Universitat Autònoma de Barcelona

Cosmology with photometric redshift

Andrea Pocino Yuste

Universitat Autònoma de Barcelona

Doctoral Thesis

Cosmology with photometric redshift

Andrea Pocino Yuste

Supervisor:
Dr. Francisco Javier Castander Serentill

Tutor:
Dr. Jordi Mompart Penina

Doctor of Philosophy in Physics

Departament de Física
Facultat de Ciències
Universitat Autònoma de Barcelona

November 18, 2020

Abstract

Current and future photometric surveys will observe a large volume of the universe that will allow us to accurately constrain the cosmological model. However, the constraining power from cosmological probes of photometric surveys highly relies on the accuracy and precision with which we can determine the galaxies redshifts. Therefore, the determination of photometric redshifts (photo-zs) and their effect in cosmological analysis should be treated and studied carefully.

In the first part of this thesis, we transform the photometry of existing simulations to mimic the photometric measurements of the Dark Energy Survey (DES). With this exercise, we expect to recover the real photo-z distribution in simulations, thus creating a more realistic environment to crosscheck the performance of DES in cosmological analyses that use photo-z. We transform the simulations using several method to transfer the statistical properties from the real observations photometry to the simulations.

In the second part of the thesis, we use the Self-Organizing Map technique to select spectroscopic targets for the C3R2 program aimed at establishing the mapping between color and redshift space. We also explore the color space defined by the photometry of galaxies from the Physics of the Accelerating Universe Survey (PAUS) in order to study the spectroscopic redshift coverage of its color space. We want to quantify the regions of color space without spectroscopic redshifts because the lack of spectroscopic representation can be a source of bias when the accuracy of photo-zs is evaluated by comparing it to spectroscopic redshifts and when the spectroscopic redshifts are used to determine the photo-z with training-based algorithms.

Lastly, we explore how the variation of the depth of ground-based observations combined with Euclid observations affects the accuracy and precision of the photo-z and thus the cosmological constraining power of Euclid focusing on photometric galaxy clustering and galaxy-galaxy lensing analyses. We also study how the number density of photometric galaxy samples affects the constraining power and which tomographic redshift binning configuration returns the maximum information to constrain the cosmological parameters. To perform such analyses, we create several realistic photo-z distributions based on the Euclid Flagship simulation and we use the Fisher forecast and the cosmological inference code, CosmoSIS, over the different configurations of the galaxy samples to determine the cosmological constraining power.

Acknowledgements

M'agradaria començar donant les gràcies a en Francisco per haver fet possible l'experiència que han sigut aquests darrers anys. Per guiar-me ja des de la carrera i aguantar-me fins al final del doctorat. Per totes les hores de discussions i aprenentatge. Per l'esprint d'aquests darrers dies. I sobretot per animar-me a veure món, a tenir més confiança i a millorar. També vull agrair a l'Isaac haver-me donat l'opció de treballar plegats. Per tot el temps dedicat a resoldre dubtes i barallar-nos amb el programa. Gràcies per guiar-me i explicar les coses tan bé, he après un munt.

He de donar les gràcies a la gent del PIC, sobretot a en Jorge i especialment a en Pau. Gràcies per les mil i una hores invertides en el codi, en trobar errades, en trencar-nos el cap i per aguantar les llargues sessions de discussions. També vull donar les gràcies a la gent del grup de cosmologia amb qui he compartit converses durant aquests anys sobretot a en Martin, en Pablo, l'Enrique, en Ricard, en Santi, l'Arnau, la Linda, l'Ismael... Especials gràcies a en Ricard per l'experiència a La Palma. Per ensenyar-me com anava tot, per portar-me a veure els paisatges del voltant i per la paciència en la primera pujada... A la Gemma i la Mar per compartir les seves experiències, escoltar i ser un referent. A en Martin, la Gosia, la Giulia i especialment la Laura per ser uns companys de meetings ben divertits.

Gràcies als companys de doctorat que han fet d'aquesta una experiència més bona, a la Clara, la Flavia, la Safoura, a en Fran i en Juan Pedro per ser uns companys de despatx genials, a l'Ivan i a en Cristian per poder parlar de totes les aficions, a en Pablo de qui trobaré a faltar les estones de croquetes i xerrades. Vull donar moltíssimes gràcies a l'Anna, la Marina, en David i la Mariona, ha sigut brutal i una sort compartir aquests anys amb vosaltres, gràcies per fer millor el dia a dia, pel vostre suport, per totes les converses, per compartir gats, memes, gifs i riures. També agrair a en Daisuke per fer forats per veure'ns quan ve de lluny. A la Clàudia pels bubbles, les converses i per estar sempre allà. A la Nuria i la Marta pels moments compartits al llarg de la vida i haver estat al meu costat perquè seguís endavant.

Finalment, vull donar les gràcies a la meva família pels ànims, els bons àpats, les estones de joc i la companyia durant tots aquests anys. I especialment, a la persona que ha fet possible aquesta tesi, l'Ángel. Per tots aquests anys d'inestimable suport, per les passejades curtes i llargues, per tots els àpats cuinats amb afecte, per alleujar-me el dia a dia de mil maneres diferents, per escoltar-me i animar-me, gràcies.

Contents

Introduction

Cosmology aims to understands the mechanisms that lie behind the evolution of the universe. The evolution depends on the components of the universe that are described in the cosmological model, which can be characterized by the cosmological parameters. These parameters are determined from the analyses of cosmological probes such as weak gravitational lensing, galaxy clustering, and their cross-correlations. To perform robust cosmological analyses and thus to reliably constrain the cosmological parameters, a large fraction of the universe should be surveyed. Cosmological surveys observe galaxies in large volumes and retrieve physical properties such as the position and redshift, which are needed to trace the underlying large-scale structure and thus to extract cosmological information from observations. Cosmological surveys are classified into spectroscopic and photometric surveys depending on whether the redshifts of galaxies are estimated with spectroscopy or using photometric techniques. Photometric surveys observe galaxies with multi-band filters instead of sampling the full energy spectral distribution with higher resolution as in spectroscopy, which requires more observational time. That way, photometric surveys can trace many more objects than spectroscopic surveys but at the expense of a degraded precision on the redshift estimates. Current photometric surveys such as the Dark Energy Survey (DES) and the Kilo-Degree Survey (KiDS), and the upcoming Stage-IV dark energy surveys such as Euclid, the Vera C. Rubin - Legacy Survey of Space and Time (Rubin–LSST), and the Nancy Grace Roman Space Telescope (formerly known as WFIRST), will trace an unprecedented large volume that will allow us to constrain with precision the cosmological model. The current standard cosmological model is the ΛCDM whose main components are dark energy and dark matter. However, the nature of theses components is still unknown. The precision of the aforementioned cosmological surveys will also allow us to better understand their nature. In order to extract the maximum information from the large amount of data coming from the cosmological surveys, efficient algorithms are needed to determine the photometric redshift in a fast and precise way.

Photometric redshifts are one of the main ingredients in cosmological analyses.

Their accurate determination is a crucial step to perform accurate science. There are many techniques to determine them, from the template fitting method to several machine-learning algorithms such as decision tree classification, random forests, neural networks, and Gaussian process regression. In order to explore, test and validate the various methods to determine photometric redshifts and study their impact in cosmological analyses, we need simulations. Therefore, simulations need to properly reproduce the observed properties of galaxies in order to be useful for photometric redshift analyses and to study the impact of photometric redshift in the cosmological analyses. However, in practice, differences in observable properties arise between simulations and real observations due to the complexity of data and the limitations and simplifications of simulations. In the first project of this thesis, we aim to transform the observables from existing simulations to model real observables. We want that these observables allow us to reproduce photometric redshift distributions as the ones found in real data. Therefore, we focus on reproducing real photometry distributions in simulations. We do that by transferring statistical properties of the photometry. We explore several methods to perform these transformations. In this thesis, we focus to faithfully reproduce the photometry of DES in order to recover realistic photometric redshift in simulations and thus to be able to crosscheck ongoing and future cosmological analyses of DES that involve photometric redshifts.

As we mentioned, the performance of the cosmological analyses depends on the accuracy and precision of the photometric redshifts used. The most usual way to determine photometric redshifts is with machine learning techniques. These types of algorithms learn from a sample with known redshift and establish a mapping function between the fluxes and redshift that is used to determine the photometric redshift. The level of photometric redshift accuracy required to fulfill the scientific goals of upcoming photometric surveys can only be achieved using spectroscopic samples to calibrate the $n(z)$ that are fully representative of the entire range of colors and redshifts of the photometric sample. The Complete Calibration of the Color-Redshift Relation survey (C3R2) is an ongoing spectroscopic effort aimed to identify and observe the galaxies that are needed to have a fully representative spectroscopic sample for the Euclid survey. To define a representative spectroscopic sample, the galaxy color–redshift relation is empirically calibrated using the machine learning algorithm called Self-Organizing Map (SOM). This method allows to detect which regions of the galaxy color space are not represented in the currently available spectroscopic sample.

In the second project of this thesis, we use the SOM to identify and target galaxies that are needed to complete the color-redshift space as part of the C3R2 effort. Knowing the color-redshift coverage is useful to detect the regions without spectroscopic counterpart, which can be a source of bias when comparing the spectroscopic redshift to the photometric one when assessing the accuracy of the latter. Therefore, we also use the SOM technique to explore the color space of the photometry of galaxies from the Physics of the Accelerating Universe (PAUS) Survey, and to study the color-redshift coverage of galaxies in the survey.

The precision and accuracy of photometric redshift particularly affect the results of galaxy clustering with photometrically-selected galaxies and weak lensing analyses. Given the instrumental specification of Euclid, to meet its science requirements its data should be complemented with ground-based observations to derive precise and accurate photometric redshifts. In the third project of this thesis, we explore how the depth of ground-based observations varies the accuracy and precision of the photometric redshifts, and thus how it affects the cosmological constraints that will be derived by the Euclid mission. We also explore which is the optimum tomographic redshift binning configuration to extract the maximum information to constrain the cosmological parameters. We focus our study on the impact on photometric galaxy clustering and galaxy-galaxy lensing analyses. For that purpose, we simulate several realistic photometric redshift distributions based on the Euclid Consortium Flagship simulation and determine the photometric redshift using a machine learning algorithm called Directional Neighborhood Fitting. We apply the Fisher matrix formalism over the simulated photometric galaxy samples to study the cosmological constraining power as a function of redshift binning, survey depth, and photometric redshift accuracy.

This thesis is organized as follows. In chapter 1, we describe the cosmological framework with basic cosmological terms to help to understand concepts that will appear along the thesis. In chapter 2, we review the DES and PAUS photometric surveys, their observational data that will be used in this thesis, and the upcoming Euclid space mission. In chapter 3, we aim to transform the statistical properties of the photometry of simulations to model the photometry of DES. We describe the methods used to achieve that, how we determine photometric redshifts from the transformed photometry, and the results on using the transformed photometry and determined redshift when applying the same galaxy sample selection as in DES. In

chapter 4, we use the SOM to identify and target useful galaxies to fill the color-redshift space defined by the C3R2 project in order to get a fully representative spectroscopic sample. We explain the photometric calibrations needed to correctly place galaxies into the SOM. We also define a SOM technique to explore the color-redshift space of PAUS in order to detect possible sources of biases in the comparison of spectroscopic redshift to the photometric one. In chapter 5, we optimize the photometric galaxy sample of Euclid for galaxy clustering and galaxy-galaxy analyses. We explain how we create the different photometric samples, determine their photometric redshifts, the optimization of the tomographic redshift bin configuration, and we study the dependency of the cosmological constraints on photometric redshift quality and sample size. Finally, we summarize and conclude in the last chapter.

Chapter 1

Cosmological Framework

In this introductory chapter, we briefly review the discoveries and theories of modern cosmology (Sec. 1.1) that lead to the standard cosmological model (described in Sec. 1.2). In Secs. 1.3 – 1.5, we describe some basic cosmological concepts that will help to better understand the thesis, while the key concepts of this thesis are explained in Secs. 1.6 and 1.7.

1.1 Modern Cosmology

Cosmology is the science that studies the universe. It aims to understand its origin and evolution through the observation and analysis of large scale structure. Modern cosmology is based on the gravitational model defined by Albert Einstein. In 1905, Einstein proposed the theory of special relativity (Einstein, 1905) that was extrapolated to include a relativistic description of gravity. This extended theory was called general relativity (Einstein, 1915). Two year later, he applied this theory to cosmology (Einstein, 1917) laying the foundations of modern cosmology. For a thorough historical review we refer the reader to Lima, and Santos (2017).

