A. Pieres, C. Pond, M. Rodriguez-Monroy, J. A. Smith, K. M. Stringer, M. A. Troxel, D. L. Tucker, J. de Vicente, W. Wester, Y. Zhang, T. M. C. Abbott, M. Aguena, J. Annis, S. Avila, S. Bhargava, S. L. Bridle, D. Brooks, D. Brout, F. J. Castander, R. Cawthon, C. Chang, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, M. E. S.

Pereira, T. M. Davis, S. Desai, J. P. Dietrich, P. Doel, K. Eckert, A. E. Evrard, I. Ferrero, P. Fosalba, J. García-Bellido, D. W. Gerdes, T. Giannantonio, D. Gruen, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, E. M. Huff, D. Huterer, D. J. James, T. Jeltema, K. Kuehn, O. Lahav, C. Lidman, M. Lima, H. Lin, M. A. G. Maia, J. L. Marshall, P. Martini, P. Melchior, R. Miquel, J. J. Mohr, R. Morgan, E. Neilsen, A. A. Plazas, A. K. Romer, A. Roodman, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, M. Smith, E. Suchyta, G. Tarle, D. Thomas, C. To, T. N. Varga, R. H. Wechsler, J. Weller, and R. D. Wilkinson, "Dark energy survey year 3 results: Photometric data set for cosmology," *The Astrophysical Journal Supplement Series*, vol. 254, no. 2, p. 24, May 2021. [Online]. Available: http://dx.doi.org/10.3847/1538-4365/abeb66 [Cited on page 20.]

[55] P. Norberg, C. M. Baugh, E. Gaztañaga, and D. J. Croton, "Statistical analysis of galaxy surveys - i. robust error estimation for two-point clustering statistics," *Monthly Notices of the Royal Astronomical Society*, vol. 396, no. 1, p. 19–38, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2966.2009.14389.x [Cited on pages 21 and 23.]

[56] S. D. Landy and A. S. Szalay, "Bias and Variance of Angular Correlation Functions," *Astrophysical Journal*, vol. 412, p. 64, Jul. 1993. [Cited on page 25.]

[57] M. Vargas-Magana, J. E. Bautista, J.-C. Hamilton, N. G. Busca, E. Aubourg, A. Labatie, J.-M. Le Goff, S. Escoffier, M. Manera, C. K. McBride, D. P. Schneider, and C. N. A. Willmer, "An optimized correlation function estimator for galaxy surveys," *Astronomy & Astrophysics*, vol. 554, p. A131, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201220790 [Cited on page 25.]

[58] F. G. Mohammad and W. J. Percival, "Creating jackknife and bootstrap estimates of the covariance matrix for the two-point correlation function," *Monthly Notices of the Royal Astronomical Society*, vol. 514, no. 1, p. 1289–1301, May 2022. [Online]. Available: http://dx.doi.org/10.1093/mnras/stac1458 [Cited on pages 28, 57, and 58.]

[59] K. W. Vugrin, L. P. Swiler, R. M. Roberts, N. J. Stucky-Mack, and S. P. Sullivan, "Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies," *Water Resources Research*, vol. 43, no. 3, 2007. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005WR004804 [Cited on pages 34, 37, and 70.]

[60] E. Sánchez, A. Carnero, J. García-Bellido, E. Gaztañaga, F. de Simoni, M. Crocce, A. Cabré, P. Fosalba, and D. Alonso, "Tracing the sound horizon scale with photometric redshift surveys," *Monthly Notices of the Royal Astronomical Society*, vol. 411, no. 1, pp. 277–288, 01 2011. [Online]. Available: https://doi.org/10.1111/j.1365-2966.2010.17679.x [Cited on pages 40 and 52.]

[61] Wikimedia Commons, "3d spherical coordinates," https://commons.wikimedia.org/wiki/File:3D_Spherical.svg, 2007, public domain image. [Cited on page 47.]

[62] E. W. Weisstein. Sphere point picking. From MathWorld–A Wolfram Resource. [Online]. Available: https://mathworld.wolfram.com/SpherePointPicking.html [Cited on page 48.]

[63] T. M. C. Abbott, M. Adamów, M. Aguena, S. Allam, A. Amon, J. Annis, S. Avila, D. Bacon, M. Banerji, K. Bechtol, M. R. Becker, G. M. Bernstein, E. Bertin, S. Bhargava, S. L. Bridle, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, A. Choi, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, T. M. Davis, J. De Vicente, J. DeRose, S. Desai, H. T. Diehl, J. P. Dietrich, A. Drlica-Wagner, K. Eckert, J. Elvin-Poole, S. Everett, A. E. Evrard, I. Ferrero, A. Ferté, B. Flaugher, P. Fosalba, D. Friedel, J. Frieman, J. García-Bellido, E. Gaztanaga, L. Gelman, D. W. Gerdes, T. Giannantonio, M. S. S. Gill, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. Huterer, D. J. James, T. Jeltema, M. D. Johnson, S. Kent, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, C. Lidman, H. Lin, N. MacCrann, M. A. G. Maia, T. A. Manning, J. D. Maloney, M. March, J. L. Marshall, P. Martini, P. Melchior, F. Menanteau, R. Miquel, R. Morgan, J. Myles, E. Neilsen, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas, C. Pond, M. Rodriguez-Monroy, A. K. Romer, A. Roodman, E. S. Rykoff, M. Sako, E. Sanchez, B. Santiago, V. Scarpine, S. Serrano, I. Sevilla-Noarbe, J. A. Smith, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, C. To, P. E. Tremblay, M. A. Troxel, D. L. Tucker, D. J. Turner, T. N. Varga, A. R. Walker, R. H. Wechsler, J. Weller, W. Wester, R. D. Wilkinson, B. Yanny, Y. Zhang, R. Nikutta, M. Fitzpatrick, A. Jacques, A. Scott, K. Olsen, L. Huang, D. Herrera, S. Juneau, D. Nidever, B. A. Weaver, C. Adean, V. Correia, M. de Freitas, F. N. Freitas, C. Singulani, G. Vila-Verde, and Linea Science Server, "The Dark Energy