In the application of the general relativity to cosmology, Einstein proposed a model of the universe where the *cosmological principle* was assumed. This principle states that the universe in the large scale average is homogeneous and isotropic. The model of Einstein also assumed the historical principle in astronomy that the universe was static. However, the equations of the model of Einstein went against this assumption. So to make the equations compatible with an static universe, he had to introduce a term in his equations, known as the *cosmological constant* Λ.

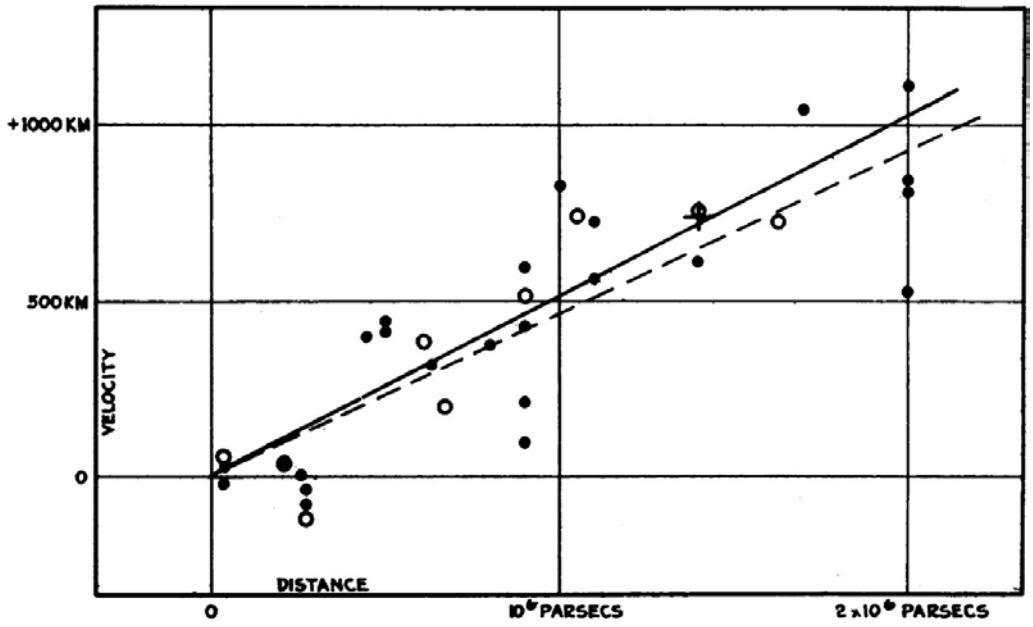

FIGURE 1.1: Original Hubble diagram from Hubble (1929) showing the velocity as a function of distance for observed "extra-galactic nebulae", former name for galaxies.

1.1.1 The Hubble law, scale factor and redshift

In 1929, Edwin Hubble found evidences that the universe was expanding. He determined a linear relationship between velocities and distances by observing Cepheids as standard candles. The original plot from Hubble showing the velocities and distances is shown in Fig. 1.1. He found out that galaxies were moving away. The further they were, the faster they moved away.

To determine the velocities and distances, Hubble measured what is called the *redshift* of a galaxy. Since the universe is expanding and the speed of light is finite, there is a shift between the observed wavelength λ_o and the wavelength emitted by a galaxy λ_e in the rest-frame. The ratio between the difference of wavelengths and the emitted wavelength is known as redshift:

$$z \equiv \frac{\lambda_o - \lambda_e}{\lambda_e} \,. \tag{1.1}$$

Redshift can be related to the recessional velocity v through the Doppler effect using an approximation for small velocities $z \simeq v/c$. Through redshift, Hubble established the linear relation between velocities and distances known as the Hubble law:

$$v \simeq cz = H_0 d \tag{1.2}$$

where d is the *proper distance* and H_0 is the Hubble constant which corresponds to the value of the Hubble rate H at present time. H_0 is parametrized by the dimensionless Hubble parameter h as $H_0 = 100\,h\,\mathrm{km\ s^{-1}Mpc^{-1}}$.

Due to the expansion of the universe, distances between objects change over time. The distance between two objects is called the physical or proper distance $\vec{d}$ and is related to the *comoving distance* $\vec{\chi}$ by the *scale factor* $a(t)$ as:

$$\vec{d} = a(t)\mathrm{d}\vec{\chi}\,, \tag{1.3}$$

where t is the cosmic time. The comoving distance is the distance that remains constant due to the expansion of the universe. The scale factor describes the expansion of the universe and thus contains the information of changes due to the expansion. The scale factor is a dimensionless parameter that for convenience it is usually defined as $a_0 \equiv a(t_0) = 1$ and was smaller in the past. Where the subscript 0 denotes present time. So at present time the proper and comoving distances are equal.

The Hubble rate can be derived from the Hubble law using the definition of velocity as the variation of distance over time $v = \mathrm{d}d/\mathrm{d}t \equiv \dot{d}$, where overdot denotes a time derivative. Considering the definition of proper distance (1.3), velocity can be written as $v = \dot{d} = \dot{a}\chi + a\dot{\chi} = \dot{a}d/a$ since at present time $\dot{\chi} = 0$. Then the *Hubble rate* is defined as:

$$H(t) \equiv \frac{\dot{a}}{a} \tag{1.4}$$

which quantifies the change in the scale factor over time.

If we consider the wavelengths of photons moving through the universe, the relation between redshift and the scale factor can be determined using the Friedmann–Lemaître–Robertson–Walker (FLRW) metric, that we will introduce later. Light travels on null geodesics, which means that the line element of light is $\mathrm{d}s = 0$. Using symmetry $\mathrm{d}\theta = \mathrm{d}\phi = 0$, the FLRW metric (1.9) becomes:

$$\frac{\mathrm{d}t}{a(t)} = \frac{\mathrm{d}r}{c\sqrt{1-kr^2}}\,. \tag{1.5}$$

Consider two successive wave crests of a monochromatic wave emitted at times t_e and $t_e + \Delta t_e$, respectively, from a position r. Position r will be the same for both crests since the position is comovil. The crests will arrive to the observer at time t_0 and $t_0 + \Delta t_0$ and position $r = 0$. Then the right side of the above equation will remain

the same and we will have:

$$\int_{t_e}^{t_0} \frac{\mathrm{d}t}{a(t)} = \int_{t_e+\Delta t_e}^{t_0+\Delta t_0} \frac{\mathrm{d}t}{a(t)} \,. \tag{1.6}$$

If an integral with interval between $t_e + \Delta t_e$ and t_0 is subtracted in both sides of the equation, we assume the scale factor is constant for small variations in time, $\Delta t_e = \lambda_e$ and $\Delta t_0 = \lambda_0$, the solved integral can be written as:

$$\frac{\Delta t_0}{\Delta t_e} = \frac{\lambda_0}{\lambda_e} = \frac{a_0}{a(t_e)} = \frac{1}{a(t)} \,. \tag{1.7}$$

Using the definition of redshift (1.1), it can be related to the scale factor as:

$$1 + z = \frac{1}{a(t)} \,. \tag{1.8}$$

This relation was firstly deduced by Georges Lemaître in 1927.

1.1.2 The Friedmann–Lemaître–Robertson–Walker metric

During the 1920s, Alexander Friedmann found an exact and general solution of the field equations of general relativity of Einstein based on a homogeneous and isotropic universe. This solution was known as the *Friedmann–Lemaître–Robertson–Walker* (FLRW) metric and it is the standard one used in cosmology:

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega\right] \tag{1.9}$$

where c is the speed of light, $\mathrm{d}\Omega = \mathrm{d}\theta^2 + sin^2\theta\mathrm{d}\phi^2$ is the solid angle in spherical comoving coordinates and r is the radial comoving distance. The solution is also characterized by the curvature k of the universe that can be:

$$k \begin{cases} > 0 & \text{closed (spherical universe)} \\ = 0 & \text{flat (Euclidean universe)} \\ < 0 & \text{open (hyperbolic universe)} \end{cases} \,. \tag{1.10}$$

Among the possible solutions of the FLRW metric, Friedmann predicted solutions that modeled an expanding universe (Friedmann, 1922; Friedmann, 1924). The expansionist solutions lead to the creation of the Big-Bang theory .

1.1.3 The Big-Bang cosmological model

Although the model was first proposed by Friedmann and Lemaître during the 1920s, the modern model was developed during the 1940s by George Gamow (Gamow, 1946) and his collaborators, Ralph Alpher and Robert Herman (Alpher, and Herman, 1948). The Big-Bang model assumes that the gravitational interactions of matter are described by general relativity and that the universe follows the cosmological principle. The theory states that the universe expands from a primordial state with extreme conditions of high energy density, temperature and pressure.

In the early stages of the universe, the universe was formed by a plasma dominated by relativistic particles and therefore by radiation. As the universe kept expanding, the temperature and density decreased, the velocities of the relativistic particles went down and their energy-density started to be dominated by their mass. Then the behavior of the universe was increasingly dominated by matter until it arrived at an epoch equally dominated by radiation and matter at redshift $z \approx 3600$, and continued dominated by matter instead of radiation. At redshift $z \approx 1100$, charged electrons and protons formed neutral hydrogen atoms, which is known as recombination. During the formation of hydrogen atoms, electrons emitted photons when transitioning to lower energy states. At the same time, photons decoupled from matter. Due to the decrease of free electrons and protons and the decoupling of photons, the mean free path of photons increased and they were able to travel more freely, thus the universe become transparent. This radiation continued through space and is known as cosmic microwave background. At redshift $z \approx 20$, matter started to collapse due to gravity forming the firsts stars and galaxies, and early large structures emerged. Galaxy clusters started to appear at approximately redshift $z \approx 2$. From redshift $z \approx 0.4$ until the present time, the universe started to accelerate and we do not know the cause. So we call dark energy to the dominant contribution of energy density of the universe that we assume that causes the acceleration.

1.1.4 Cosmic microwave background

The recession of galaxies discovered by Hubble was the first observational evidence of the expansion of the universe and the Big-Bang cosmological model. The detection of the cosmic microwave background (CMB) by Arno Penzias and Robert Wilson (Penzias, and Wilson, 1965) provided another evidence to support the Big-Bang theory.

Since the decoupling of the CMB photons, their temperature gradually diminished due to the expansion of the universe as a function of the scale factor. The determination of the CMB temperature nowadays helps to assess the temperature and time of the universe when the photons of the CMB were decoupled. The Wilkinson Microwave Anisotropy Probe (WMAP; 2001) satellite and the Planck (2009) determined that the temperature of the CMB was about 2.726 K.

Although the CMB radiation looks almost the same in all directions, WMAP and Planck observations showed small fluctuations or anisotropies in the CMB temperature of the order of 10^{-5} in different places of the sky. So the CMB is not perfectly isotropic. These fluctuations are caused by density fluctuations in the early universe, which eventually generated the large scale structures. Therefore mapping the temperature of the CMB gives information about the initial conditions that formed large structures, which can help to constrain basic parameters of the cosmological model.

1.1.5 Dark matter

Several observational evidences have suggested that there is more mass in the universe than what can be observed. This non-visible matter that can only be detected through its gravitational effects is called *dark matter*.

One of the first evidences of the existence of dark matter was found in the velocity dispersion of galaxies in clusters. The virial theorem relates the kinetic and potential energy of a system, so by measuring the velocities of galaxies gravitationally bound to a cluster, the gravitational mass of the cluster can be inferred. Fritz Zwicky (Zwicky, 1933) was the first one to use the virial theorem to infer the gravitational mass of a cluster. He compared the inferred dynamical mass with the observed luminous mass and realized

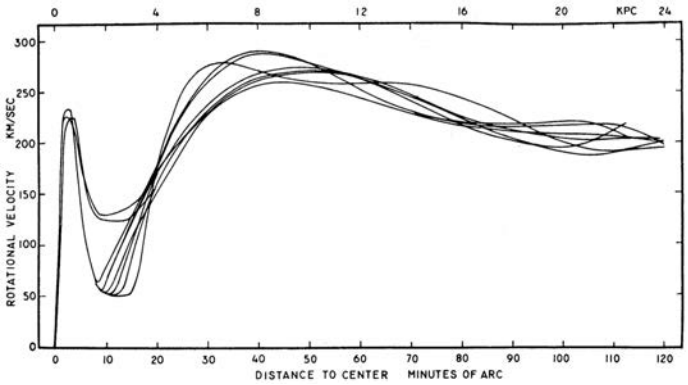

FIGURE 1.2: Original plot from Rubin, and Ford (1970) showing the fits over the rotation curves of several emission points from galaxy M31. The fits were computed with different methodologies.

that the dynamical mass was larger than the luminous one hinting the existence of non-observable matter.

At the beginning of the 1970s, Keneth Freeman (Freeman, 1970), Vera Rubin and Kent Ford (Rubin, and Ford, 1970) confirmed the existence of dark matter. They discovered that the rotational velocities v_c of spiral galaxies did not decrease as a function of the distance to the center of the galaxy r, as one would expect from the observed luminous matter. Remember that the mass distribution $M(r)$ of a galaxy is related to the rotational velocity as $v_c = \sqrt{GM(r)/r}$ where G is the gravitational constant. Instead, the rotational velocity with distance was approximately constant as shown in Fig. 1.2. The only way for that to happen is that matter without luminosity existed in the edges of the galaxies.

Another evidence of dark matter is found in the CMB. The fluctuations in the temperature of the CMB radiation are the result of oscillations of matter and radiation in dense regions caused by the competition between the gravitational force and the pressure of radiation in the time the CMB was emitted. So these fluctuations are a map of the matter density distribution. The measured anisotropies of the CMB can be decomposed into the angular power spectrum. Matter leaves a series of acoustic peaks in the power spectrum that are related to the density of baryonic and dark matter. The WMAP satellite, in 2001, and the Planck mission, in 2009, provided measurements of the peaks that were compatible with the existence of dark matter. The power spectrum of the CMB measured by the Planck mission showing the intensity of fluctuations across different scales can be seen in Fig. 1.3.

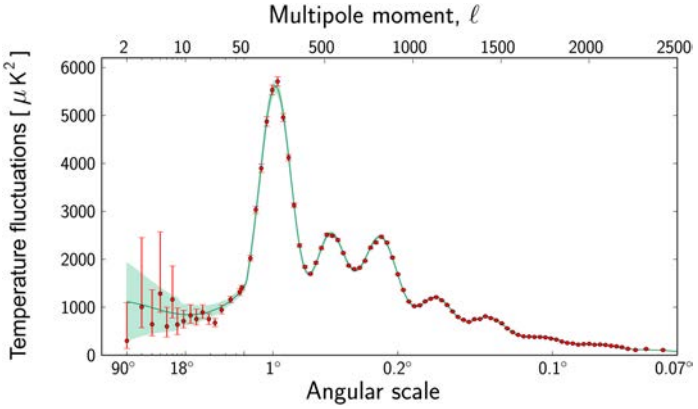

FIGURE 1.3: Power spectrum of temperature fluctuations in the CMB detected by Planck at different angular scales on the sky. *Credit:* ESA and the Planck Collaboration.

There are more cosmological probes that hint to the existence of dark matter such as the use of type Ia supernovae or the baryonic acoustic oscillations, which are sensitive to the components of the universe and so to dark matter, and the weak gravitational lensing (see 1.6.1). Over the past decades, a lot of effort has been put to

map the dark matter of the universe. Current and future generation of cosmological surveys will help to better understand the nature of dark matter.

1.1.6 Accelerated universe

In 1998, the teams of Adam Riess (Riess et al., 1998) and Saul Perlmutter (Perlmutter et al., 1999) measured the distances to type Ia supernovae (SNe Ia) at high redshift and found they were farther away than expected for a predicted matter only dominated universe, discovering that not only the universe was expanding but it was accelerating. This caused a change of paradigm in cosmology. Identifying the cause and origin of the acceleration is one of the fundamental questions in cosmology nowadays.

The required existence of dark energy

The acceleration of the universe disagrees with a universe that is only matter dominated. If we look at the equation that describes the dynamics of the universe as a function of its components 1.12, (which will be presented in detail in the next section), we see that for the universe to expand, the acceleration of the scale factor $\ddot{a}$ should be positive, and that requires of a component with negative pressure that satisfies $p < -\rho/3$. This component with negative pressure and responsible of the acceleration of the universe is called *dark energy*. Nowadays, the cosmological constant is considered to be tied to the dark energy.

Given that dark energy has not been detected yet nor it has been fully characterized, the origin of the acceleration is still uncertain. For the past two decades, theoretical and observational cosmology has made a huge effort to model the dark energy and to observe its physical effects in order to understand how the acceleration of the universe works.

1.2 The standard cosmological model

The Lambda Cold Dark Matter (ΛCDM) universe is currently considered the standard cosmological model. This model assumes the Big Bang as the origin of the universe, the cosmological principle, the flatness of the universe and an accelerated expansion. It is described by the FLRW metric and its dynamics follows the Friedmann equations. The model considers that the universe is composed of radiation

(photons and relativistic particles), dark energy that behaves as the cosmological constant i.e. as the vacuum energy, and matter as the sum of baryonic matter (protons and neutrons) and cold dark matter (CDM). The cold dark matter is a hypothetical matter that is non-baryonic, with velocities slower than the speed of light (providing the adjective cold), and only interacts through gravity.

The standard cosmological model is the result of trying to describe and justify several observational cosmological effects such as the ones presented along this section. The ΛCDM model can be described by six parameters: the physical density parameters of baryonic $\Omega_{\rm b}h^2$ and cold dark matter $\Omega_{\rm cdm}h^2$ that indicate the amount of energy density that each components contributes to the universe; the dimensionless Hubble constant $h \equiv H_0/100$ s^{-1}Mpc^{-1}km that indicates the age of the universe; the initial power spectrum amplitude $A_{\rm s}$ and the scalar spectral index $n_{\rm s}$ obtained from the amplitude and slope of fluctuations produced during inflation; and the reionization optical depth τ that quantifies the possible alterations of the CMB anisotropies produced by highly ionized electrons that scatter photons at low redshift. The values of the cosmological parameters define the contents and evolution of the universe, therefore they can be constrained from cosmological probes such as the CMB, SNe Ia, gravitational lensing or galaxy clustering.

1.2.1 Dynamics of the expansion

The standard cosmological model is based on the FLRW solution (1.9) to the field equations of gravitation of Einstein for a homogeneous and isotropic universe. The dynamics of the expansion of the FLRW universe are modeled by the *Friedmann equations* (Friedmann, 1922; Friedmann, 1924) for a perfect fluid. The first one is commonly knows as the *Friedmann equation* per se

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \tag{1.11}$$

and the second equation is also known as the *acceleration equation*

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda c^2}{3} \tag{1.12}$$

where G is the gravitational constant of Newton, k is the spatial curvature, c is the speed of light in vacuum, ρ is the total *energy density* of the universe composed of

the sum of energies of each component, and p is the total *pressure* given by the sum of pressures. These equations describe the evolution of the scale factor $a(t)$ as a function of the total energy and pressure of the universe and allows to derive the Hubble parameter H. The cosmological constant Λ was expressed separately before the discovery of the acceleration of the universe, but can be incorporated to the energy density and pressure with the changes $\rho \to \rho - \Lambda c^2/8\pi G$ and $p \to p + \Lambda c^4/8\pi G$.

A useful third equation can be derived from equations 1.11 and 1.12. This equation is called the *continuity equation*

$$\dot{\rho} = -3H\left(\rho + \frac{p}{c^2}\right) \tag{1.13}$$

and it expresses the evolution of the energy density. In a cosmological model for a universe filled with a perfect fluid, the relation between the energy density and pressure is characterized by the *equation of state*

$$\omega \equiv \frac{p}{\rho c^2} \tag{1.14}$$

where ω is a dimensionless constant. The equation of state for each component of the universe is denoted as $\omega_i = p_i/\rho_i c^2$.

The *critical density* ρ_c of the FLRW metric is defined as the density of a flat universe, thus having zero curvature $k = 0$, without cosmological constant, $\Lambda = 0$. Applying this criteria in equation 1.11, the critical density is

$$\rho_c = \frac{3H^2}{8\pi G}\,. \tag{1.15}$$

The dynamics of the universe is determined by its content, so another useful cosmological parameter to define is the *density parameter* Ω, which is defined as the ratio of the energy density ρ and the critical density

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}\,. \tag{1.16}$$

This parameter for each component is expressed as $\Omega_i = \rho_i/\rho_{i,c}$ where ρ_i is the energy density of the component i. The density parameter is useful to compare different cosmological models. Considering the Friedmann equation (1.11) with the cosmological constant introduced in the total energy density, the equation can be

rewritten as

$$1 - \Omega = -\frac{kc^2}{a^2 H^2} \equiv \Omega_{\mathrm{k}} \,, \tag{1.17}$$

where Ω_{k} is associated to the curvature of the universe. Then the density parameter indicates the spatial geometry of the universe:

$$\begin{cases} k = 1 \text{ (closed)} & \text{if } \Omega > 1 \, (\rho > \rho_{\mathrm{c}}) \\ k = 0 \text{ (flat)} & \text{if } \Omega = 1 \, (\rho = \rho_{\mathrm{c}}) \\ k = -1 \text{ (open)} & \text{if } \Omega < 1 \, (\rho < \rho_{\mathrm{c}}) \end{cases}$$

Since the ΛCDM model assumes a flat universe, the sum of the energy density parameters of all the component of the universe must be equal to 1.

To quantify the acceleration of the expansion in the FLRW universe, the *deceleration parameter* was defined

$$q(t) \equiv -\frac{\ddot{a}(t)a(t)}{\dot{a}^2(t)} = -\frac{\ddot{a}(t)}{a(t)H^2(t)} = \frac{1}{2}\Omega(1 + 3\omega) \tag{1.18}$$

where the acceleration equation, the definition of the critical density, the density parameter and the equation of state have been used in the last equality. A negative value of q happens when the universe is expanding since the accelerating $\ddot{a}$ is positive.

1.2.2 Evolution of the components of the universe

The dynamics of the expansion and the geometry of the universe are determined by the evolution of its components. If the evolution of the energy density of each component over time is known, we can describe the evolution of the universe. The evolution of the total energy density can be derived from the continuity equation (1.13)

$$\frac{\mathrm{d}\rho}{\rho} = -3(1 + \omega)\frac{\mathrm{d}a}{a} \tag{1.19}$$

where the expression of the equation of state (1.14) has been used for convenience because ω is a constant. Integrating this relation gives the variation of the energy density with time as a function of the scale factor

$$\rho = \rho_0 a^{-3(1+\omega)} \rightarrow \rho \propto a^{-3(1+\omega)} \,. \tag{1.20}$$

In addition, solving the Friedmann equation (1.11) for a flat universe gives the evolution of the scale factor with time as a function of the constant of the equation of state

$$a(t) \propto t^{\frac{2}{3(1+\omega)}} . \tag{1.21}$$

To determine the variation of each component, the solution of the continuity equation must be considered assuming that only one component dominates the total energy density of the universe. In the standard cosmological model the universe contains radiation, matter (as the sum of baryonic and cold dark matter) and a cosmological constant. Radiation is assumed to consist on photons and other relativistic particles so it exerts a pressure of $p_{\rm r} = \rho/3$, i.e. $\omega = 1/3$, where the subscript r refers to radiation. Therefore its energy density and the scale factor for a radiation dominated universe evolve as

$$\rho_{\rm r} \propto a^{-4}, \quad a \propto t^{1/2} . \tag{1.22}$$

On the other hand, matter is considered to be formed of non-relativistic particles, so it does not exerts pressure $p_{\rm m} = 0$, i.e. $\omega_{\rm m} = 0$, where the subscript m refers to matter. Then, a matter dominated universe evolves as

$$\rho_{\rm m} \propto a^{-3}, \quad a \propto t^{2/3} . \tag{1.23}$$

Finally, a universe dominated by the cosmological constant is assumed to be mainly driven by the vacuum energy with pressure $p_\Lambda = -\rho_\Lambda$, i.e. $\omega_\Lambda = -1$. Which leads to a constant value of the energy density over time

$$\rho_\Lambda = \text{constant}, \quad a \propto e^{\sqrt{\Lambda/3}t} = e^{Ht} . \tag{1.24}$$

The subscript Λ refers to the cosmological constant. Since the universe is expanding, the cosmological constant represents the dark energy in the ΛCDM model. Although the ΛCDM is consistent with current observations, there are other cosmological models that consider the dark energy as a component instead of the cosmological constant and predict a dynamical equation of state for dark energy. To describe the evolution of the equation of state for dark energy, the parametrization most commonly used is the Chevallier-Polarski-Linder (CPL; Chevallier, and Polarski 2001; Linder 2005):

$$\omega(a) = \omega_0 + (1-a)\omega_a = \omega_0 + \omega_a \frac{z}{1+z} , \tag{1.25}$$

where ω_0 and ω_a are constants. In this cosmological model the equation of state of dark energy is $\omega_{\rm de} = \omega(a) < -1/3$, where the inequality refers to the necessary value for the universe to expand as seen in Sec. 1.1.6. The energy density parameter of dark energy will evolve as a function of the equation of state as $\rho_{\rm de} \propto a^{-3[1+\omega(a)]}$.