Survey Data Release 2," *Astrophysical Journal, Supplement Series*, vol. 255, no. 2, p. 20, Aug. 2021. [Cited on page 51.]

[64] R. Euclid Collaboration: Scaramella, J. Amiaux, Y. Mellier *et al.*, "Euclid preparation. I. The Euclid Wide Survey," *Astronomy and Astrophysics*, vol. 662, p. A112, Jun. 2022. [Cited on page 51.]

[65] Y. P. Jing and Y. Suto, "Triaxial modeling of halo density profiles with high□resolutionn□body simulations," *The Astrophysical Journal*, vol. 574, no. 2, p. 538–553, Aug. 2002. [Online]. Available: http://dx.doi.org/10.1086/341065 [Cited on page 51.]

[66] R. Scoccimarro, R. K. Sheth, L. Hui, and B. Jain, "How Many Galaxies Fit in a Halo? Constraints on Galaxy Formation Efficiency from Spatial Clustering," *The Astrophysical Journal*, vol. 546, no. 1, pp. 20–34, Jan. 2001.

[67] A. A. Berlind and D. H. Weinberg, "The halo occupation distribution: Toward an empirical determination of the relation between galaxies and mass," *The Astrophysical Journal*, vol. 575, no. 2, p. 587, aug 2002. [Online]. Available: https://doi.org/10.1086/341469 [Cited on page 51.]

[68] A. Vale and J. P. Ostriker, "Linking halo mass to galaxy luminosity," *Monthly Notices of the Royal Astronomical Society*, vol. 353, no. 1, pp. 189–200, Sep. 2004. [Cited on page 51.]

[69] A. Tasitsiomi, A. V. Kravtsov, R. H. Wechsler, and J. R. Primack, "Modeling galaxy-mass correlations in dissipationless simulations," *The Astrophysical Journal*, vol. 614, no. 2, p. 533, oct 2004. [Online]. Available: https://doi.org/10.1086/423784

[70] C. Conroy, R. H. Wechsler, and A. V. Kravtsov, "Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time," *The Astrophysical Journal*, vol. 647, no. 1, pp. 201–214, Aug. 2006. [Cited on page 51.]

[71] J. P. Huchra and M. J. Geller, "Groups of Galaxies. I. Nearby groups," *The Astrophysical Journal*, vol. 257, pp. 423–437, Jun. 1982. [Cited on page 51.]

[72] W. H. Press and M. Davis, "How to identify and weigh virialized clusters of galaxies in a complete redshift catalog," *The Astrophysical Journal*, vol. 259, pp. 449–473, Aug. 1982.

[73] E. Tago, J. Einasto, and E. Saar, “Structure of superclusters and supercluster formation - IV. Spatial distribution of clusters of galaxies in the Coma Supercluster and its large-scale environment.” *Monthly Notices of the Royal Astronomical Society*, vol. 206, pp. 559–587, Feb. 1984. [Cited on page 51.]

[74] P. Fosalba, M. Crocce, E. Gaztañaga, and F. J. Castander, “The MICE grand challenge lightcone simulation - I. Dark matter clustering,” *Monthly Notices of the Royal Astronomical Society*, vol. 448, pp. 2987–3000, Apr. 2015. [Cited on page 51.]

[75] M. Crocce, F. J. Castander, E. Gaztañaga, P. Fosalba, and J. Carretero, “The MICE Grand Challenge lightcone simulation - II. Halo and galaxy catalogues,” *Monthly Notices of the Royal Astronomical Society*, vol. 453, pp. 1513–1530, Oct. 2015.

[76] P. Fosalba, E. Gaztañaga, F. J. Castander, and M. Crocce, “The MICE Grand Challenge light-cone simulation - III. Galaxy lensing mocks from all-sky lensing maps,” *Monthly Notices of the Royal Astronomical Society*, vol. 447, pp. 1319–1332, Feb. 2015. [Cited on page 51.]

[77] J. Carretero, P. Tallada, J. Casals, M. Caubet, F. Castander, L. Blot, A. Alarcón, S. Serrano, P. Fosalba, C. Acosta-Silva, N. Tonello, F. n. Torradeflot, M. Eriksen, C. Neissner, and M. Delfino, “CosmoHub and SciPIC: Massive cosmological data analysis, distribution and generation using a Big Data platform,” in *Proceedings of the European Physical Society Conference on High Energy Physics. 5-12 July*, Jul. 2017, p. 488. [Cited on page 51.]

[78] A. C. Alfano, O. Luongo, and M. Muccino, “Breaking the baryon-dark matter degeneracy in a model-independent way through the sunyaev-zeldovich effect,” 2023. [Online]. Available: https://arxiv.org/abs/2311.05324 [Cited on page 74.]