Once the evolution of the energy density of each component is known, the total energy density of the universe can be described rewriting the Friedmann equation as

$$H(a) = H_0 \sqrt{\Omega_{\rm r} a^{-4} + \Omega_{\rm m} a^{-3} + \Omega_\Lambda + \Omega_{\rm k} a^{-2}}\,, \tag{1.26}$$

where $\Omega_{\rm k}$ is defined in equation 1.17. Recent observational cosmological probes have constrained the current values of Ω_Λ and $\Omega_{\rm m}$ to be approximately 0.7 and 0.3, respectively, and $\Omega_{\rm r}$ is negligible. Since $\Omega_\Lambda + \Omega_{\rm m} + \Omega_{\rm r} + \Omega_{\rm k} = 1$, then $\Omega_{\rm k}$ is zero. The measured value of the baryonic matter energy density parameter $\Omega_{\rm b}$ is about 0.049, which means that the cold dark matter $\Omega_{\rm cdm}$ greatly contributes to the total matter energy density since $\Omega_{\rm m} = \Omega_{\rm b} + \Omega_{\rm cdm}$.

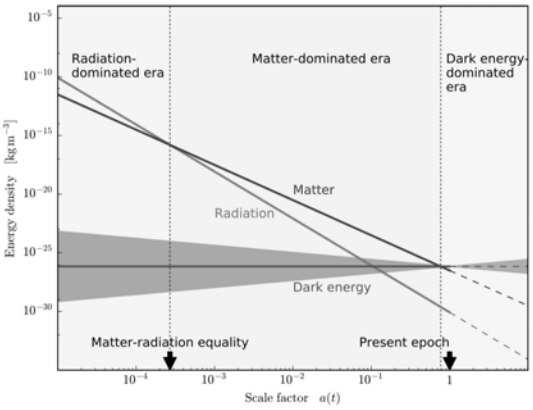

FIGURE 1.4: Energy density as a function of the scale factor for each component of the universe. *Source:* Debono, and Smoot (2016).

Extrapolating back in time the measurements of the energy density of the components nowadays, we see that at the early stages the universe was radiation dominated as shown in Fig. 1.4. After the Big bang, the universe kept expanding and the temperature decreased. The energy-density of relativistic particles started to be dominated by their mass. Then the behavior of the universe was increasingly dominated by matter until it exceed the radiation contribution. Both contributions decrease over time. At times close to the present time the dark energy started to dominate the universe, which also has contributions from matter and a negligible contribution from radiation.

1.3 Distances in the universe

In an expanding universe, distances increase over time. Due to the expansion, light from distant sources gets redshifted when traveling towards the observer since its

speed is limited and it takes time to reach the observer. Determining the distances of objects is not trivial since it is not possible to perform a direct measurement. However, there are observable that allow us to determine distances. In this section we will review the concept of redshift and present the different types of distance that are defined in cosmology.

1.3.1 Redshift

In Sec. 1.1.1, we explained that the expansion of the universe and the finite speed of light cause a shift between the observed wavelength λ_o and the wavelength emitted by a galaxy λ_e in the rest-frame. The difference between these wavelength divided by the emitted one is called redshift z, and it can be associated to the scale factor:

$$z \equiv \frac{\lambda_o - \lambda_e}{\lambda_e} = \frac{1}{a(t)} - 1\,. \tag{1.27}$$

The redshift can be used to determine distances if the Hubble parameter given by the cosmological model is known and vice versa. The methodology to compute redshift from observations will be reviewed in more detail in Sec. 1.7.

1.3.2 Comoving distance

As mentioned in Sec. 1.1.1, the comoving distance $\vec{\chi}$ remains the same with the expansion but can change due to other effects such as peculiar velocities. It is related to the physical distance $\vec{d}$, that increases with the expansion of the universe, through the scale factor (equation 1.3). The comoving distance between an object and an observer at the present time is

$$\chi = \int_{t(a)}^{t_0} \frac{c\mathrm{d}t'}{a(t')} = \int_a^1 \frac{c\mathrm{d}a'}{a'^2 H(a')} = \int_0^z \frac{c\mathrm{d}z'}{H(z')}\,, \tag{1.28}$$

where a change of variable has been applied between steps.

1.3.3 Angular diameter distance

Another distance used in astronomy is the *angular diameter distance*, which is obtained by measuring the angle subtended θ by an object that has a known physical

size l

$$d_{\mathrm{A}} = \frac{l}{\theta}\,. \tag{1.29}$$

This relation only holds for small angles. The comoving distance to the object is given by equation 1.28, while the comoving size is l/a, therefore the subtended angle is $\theta = (l/a)/\chi(a)$. Comparing this with the definition of the angular diameter distance, it becomes

$$d_{\mathrm{A}} = a\chi = \frac{\chi}{1+z}\,. \tag{1.30}$$

This relation is only valid under the assumption of a flat universe. For an open and a closed universe, this relation can be generalized as

$$d_{\mathrm{A}} = aS_k(\chi) \quad \text{where} \quad S_k(\chi) = \begin{cases} \sin\left(\sqrt{-\Omega_k}H_0\chi\right)/(H_0\sqrt{|\Omega_k|}) & \Omega_k < 0 \\ \chi & \Omega_k = 0 \\ \sinh\left(\sqrt{\Omega_k}H_0\chi\right)/(H_0\sqrt{|\Omega_k|}) & \Omega_k > 0 \end{cases} \tag{1.31}$$

where Ω_k is defined in equation 1.17.

1.3.4 Luminosity distance

Another astronomical distance that depends on the cosmological model is the *luminosity distance*. The luminosity distance is obtained by measuring the flux F from an object with known luminosity L such as Cepheids or SNe Ia. The observed flux of a source with known luminosity at a luminosity distance $d_{\mathrm{L}} \equiv \chi/a$ for an expanding universe is given by

$$F = \frac{L_{\mathrm{s}}}{4\pi d_{\mathrm{L}}^2(a)}\,. \tag{1.32}$$

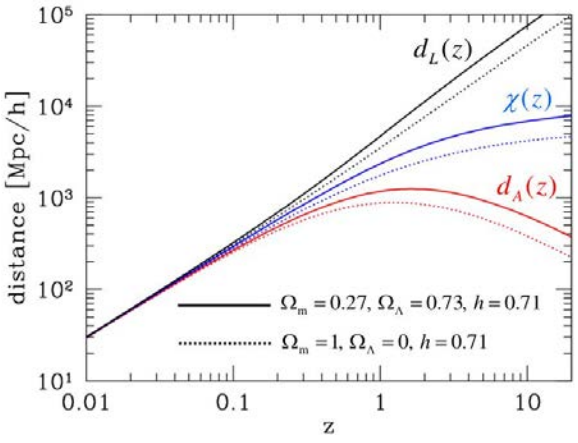

FIGURE 1.5: Luminous, comoving and angular diameter distances as a function of redshift for two different cosmological models. *Source:* Atsushi (2006).

The evolution of the luminosity distance as a function of redshift is shown in Fig. 1.5 in comparison to the comoving and angular diameter distances for two cosmological models. As can be seen, distances strongly depend on the cosmological model. The more the cosmological

constant contributes to the total energy density, the more the distances increase with redshift. The dependency of distances in the cosmological parameters allows to use cosmological probes such as SNe Ia to constrain the cosmological parameters and thus the cosmological model.

1.4 Structure formation

The universe looks approximately homogeneous and isotropic at large scales. But at small scales matter accumulates in denser regions forming galaxies, cluster of galaxies, walls and filaments – referred as large scale structure (LSS). In Fig. 1.6, the map of galaxies observed by the 2dF Galaxy Redshift Survey (2dFGRS) shows the existence of variations in the matter distribution.

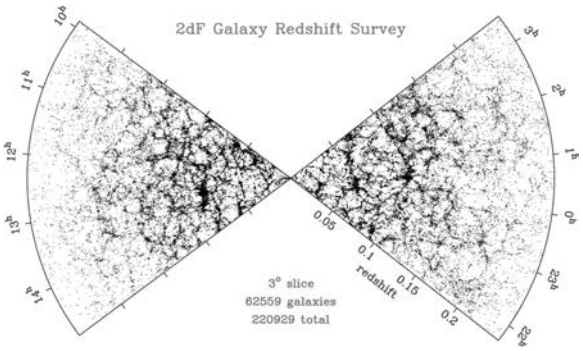

FIGURE 1.6: Distribution of galaxies observed by 2dFGRS. *Credit:* 2dFGRS Team.

It is theorized that the inhomogeneities in the matter distribution originated in the early stages of the universe due to quantum fluctuations, which produced the anisotropies observed in the CMB radiation. The quantum fluctuations were amplified to larger scales by inflation (a theory that assumes an exponential expansion of space in the early stages of the universe during a short period of time) producing small density perturbations that later became a gravitational attraction well that led to the formation of large-scale structures. The formation and evolution of large-scale structures are determined by the cosmology of the universe, so these structures can be used to constrain the cosmological model.

1.4.1 Characterization of the density field

The initial fluctuations, originated during inflation, are assumed to be Gaussian. If a field is Gaussian, it can be fully characterized by its statistical properties, i.e. the mean and the variance of the field. The density perturbation field can be described using the *overdensity field* δ

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}}\,, \tag{1.33}$$

which characterizes the energy density perturbations in a three-dimensional position $\rho(\mathbf{x})$ in comparison to the mean energy density $\bar{\rho}$ of the universe. A positive value of the density field denotes an overdense region, while a negative value indicates an underdense region. Since the universe was homogeneous and isotropic at the moment the initial fluctuations were formed and still is at large scales, the overdensity field is expected to have the same statistical properties, i.e. to have the same symmetries (invariant under rotation and translation). Therefore, considering the whole density field, the average inhomogeneities should be zero $\langle\delta(\mathbf{x})\rangle = 0$.

If the density field is Gaussian, it can be fully characterized by the two point correlation function ξ. Which measures the degree of clustering between two different points of the field, i.e. the probability of finding an excess of density compared to the density one would expect for a random distribution. The two point correlation function is defined as

$$\xi(\mathbf{r}) = \langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle\,. \tag{1.34}$$

The two point correlation function can be computed from data. While its Fourier transform, called power spectrum $P(k)$, is the quantity predicted by theories for the formation of large-scale structure. Since the density field is assumed to be Gaussian, the density fluctuations are completely statistically described by the power spectrum. $P(k)$, which indicates the amount of clustering at different scales, depends on the wavenumber k that is related to the scale or wavelength λ of a fluctuation by $k = 2\pi/\lambda$. Since the power spectrum and the correlation function are the Fourier transforms of each other, they are related as

$$\xi(\mathbf{r}) = \int P(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{r}}\frac{\mathrm{d}^3k}{(2\pi)^3} \quad \text{and} \quad P(\mathbf{k}) = \int \xi(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}\mathrm{d}^3r\,. \tag{1.35}$$

The overdensity field and its conjugate can be expressed in a Fourier series as $\delta(\mathbf{r}) = \int\tilde{\delta}(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{r}}\mathrm{d}^3k$ and $\tilde{\delta}(\mathbf{k}) = \int\delta(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{r}}\mathrm{d}^3r$. Then, assuming an isotropic universe, the correlation function and the power spectrum can be written as

$$\xi(r) = \int_0^\infty \Delta^2(k)j_0(kr)\mathrm{d}(\ln k) \quad \text{and} \quad \Delta^2(k) = \frac{2k^3}{\pi}\int_0^\infty \xi(r)j_0(kr)r^2\mathrm{d}r \tag{1.36}$$

where $\Delta^2(k) \equiv P(k)k^3/(2\pi^2)$ is a commonly used normalization for the power spectrum and $j_0(kr) = \sin(kr)/kr$ is the zeroth order spherical Bessel function.

Since the amplitudes for different wavenumber are orthogonal, the expected value of the overdensity field across all directions in the k-space can also be expressed as

$$\langle\tilde{\delta}(\mathbf{k})\tilde{\delta}(\mathbf{k}')\rangle = (2\pi)^3 P(k)\delta_{\mathrm{D}}^{(3)}(\mathbf{k}+\mathbf{k}') \tag{1.37}$$

where $\delta_{\mathrm{D}}^{(3)}$ is the three-dimensional Dirac delta function.

The variance of the density field is given by

$$\sigma^2 \equiv \langle\delta^2(\mathbf{x})\rangle = \int_0^\infty \Delta^2(k)\mathrm{d}(\ln k)\,. \tag{1.38}$$

Sometimes it is convenient to smooth the matter distribution on a certain scale R by evaluating the density field in a small sphere. For that, three-dimensional spherical top hat filter function W_R is used. The smoothed variance is written as

$$\sigma_R^2 = \int_0^\infty \Delta^2(k)\tilde{W}^2(kR)\mathrm{d}(\ln k)\,. \tag{1.39}$$

Where the filter function is

$$\tilde{W}(kR) = \frac{3}{kR}j_1(kR) = 3\frac{\sin(kR) - kR\cos(kR)}{(kR)^3}\,. \tag{1.40}$$

When the filter is used to smooth the matter density in a sphere of radius $R = 8$ h^{-1}Mpc, the cosmological parameter σ_8 is obtained:

$$\sigma_8^2 = \int_0^\infty \Delta^2(k)\tilde{W}^2(kR_8)\mathrm{d}(\ln k)\,. \tag{1.41}$$

This parameter indicates the density fluctuations on a scale of 8 h^{-1}Mpc, which is the typical scale of massive galaxy clusters.

The two point correlation function and the power spectrum give information of the matter density in the three-dimensional space. However, we do not always have enough observational precision to map the three-dimensional space. Therefore, it is convenient to split the universe in layers – usually in redshift slices – and project the three-dimensional density field into the layers. The equivalent statistics to describe the density perturbations in a two-dimensional space are the two point angular correlation function, $\omega(\theta)$, and its Fourier transform, the angular power spectrum $C(l)$.

To work with the projected density distribution, first we integrate the density of galaxies along the line of sight considering that the distance to a galaxy is given by

its comoving distance χ:

$$\delta(\boldsymbol{\theta}) = \int_0^\infty \delta(\mathbf{x}(\chi, \boldsymbol{\theta}))W(\chi)\mathrm{d}\chi \tag{1.42}$$

where $\mathbf{x}(\chi, \boldsymbol{\theta})$ is the three dimensional position of a galaxy, and $W(\chi)$ is the normalized filter function $\int W(\chi)\mathrm{d}\chi = 1$ that contains the probability of observing a galaxy at a certain distance, i.e inside a redshift slice, $n(z)\mathrm{d}z = n(\chi)\mathrm{d}\chi$. Similar to the three dimensional power spectrum (1.37), the angular power spectrum can be defined as

$$\langle\tilde{\delta}(\boldsymbol{l})\tilde{\delta}(\boldsymbol{l}')\rangle = (2\pi)^3 C(l)\delta_{\mathrm{D}}^{(2)}(\boldsymbol{l} + \boldsymbol{l}') \tag{1.43}$$

where l is the two dimensional wave-number conjugate of the angular separation θ. If the width of the layer defined in the windows function is much larger than the length scale, the Limber approximation can be used and thus the angular power spectrum can be expressed as a function of the three dimensional power spectrum

$$C(l) = \int_0^\infty \frac{W(\chi)}{\chi^2} P(k = \frac{l}{\chi})\mathrm{d}\chi\,. \tag{1.44}$$

To work with fluctuations in a two-dimensional space, the projected field is decomposed into spherical harmonics. The spherical harmonic component of the projected distribution is

$$a_{lm} = \frac{4\pi}{(2\pi)^3}\int \delta(\mathbf{k}, 0) i^l W_{g,l}(k) Y_{l,m}(\hat{\mathbf{k}})\mathrm{d}^3k \tag{1.45}$$

where $Y_{l,m}(\hat{\mathbf{k}})$ is a Laplace spherical harmonic function of degree l and order m and the filter function is given by

$$W_{g,l}(k) = \int b(z)n(z)D(z)j_l(k\chi(z))\mathrm{d}z \tag{1.46}$$

where $b(z)$ is the bias (which will be introduced in Sec. 1.5), $D(z)$ is the growth factor and $j_l(k\chi(z))$ is the spherical Bessel function of order l. The filter function is used to account for effects on the large scales and on the multipole moments and compares the observed power spectrum δ_{g} to the true underlying distribution δ_{m} determined by predictions in a certain position $\mathbf{x}$:

$$\delta_{\mathrm{g}}(\mathbf{x}) = W(\mathbf{x})\delta_{\mathrm{m}}(\mathbf{x})\,. \tag{1.47}$$

Then the angular power spectrum is expressed as

$$C_l^{ij} \equiv \langle a_{lm}^i a_{lm}^{j*} \rangle = \frac{2}{\pi} \int W_{g,l}^i(k) W_{g,l}^j(k) k^2 P(k,0) \mathrm{d}k \,. \tag{1.48}$$

The superscript i and j denote different redshift layers. So the windows function will depend on the bias, growth factor and density of galaxies of each tomographic bin. For large scales (small l) and narrow slices of redshift additional corrections due to redshift space distortions should be used. Sometimes the dimensionless angular power spectrum is expressed as $l(l+1)C_l/(2\pi)$.

1.4.2 Evolution of the density field

Theoretically, the primordial fluctuations, that generated the perturbations leading to gravitational wells that attracted matter and formed the large-scale structure, were originated during inflation. The primordial fluctuations are described by a power spectrum that follows a power law tendency (firstly determined by Harrison (1970), Zeldovich (1972), and Peebles, and Yu (1970)):

$$P_{\mathrm{in}}(k) = A k^{n_{\mathrm{s}}-1} \tag{1.49}$$

where A is the scalar amplitude or normalization and n_s is the scalar spectral index. The initial power spectrum can only be detected at large scales. Observations have measured a spectral index close 1. For example, the Planck Collaboration has found a value of $n_{\mathrm{s}} = 0.9649 \pm 0.0044$ (Planck Collaboration et al., 2020b).

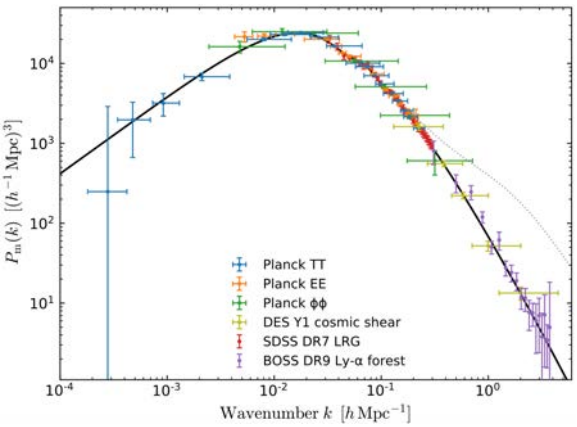

FIGURE 1.7: Power spectrum at $z = 0$ determined from different cosmological probes and surveys. The solid line is the best fit for a linear-theory power spectrum model with a ΛCDM cosmology of $\Omega_\Lambda = 0.6889 \pm 0.0056$ and $\Omega_{\mathrm{m}} = 0.3111 \pm 0.0056$. The dotted line is the impact of non-linear overdensities. *Source:* Planck Collaboration et al. (2020a).

Using linear perturbation theory, it can be determined how the initial power spectrum changes with the evolution of the universe. The change in P_{in} is quantified in terms of the *transfer function* $T(k)$ that relates the initial power to the power

spectrum today

$$P(k) = T^2(k) P_{\rm in}(k)\,. \tag{1.50}$$

The transfer function is constant for scales that enter the Hubble horizon – the boundary between particles with speed slower than the speed of light and particles faster than the speed of light as seen by an observer – before the matter-radiation equality (at a comoving wavenumber $k = 0.01\ h\ \mathrm{Mpc}^{-1}$) and goes as k^{-2} for scales that entered the horizon during the radiation domination epoch. An example of the shape of the power spectrum is shown in Fig. 1.7.

To understand the evolution of large-scale structure, the effect of the gravitational field over the fluctuations of the density field must be quantified. In the ΛCDM cosmological model, matter can be considered an ideal fluid. Then its motion can be described by the continuity, Euler, and Poisson equations. Which are

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{1.51}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}\nabla_r P - \nabla\Phi \tag{1.52}$$

$$\nabla^{\mathbf{2}}\Phi = 4\pi G\rho \tag{1.53}$$

where ρ is the density, $P \ll \rho$ is the pressure, $\mathbf{v}$ is the physical velocity, and Φ is the gravitational potential. The equations describe the variation of these quantities in physical coordinates.

Linear regime

To describe the dynamical evolution of density perturbations of the universe it is convenient to use two regimes: the linear and non-linear perturbation theory. In the former regime, inhomogeneities can be considered as small perturbations ($|\delta| \ll 1$). Changing the dynamical equations from physical to comoving distance, perturbing the variables ($x_i = \bar{x_i} + \delta x_i + \mathcal{O}(\delta^{(2)} x_i) + ...$) and only considering terms with first order in δ_i (to produce analytical equation) the equations become

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla \cdot \mathbf{u} = 0 \tag{1.54}$$

$$\frac{\partial \mathbf{u}}{\partial t} + H\mathbf{u} = -\frac{1}{a\bar{\rho}}\nabla\delta P - \frac{\nabla\varphi}{a} \tag{1.55}$$

$$\nabla^{\mathbf{2}}\varphi = 4\pi G\bar{\rho}^2\delta \tag{1.56}$$

where $\mathbf{v} = \dot{a}\mathbf{x} + a\dot{\mathbf{x}} = \dot{a}H\mathbf{x} + \mathbf{u}$, $P = \bar{P} + \delta P$, $\rho = \bar{\rho} + \delta\rho = \bar{\rho}(1+\delta)$, and the gravitational potential has been split in the perturbation and background potential term $\Phi = \varphi + a\ddot{a}x^2/2$. In the linear regime, the growth of structure is homogeneous.

Taking the time derivative in the linearized form of the continuity equation, applying a divergence to the linearized Euler equation, subtracting both equations, substituting the continuity and Poisson equation, and removing the pressure term since $\nabla P/\rho \ll \nabla\varphi$, the linear perturbation equation is obtained

$$\frac{\partial^2\delta}{\partial t^2} + 2H\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta\,, \tag{1.57}$$

which is a second order differential equation for $\delta(\mathbf{x}, t)$. For a matter dominated universe, the equation of evolution of the overdensity parameter can be written as

$$\ddot{\delta_{\rm m}}(\mathbf{k}, z) + 2H(z)\dot{\delta_{\rm m}}(\mathbf{k}, z) - \frac{3}{2a^3}\Omega_{\rm m,0}H_0^2\delta_{\rm m}(\mathbf{k}, z) = 0 \tag{1.58}$$

where the definition of the critical parameter $\rho_{\rm c}$ (1.15) and the density parameter Ω (1.16) have been applied. Considering a universe with no massive neutrinos, the equation can be rewritten as a function of redshift

$$\delta''_{\rm m}(\mathbf{k}, z) + \left[\frac{H'(z)}{H(z)} - \frac{1}{1+z}\right]\delta'_{\rm m}(\mathbf{k}, z) - \frac{3}{2}\frac{\Omega_{\rm m}(z)}{(1+z)^2}\delta_{\rm m}(\mathbf{k}, z) = 0 \tag{1.59}$$

where prime indicates the derivative with respect to the redshift and the Friedmann equation for the total energy component has been used (1.26). The linear solution for the evolution of overdensitites can be defined in terms of the growth factor $D(z)$:

$$\delta_{\rm m}(\mathbf{k}, z) = \delta_{\rm m}(\mathbf{k}, z_i)\frac{D(z)}{D(z_i)} \tag{1.60}$$

where z_i is a reference redshift in a matter dominated era. The growth factor depends on the components of the universe and indicates the evolution of density growth for different redshift. So the linear perturbation equation can be written in terms of the growth factor:

$$\ddot{D} + 2H(z)\dot{D} - \frac{3}{2}\Omega_{\rm m,0}H_0^2(1+z)^3 D = 0\,. \tag{1.61}$$

This equation can also be written as a function of the growth rate parameter that is defined as:

$$f(\Omega_{\rm m}) \equiv \frac{1}{H}\frac{\dot{D}}{D} = \frac{{\rm d}\ln D(a)}{{\rm d}\ln a} = -\frac{{\rm d}\ln D(z)}{{\rm d}\ln(1+z)}\,. \tag{1.62}$$

Non-linear regime

The transition to non-linear growth of structure occurs when $\delta \sim 1$, where the dynamical equations do not have an analytical solution and require of numerical approaches such as the Zeldovich approximation (Zeldovich, 1970). f=For the non-linear regime, it is even more difficult to obtain a solution. The property of Gaussianity of fluctuations remained in the linear regime, but the fluctuation become non-Gaussian in the non-linear regime due to the non-linear structure formation and the break of homogeneity. The non-linear structure formation ($\delta > 1$) requires of higher order elements in the perturbation decomposition to describe it and thus the dynamical equations need of N-body simulations to fit the formulae for the non-linear power spectrum.

1.5 Galaxy Bias

A precise knowledge of the distribution of matter in the universe is essential to determine the cosmological model and history of the universe. Galaxies are observable tracers of the underlying matter distribution. However, effects such as the bending of light due to gravitational attraction suggest that there is more matter than what can be observed. Then galaxies are a biased tracer of the distribution of matter. The simplest model used to account for the bias is called *linear bias*

$$\delta_{\rm g}(\mathbf{x}) = b\delta_{\rm m}(\mathbf{x})\,. \tag{1.63}$$

The bias b is a constant that relates the matter distribution of the universe $\delta_{\rm m}$ with the galaxy distribution $\delta_{\rm g}$ in a certain position $\mathbf{x}$. Considering the relation between the power spectrum and the overdensity parameter (1.37), the relation between the power spectrum of galaxies and matter become

$$P_g = b^2(k)P_{\rm m}(k)\,. \tag{1.64}$$

The linear bias parameter model is accurate enough for large scales, but in general it is too simplistic because different type of galaxies cluster in different ways depending on the scale and redshift. To parameterize the bias more accurately, more complex bias models can be used with additional nuisance parameters.

1.6 Cosmological probes

Cosmology aims to extract the maximum information from observations to constrain the cosmological parameters as much as possible. The process of conversion from observables to parameters usually require prior assumptions on the cosmological model. Different observable effects have different sensitivities to the cosmological parameters. In this section, we will introduce the cosmological probes related to perturbations.

1.6.1 Weak gravitational lensing

The bending of light due to the gravitational attraction field created by matter is called *gravitational lensing*. Massive objects exert a gravitational field proportional to their mass that warps the space-time. Light travels in geodesics paths, which is the path with the locally minimum distance between two points. In the vacuum, light travels in a straight line but if the light passes by a warped space-time it gets bended due to a gravitational field.

The amount of deflection of light depends on the mass that creates the gravitational field. The distortion produced by small objects is difficult to detect, while the lensing created by large scale structures such as galaxies and clusters is easier to observe. The distant galaxies whose light gets distorted are known as source galaxies and the objects that produce the distortion are known as lenses. Depending on the amount of mass, light coming from background galaxies gets distorted producing different observable effects. For example, light from a source galaxy can be stretched and form an arc around the lensing cluster, the image of a distant galaxy can be duplicated and displaced in another position, or the shape of the galaxies can be distorted. Some of these effects are shown in Fig, 1.8.

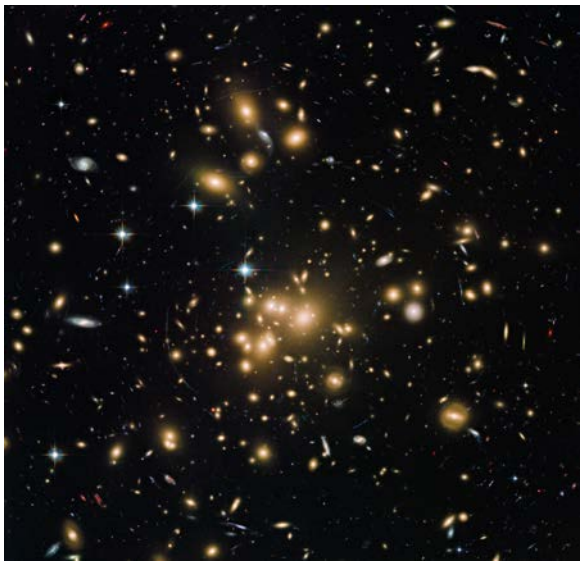

FIGURE 1.8: Image of the Hubble Space Telescope showing the galaxy cluster Abell 1689. The cluster lenses the background galaxies distorting their images. *Credit:* NASA, ESA, STScI.

There are two regimes of gravitational lensing: strong and weak. The former takes places when the lensing effects are strong enough to be visually detected on an astronomical image. Strong lensing is produced by highly overdense regions, such as clusters of galaxies, caused by non-linear perturbations. This regime is able to produce multiple images of background objects and heavily distort the shape of distant objects. On the other hand, in the weak lensing regime shapes and sizes are typically modified by the order of only 1%. In this regime, the lensing is the result of the bending created by all the distribution of matter along the light path instead of a heavy overdense region. Since the alterations of shapes are almost imperceptible, the properties of the density field must be characterize using a statistical approach.

Weak gravitational lensing can be used to trace the underlying matter distribution even if the matter is not visible by relating the distortions of images to the mass power spectrum of the universe. So it is a useful cosmological probe to constrain the cosmological parameters and probe the existence of dark matter.

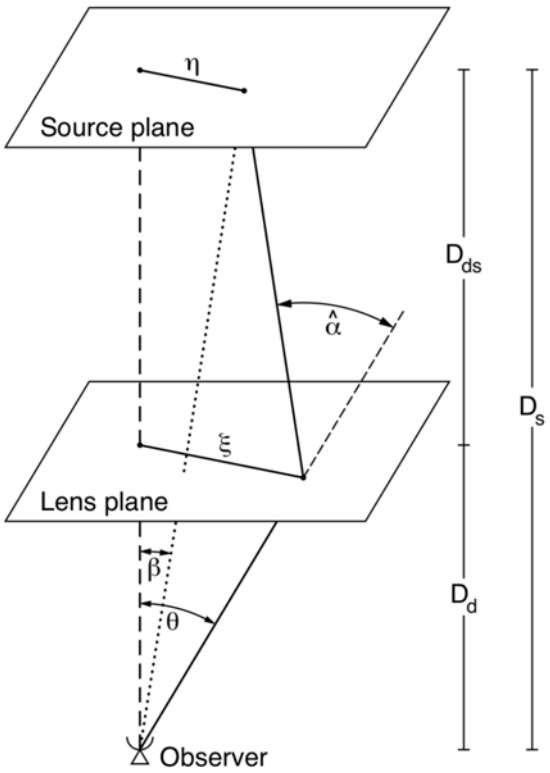

FIGURE 1.9: Scheme of the deflection of light by a gravitational lens. *Source:* Bartelmann, and Schneider (2001).

The deflection of light and the lens equation

For gravitational lensing studies, it is important to describe the deflection of the light path due to a gravitational potential field. The light travels from a source at an angular position $\boldsymbol{\beta}$ to the observer crossing a gravitational potential Φ generated by a lens source. Due to the potential, light is deflected an angle $\boldsymbol{\alpha}$, which makes that the observer perceives the source object at an angle $\boldsymbol{\theta}$. The *deflection angle* $\boldsymbol{\alpha}$ can be obtained by integrating the gravitational potential perpendicular to the line of sight:

$$\boldsymbol{\alpha}(\boldsymbol{\xi}) = -\frac{2}{c^2}\int \boldsymbol{\nabla}_{\perp}\Phi(r')\mathrm{d}r' \tag{1.65}$$

where $\boldsymbol{\xi}$ is the impact parameter, which for large distance and small angles is $\boldsymbol{\xi} = \boldsymbol{\theta} D_d$.

When observing galaxies, in general the size of the lens source is much smaller than the distances between observer and lens D_d, lens and source D_{ds}, and observer and source D_s. Hence the lens source and the source can be approximated as thin planes. If $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are small, the relation between angles and distances can be described by the *lens equation*: $\boldsymbol{\theta} D_s = \boldsymbol{\beta} D_s + \hat{\boldsymbol{\alpha}} D_{ds}$. Defining the reduced deflection angle $\boldsymbol{\alpha}(\boldsymbol{\theta}) \equiv \hat{\boldsymbol{\alpha}} D_{ds}/D_s$, the lens equation can be expressed as

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})\,. \tag{1.66}$$

The lensing potential, convergence and shear

It is useful to project the three-dimensional Newtonian potential on the lens plane. To do that, the deflection angle (1.65) should be expressed as a function of the angular distance. So the perpendicular gradient must be replaced by the gradient with respect to the angular distance as $\boldsymbol{\nabla}_\perp = \boldsymbol{\nabla}_\theta / D_d$. Then the deflection angle can be expressed as

$$\boldsymbol{\alpha} = \boldsymbol{\nabla}_\theta \Psi\,, \tag{1.67}$$

where Ψ is the *lensing potential*

$$\Psi(\boldsymbol{\theta}) = \frac{2}{c^2} \frac{D_{ds}}{D_d D_s} \int \Phi(D_d \boldsymbol{\theta}, z) \mathrm{d}z\,, \tag{1.68}$$

which is the projected potential we were looking for. Besides the gradient of Ψ giving the deflection angle, the lensing potential has a second interesting property which is that its Laplacian gives twice the *convergence* κ:

$$\nabla_\theta^2 \Psi = \frac{2}{c^2} \frac{D_d D_s}{D_{ds}} \int \nabla_\xi^2 \Phi \mathrm{d}z = \frac{2}{c^2} \frac{D_d D_s}{D_{ds}} 4\pi G \Sigma = 2 \frac{\Sigma(\boldsymbol{\theta})}{\Sigma_{\mathrm{cr}}} \equiv 2\kappa(\boldsymbol{\theta}) \tag{1.69}$$

where Σ is the surface-mass density, the Poisson equation has been used to relate the gravitational potential Ψ to the mass density, and Σ_{cr} is the critical surface density:

$$\Sigma_{\mathrm{cr}} = \frac{c^2}{4\pi G} \frac{D_s}{D_d D_{ds}}\,. \tag{1.70}$$

which depend on the angular diameter distance between lens and source. If $\Sigma \geq \Sigma_{\mathrm{cr}}$, a mass distribution produces multiple images of the source. So the value of the surface mass density Σ_{cr} helps to distinguish between strong and weak lenses.

Gravitational lensing bends the light introducing distortions into the shape of the sources. The distortion of the sources can be described by the *magnification matrix*:

$$A \equiv \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} = \delta_{ij} - \frac{\partial \alpha_i}{\partial \theta_j} = \delta_{ij} - \frac{\partial^2 \Psi}{\partial \theta_i \partial \theta_j} = \begin{pmatrix} 1-\kappa-\gamma_1 & -\gamma_2 \\ -\gamma_2 & 1-\kappa+\gamma_1 \end{pmatrix}. \tag{1.71}$$

The convergence κ describes how the size of images is magnified isotropically. While the components of the *shear* tensor $\gamma = \gamma_1 + i\gamma_2 = |\gamma| e^{2i\phi}$ describe the elliptical deformation of the shape of images, where ϕ corresponds to the orientation angle of the major axis. The shear transforms the image to an ellipse with a major axis $a = (1-\kappa-|\gamma|)^{-1}$ and a minor axis $b = (1-\kappa+|\gamma|)^{-1}$, which are the eigenvalues of the inverse of the magnification matrix A. Using the simplified notation $\Psi_{ij} \equiv \partial^2\Psi/\partial\theta_i\partial\theta_j$, the convergence and the shear components can be expressed as

$$\kappa = \frac{1}{2}\left(\Psi_{11} + \Psi_{22}\right) = \frac{1}{2}\mathrm{Tr}\,\Psi_{ij} \tag{1.72}$$

$$\gamma_1 = \frac{1}{2}\left(\Psi_{11} - \Psi_{22}\right) \equiv \gamma(\boldsymbol{\theta})\cos\left[2\phi(\boldsymbol{\theta})\right] \tag{1.73}$$

$$\gamma_2 = \Psi_{12} = \Psi_{21} \equiv \gamma(\boldsymbol{\theta})\sin\left[2\phi(\boldsymbol{\theta})\right] \tag{1.74}$$

where ϕ is the position angle of the ellipse.

Measuring the ellipticity of observed galaxies ϵ, the shear can be determined through the relation

$$\epsilon = \frac{\epsilon_\mathrm{s} + g}{1 + g^* \epsilon_\mathrm{s}} \tag{1.75}$$

where ϵ_s is the intrinsic ellipticity of the source galaxies and $g = \gamma/(1+\kappa)$ is called the reduced shear. In the weak lensing regime, small distortions ($|\gamma| \ll 1$ and $|\kappa| \ll 1$) are assumed, so $\epsilon \simeq \epsilon_\mathrm{s} + g \simeq \epsilon_\mathrm{s} + \gamma$. The intrinsic ellipticity can not be directly measured, so usually it is assumed that their orientation is random thus the average intrinsic ellipticity for a large sample should be zero, $\langle\epsilon_\mathrm{s}\rangle = 0$. In that case, the average of the observed ellipticities of background galaxies is a direct measure of the shear induced by the foreground mass distribution $\langle\epsilon\rangle = \langle\gamma\rangle$. However, background galaxies have indeed intrinsic alignments that are not random and are one of the main systematics in the determination of the galaxy shear.

Convergence power spectrum

So far we have used the single lens approximations to derive the lensing parameters. However, to determine the weak lensing parameters, the large scale structure of the universe will be observed. Then, the derived result should be extended to extended lenses. To do so, the angular diameter distance used so far should be converted to comoving distances χ through $D_{ds}/(D_d D_s) \rightarrow (\chi_\mathrm{s} - \chi)/(\chi_\mathrm{s}\chi)$ where χ_s is the distance to the source. Then the lensing potential defined in equation 1.68 for a thin lens approximation becomes

$$\Psi(\boldsymbol{\theta}) = \frac{2}{c^2}\int_0^{\chi_\mathrm{s}} \frac{\chi_\mathrm{s} - \chi}{\chi_\mathrm{s}\chi}\Phi(\chi\boldsymbol{\theta}, \chi)\mathrm{d}\chi \tag{1.76}$$

for an extended lens, where $\chi\boldsymbol{\theta}$ is the perpendicular direction and χ is the line of sight position.

Through the Poisson equation in comoving coordinates $\nabla^2\Phi = 4\pi G\bar{\rho}a^2\delta/c^2$, the three dimensional gravitational potential can be related to the overdensity field as

$$\nabla^2\Phi = \frac{3\Omega_m H_0^2}{2a}\delta \tag{1.77}$$

where the mean matter density $\bar{\rho}_\mathrm{m} = (3H_0^2/8\pi G)\Omega_\mathrm{m}a^{-3}$ has been used. Using this expression and the relation between the lensing potential and convergence (1.69) and equation 1.76, the convergence can be related to the overdensity field

$$\kappa = \frac{3\Omega_m H_0^2}{2c^2}\int_0^{\chi_\mathrm{s}} \frac{\chi(\chi_\mathrm{s} - \chi)}{\chi_\mathrm{s}}\frac{\delta(\chi\boldsymbol{\theta}, \chi)}{a}\mathrm{d}\chi\,. \tag{1.78}$$

This expression of κ for an extended source is called *effective convergence* because it is equivalent to the convergence caused by the extended matter distribution.

A common method to constrain the cosmological parameters is using second order statistics (not first order statistic because the expected average of the overdensity field is zero then the field is characterized by the second order statistics) related to the two dimensional convergence power spectrum $P_\kappa(l)$ defined as

$$\langle\tilde{\kappa}(\boldsymbol{l})\tilde{\kappa}(\boldsymbol{l'})\rangle = (2\pi)^2\delta_\mathrm{D}(\boldsymbol{l} - \boldsymbol{l'})P_\kappa(l) \tag{1.79}$$

where $\tilde{\kappa}$ is the Fourier transform of the two dimensional convergence

$$\tilde{\kappa}(\boldsymbol{l}) = -\frac{l^2}{2}\tilde{\Psi}(\boldsymbol{l})\,. \tag{1.80}$$

The two dimensional convergence power spectrum can be expressed as a function of the three dimensional power spectrum as

$$P_\kappa(l) = \left[\frac{3H_0^2\Omega_m}{2c}\right]^2 \int_0^{X'} \left[\frac{g(\chi)}{a}\right]^2 P\left(k = \frac{l}{\chi}, \chi\right) \mathrm{d}\chi \tag{1.81}$$

where χ' is the comoving distance of the source galaxies and $g(\chi)$ is the weighted lens efficiency function defined as

$$g(\chi) = \int_\chi^{\chi'} n_\chi(\chi)\frac{\chi - \chi_\mathrm{s}}{\chi}\mathrm{d}\chi \tag{1.82}$$

where n_χ is the redshift distribution of sources in comoving coordinates.

The convergence and shear have the same power spectrum. First we transform the shear components into Fourier space

$$\tilde{\gamma_1}(\boldsymbol{l}) = -\frac{l_1^2 - l_2^2}{2}\tilde{\Psi}(\boldsymbol{l}) \quad \text{and} \quad \tilde{\gamma_2}(\boldsymbol{l}) = -l_1^2 l_2^2 \tilde{\Psi}(\boldsymbol{l}) \tag{1.83}$$

where $\boldsymbol{l} = (l_1, l_2)$ is the wave-vector scale. If we use these expressions as well as the Fourier transform of the convergence (1.80):

$$4|\tilde{\gamma}|^2 = \left[\left(l_1^2 - l_2^2\right)^2 + 4l_1^2 l_2^2\right]|\tilde{\Psi}|^2 = \left(l_1^2 - l_2^2\right)^2 |\tilde{\Psi}|^2 = 4|\tilde{\kappa}|^2\,. \tag{1.84}$$

Therefore the shear power spectrum will also be given by equation 1.81.

Shear correlation function

In the weak gravitational regime, the deformation of images is caused by small distortions in the shape and size. The small distortion, known as shear or cosmic shear, can be inferred from observations. To determine the shear caused by lensing, it is necessary to measure the ellipticities of the source galaxies. However, not all the galaxy measured ellipticities are caused by gravitational lensing. Most galaxies have elliptical shapes that are intrinsic to the galaxy. Only very few are round. These

ellipticities are very large compared to the ones induced by lensing or intrinsic alignments (IA) – that are the induced ellipticities due to other astrophysical effects –. The intrinsic ellipticities of galaxies are in principle random and vanish to zero when statistically combined. What remains then is the effect of weak lensing distortions and intrinsic alignment distortions.

The shear can be decomposed into the tangential shear γ_t and the cross-component γ_x:

$$\gamma_t = \gamma \cos(2\beta) \quad \text{and} \quad \gamma_x = \gamma \sin(2\beta) \tag{1.85}$$

where β is the polar angle for a given direction $\boldsymbol{\theta}$. With this two quantities there are three two point correlation function and power spectrum, formed by the auto-correlation of each component:

$$\xi_{tt}(\theta) = \langle \gamma_t \gamma_t^* \rangle = \frac{1}{2} \int \frac{l \mathrm{d}l}{2\pi} P_\kappa(l) \left[j_0(l\theta) + j_4(l\theta) \right] , \tag{1.86}$$

$$\xi_{xx}(\theta) = \langle \gamma_x \gamma_x^* \rangle = \frac{1}{2} \int \frac{l \mathrm{d}l}{2\pi} P_\kappa(l) \left[j_0(l\theta) - j_4(l\theta) \right] , \tag{1.87}$$

and the cross-correlation of both components $\xi_{tx}(\theta) = 0$, which vanished due to symmetries. Then it makes sense to define the correlation functions $\xi_\pm(\theta) \equiv \langle \gamma_t \gamma_t^* \rangle \pm \langle \gamma_x \gamma_x^* \rangle$:

$$\xi_+(\theta) = \int \frac{l \mathrm{d}l}{2\pi} P_\kappa(l) j_0(l\theta) , \tag{1.88}$$

$$\xi_-(\theta) = \int \frac{l \mathrm{d}l}{2\pi} P_\kappa(l) j_4(l\theta) . \tag{1.89}$$

where j_0 and j_4 are the spherical Bessel function of order 0 and 4.

1.6.2 Galaxy clustering

The initial quantum perturbations originated during inflation grew into overdensities that generated gravitational attraction wells and ended up forming the large scale structures, which continue evolving. The distribution over time and across the universe of galaxies have imprinted the information of the cosmological model. Cosmological surveys map the distribution of observable galaxies, which can be related to the underlying matter distribution, i.e. dark matter, if it is weighted by a galaxy bias function. The bias can be determined through prior knowledge from simulations of galaxy formation or from gravitational lensing effects.

The distribution of galaxies is determined in redshift space. So prior assumptions on the cosmological parameters are needed to determine the distance of galaxies. In addition, dynamical effects that distort the redshift space should be taken into account to correctly determine the distance such as the line-of-sight peculiar velocity that elongate the galaxy clusters (effect known as Fingers of God) and the motion of galaxies when they fall inwards their cluster that make the clusters look compressed along the radial direction (known as the Kaiser effect).

To constrain the cosmological parameters using the observed galaxy distribution, the angular clustering correlation and the galaxy power spectrum can be used.

1.6.3 Galaxy-galaxy lensing

The cosmological probe that correlates the measured lensing of background galaxies, produced by the foreground mass distribution, with the observed foreground galaxy distribution is called galaxy-galaxy lensing. At small scales, this combined probe allows to trace the underlying matter distribution. At large scales, since the lensing allows to directly determine the total matter content, the combined probe allows to correlate the observed galaxy distribution with the total matter distribution. This correlation is useful to determine the large-scale bias.

The signal from galaxy clustering scales as $(b\sigma_8)^2$ (remember σ_8 indicates the density fluctuation on a scale 8 h^{-1}Mpc which is the typical scale of massive galaxy clusters) and at large scales the signal from galaxy-galaxy lensing scales as $(\Omega_\mathrm{m} b\sigma_8)^2$. Then the combination allows to determine the bias and thus to constrain $\Omega_\mathrm{m}\sigma_8$. Hence this combination of probes breaks the degeneracy of the cosmological parameters from each probe. Since these parameters are constrained, galaxy-galaxy lensing also helps to decrease the uncertainty on the recovered dark energy equation of state parameter (ω_0 and ω_a when considering cosmological models that parameterize the dark energy).

To extract more cosmological information from the power spectrum to constrain the cosmological constrains, galaxies are split in tomographic redshift bins. Photometric redshifts are usually used to split the data and determine the location of galaxies. Therefore accurate and precise photometric redshifts are a key component affecting the amount of information that can be obtained from cosmological probes.

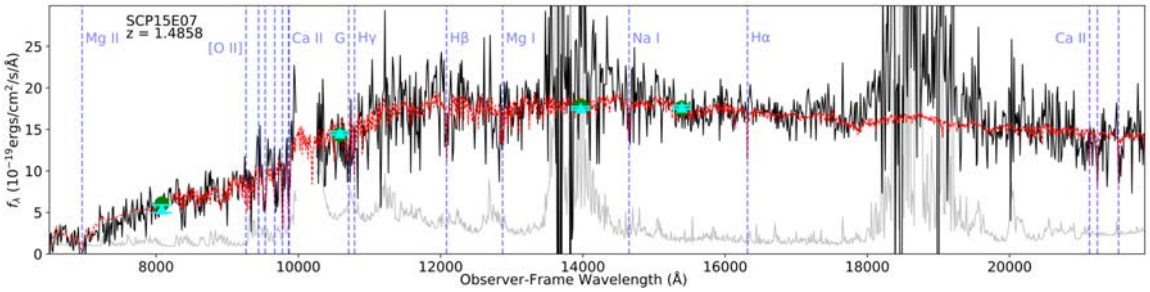

FIGURE 1.10: Plot from Williams et al. (2020). The spectral energy distribution of the host galaxy of SN Ia SCP15E07 is shown in the solid black line. The best template fit (red dashed line) corresponds to a template redshifted to $z = 1.4858$.

1.7 Determination of redshift

In order to have a picture of the three-dimensional distribution of galaxies across the observable universe, their distances must be determined. Inferring the distances from observations is not a trivial task. So usually the redshift is used as an estimate of the distance, although its determination is not simple. Remember that due to the expansion of the universe and the finite velocity of light, the wavelength emitted by a source λ_e suffers a variation from the moment it is emitted until it is observed λ_o. The ratio between the difference of wavelengths and the emitted wavelength is defined as redshift

$$z \equiv \frac{\lambda_o - \lambda_e}{\lambda_e} \,. \tag{1.90}$$

To infer the redshift, one observes the flux density f_λ, which is the energy received per time and unit of area as a function of wavelength. There are two methodologies to derive the redshift: one is spectroscopy that measures the flux along the wavelength range and the other one is photometry, which observes the flux just through a few filters. In the spectroscopic technique, the observation of the flux as a function of the wavelength – called spectral energy distribution (SED) – allows to determine with precision the redshift but at the expenses of needing more observational time for each object to measure the full SED. On the other hand, in the photometric methodology, by only measuring the flux through a few filters, the required observational time per object drastically decreases allowing to observe a larger amount of objects but at the expenses of loosing precision in the redshift.

1.7.1 Spectroscopic redshift

In spectroscopy, the complete spectral energy distribution (SED) of an object is obtained by using a spectrograph that splits the light of the object into narrow bins of wavelength through dispersion. Then the SED is compared to a set of observed galaxy spectra with known redshift or theoretical templates that include models for different types of galaxies such as ellipticals, spirals, irregulars, and starbursts. To choose the best template fit, typical spectral features are matched with the same features from the templates. As can be seen in Fig. 1.10, such features are absorption lines, emission lines, or other variations in light intensity such as the H_α emission line at 6563 Å or the 4000 Å break. The redshift can be inferred using the difference in wavelength obtained by comparing the position in wavelength of the spectral features of the best fit template with the original position at restframe.

1.7.2 Photometric redshift

In the photometric methodology, instead of the full spectral energy distribution, the spectra are only measured through a few wavelength filters, obtaining a discrete relation between intensity and wavelength and thus spectra with very low resolution. Each of the filters has a transmission function $T_i(\lambda)$ that determines the fraction of photons that passes through the filter. For an object that has a spectral flux density f_λ, the photon flux through a filter would be

$$f_i = \frac{\int T_i(\lambda) f_\lambda \lambda \mathrm{d}\lambda}{c \int T_i(\lambda) \frac{\mathrm{d}\lambda}{\lambda}} . \tag{1.91}$$

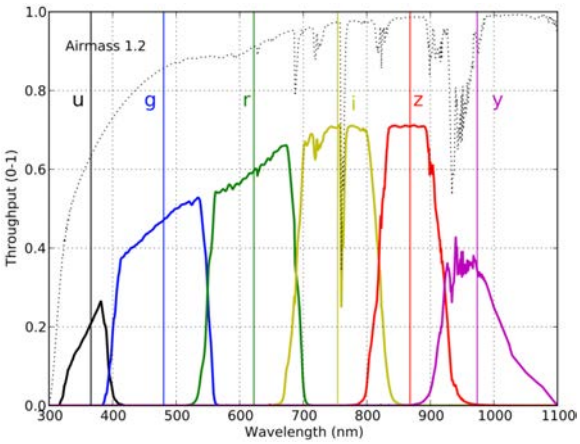

FIGURE 1.11: Total throughput of the Vera C. Rubin Observatory - Large Synoptic Survey Telescope (LSST) bands once the atmospheric transmission (dotted line), the optics, and the detector sensitivity have been taken into account. *Source:* Ivezić et al. (2019).

In Fig. 1.11 an example of the filters throughput is shown for a set of broad-bands filters. Limitations of photometric redshift are that the effective wavelength resolution is only as accurate as narrow the filter bandpasses are and that the ability to see individual spectral features is lost. In

addition, it is difficult to transfer the information between surveys that use different filter set.

The main advantage of this method is a better signal-to-noise ratio due to the concentration of light from a source onto one spot on the detector over one filter band, which reduces the observational time needed for each source. In addition, flux can be measured for several objects at the same time, increasing the amount of objects that can be observed in the same amount of time. Using a similar observational time for photometric observation than the time used for spectroscopic, deeper fluxes can be measured since the signal-to-noise ratio is better for fainter objects that it would be for spectroscopic observations. Therefore, photometric observations allow to observe farther away and so to better trace the large-scale structures. Next, we briefly overview the two main approaches to determine photometric redshifts: template fitting and training methods.

Template fitting

Similar to the spectroscopic methodology, the template fitting approach requires having a library of theoretical or empirical templates for many galaxy types. This set of templates are redshifted and the expected flux calculated in several filters and redshift values. These values are compared to the observed flux points of the different bands of the observed object until a best match is found. The best match is the template that minimizes the value

$$\chi^2(z) = \sum_{i=1}^{N_{\text{filters}}} \left[\frac{f_i - a f_{\text{template},i}(z)}{\sigma_i} \right]^2 \tag{1.92}$$

where σ_i is the flux uncertainty in filter i, a is a normalization constant, and f_{template} is the predicted flux of the template in a certain redshift computed by integrating the transmission function of each filter over the predicted template. Finding the template that best fits the observed object allows to determine both the galaxy type and the redshift of the object.

Due to the finite number of templates and the limited resolution of the spectral energy distribution, there can be mismatches between templates and galaxies of the survey. This problem can not be solved by adding more templates since too many templates can lead to degeneracies where the template library can give two different redshift for the same fluxes. These mismatches can cause systematic errors in the

redshift estimation. In addition, observations are affected by other effects such as the contribution of emission lines, dust or AGN, requiring adapted templates for different cases that usually we do not have. Therefore, it is specially important that the template set is representatives of the observed galaxies, avoiding to suffer of template incompleteness. A thorough discussion can be found in Walcher et al. (2011).

Training technique

In training-based methods, first a subset of objects with known redshift, called training set, is chosen. The training method consist on using this subsample to describe the redshift distribution in flux and color space and thus to find a mapping function between the fluxes and redshift. This function will be applied over another part of the sample, called validation sample, to see whether the algorithm is well calibrated. Then the mapping function will be applied over the rest of the sample, called test set, with unknown redshifts to determine their redshift. The algorithms that perform training-based methods are called machine-learning algorithms. Some examples of the most used machine-learning algorithms are decision tree classification, random forests, neural networks and Gaussian process regression. Advantages of training methods are that the correlations between the flux, color and redshift can be determined with a high degree of confidence (depending on the training set available), the algorithms can handle large training sets and they return strong probabilistic estimates on the redshift. Drawbacks of these methods are that the redshift estimation is only accurate when objects in the training set have the same observables as the rest of the sample and that the training set must be large enough to ensure that the parameter space in color, flux, galaxy type and redshift is well covered. For further discussion we refer the reader to Walcher et al. (2011).

Chapter 2

Galaxy surveys

Determining cosmological parameters requires to sample a large volume of the universe. This volume is traced with galaxies. So we need to observe a large amount of galaxies to do cosmology. Galaxy surveys observe large data sets of galaxies and retrieve physical properties such as the position and redshift, which are needed to trace the underlying large-scale structure and thus to extract cosmological information from observations. Galaxy surveys can be classified into spectroscopic and photometric (also called imaging) surveys, depending on whether the redshift of the observed objects is estimated with spectroscopy or using photometric techniques. Photometric surveys observe galaxies with multi-band filters instead of observing their full spectral energy distribution as in spectroscopy that requires more observational time. Nevertheless, photometric surveys provide measurements for many more objects than spectroscopic surveys but at the expense of a degraded precision on the redshift estimates. Photometric surveys such as the Dark Energy Survey[1] (DES; Dark Energy Survey Collaboration 2005), the Kilo-Degree Survey[2] (KiDS; de Jong et al. 2013), the Hyper Suprime-Cam Subaru Strategic Program[3] (HSC-SSP; Aihara et al. 2018), and the upcoming Euclid[4] (Laureijs et al., 2011) and the Vera C. Rubin Observatory Legacy Survey of Space and Time[5] (Rubin-LSST; LSST Science Collaboration: Abell et al. 2009), will be an essential tool to trace the large-scale structure of the universe. They will allow us to constrain a large variety of cosmological models.

In this chapter we describe the main photometric surveys whose data is used along this work. We present the already available data of DES and the Physics of the

[1] https://www.darkenergysurvey.org
[2] http://kids.strw.leidenuniv.nl
[3] https://hsc.mtk.nao.ac.jp/ssp/
[4] https://www.euclid-ec.org
[5] https://www.lsst.org

Accelerating Universe (PAU) photometric surveys, and the upcoming Euclid space mission.

2.1 Dark Energy Survey

The Dark Energy Survey (DES; Dark Energy Survey Collaboration 2005) was designed to better understand the nature of dark energy and to probe the acceleration of the universe by constraining the cosmological parameters with four complementary techniques: large-scale galaxy angular clustering, weak gravitational lensing, type Ia supernovae and baryon acoustic oscillations. The last two probes give information of the expansion of the Universe while the first two probes help to constrain the geometry and the evolution of the Universe. fjctwhile the first two probes help to constrain the growth of structure and the geometry of the Universe.

The Dark Energy Camera (DECam; Flaugher et al. 2015) was built to carry out a large survey of 5000 deg^2 in the Southern Hemisphere. The 570 megapixel focal plane of the camera has 62 Charge Coupled Devices (CCDs) for imaging spanning a total of 3 deg^2 field of view and 12 CCDs for guiding and focus. The camera was mounted on the Blanco 4-meter telescope at the Cerro Tololo Inter-American Observatory (CTIO) in northern Chile. After observations, the images were processed at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign.

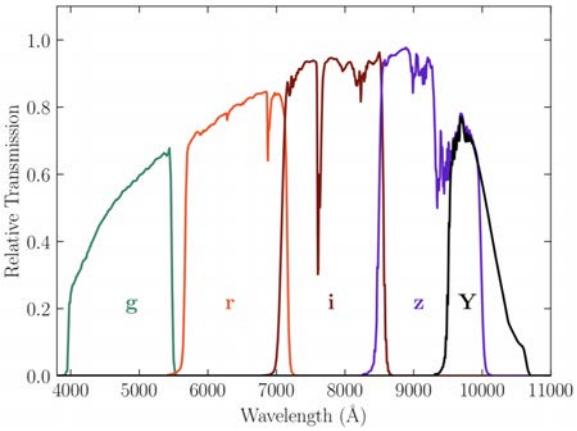

FIGURE 2.1: Total system response for DES broad band filters, including atmospheric transmission (airmass = 1.2) and the average instrumental response across the CCDs (Abbott et al., 2018b).

DES started observations in August 2013 and finished in January 2019, after 758 nights of observations. The survey covers a main wide field (WF) of 5000 deg^2 around the southern galactic cap. The DES footprint is shown in Fig. 2.2. The sky was observed through 5 broad band filters, $grizY$, with wavelengths ranging from 400 to 1065 nm, shown in Fig. 2.1. The wide field survey used single-exposures times of 90 seconds in $griz$ and 45 seconds in the Y band, yielding a nominal limiting

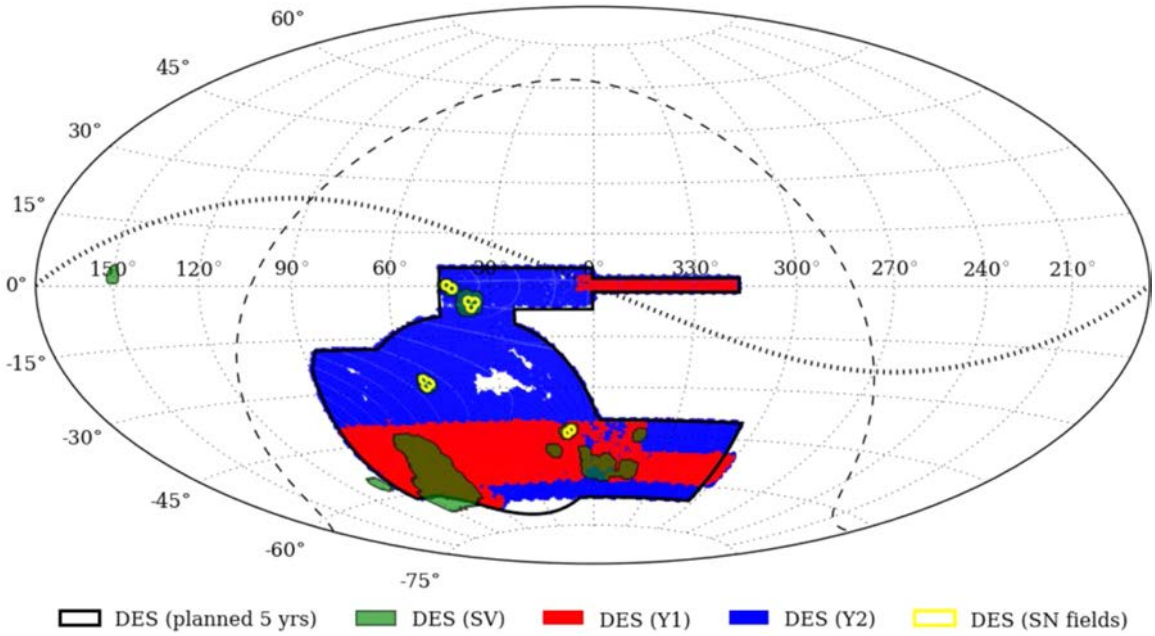

FIGURE 2.2: The Dark Energy Survey observational footprint. Outlined in black is the 5000 deg^2 wide area survey of DES by the end of observations. Areas in green correspond to regions used for Science Verification (SV) as described in Dark Energy Survey Collaboration: Abbott et al. (2016). The sky coverage during the first (Y1) and second (Y2) years of observations is colored in red and blue, respectively. Outlined in yellow are the Supernova fields. The Milky Way or galactic plane is shown as a horizontal dashed line. The ecliptic plane is shown as a vertical dashed line. *Credit:* Diehl et al. (2017).

magnitude for single-epoch point spread function (PSF) at signal to noise 10σ of $g = 23.57$, $r = 23.34$, $i = 22.78$, $z = 22.10$, and $Y = 20.69$ (Morganson et al., 2018). The final coadded depths are expected to be roughly one magnitude deeper, with the coaddition of 10 images in each band. Observations allowed DES to detect objects up to redshift 1.4. The survey also observed 10 time-domain fields of 27 deg^2 in total in the *griz* bands with a 7 day cadence, to discover and study supernovae (Dark Energy Survey Collaboration: Abbott et al., 2016).

2.1.1 Gold Catalog

For this thesis, we use data from the DES Year 3 Gold catalog release version 2.2 and refer to it as DES data. This catalog will be described in more detail in Sevilla-Noarbe et al. (in prep.). As a reference, the catalog is a subsample of the first public data release of DES (Dark Energy Survey Collaboration: Abbott et al., 2018) that contains information from the first three years of observations (from August

TABLE 2.1: Selection criteria for DES Y3 Gold subsample.

Description	Selection
High confidence galaxies	extended_class_sof = 3
Extreme colors	$-1 < \mathrm{g} - \mathrm{r} < 3$
	$-1 < \mathrm{r} - \mathrm{i} < 2.5$
	$-1 < \mathrm{i} - \mathrm{z} < 2$
Extreme magnitudes	$\{\mathrm{g}, \mathrm{r}, \mathrm{i}, \mathrm{z}\} < 30$
Photometric redshift sentinel value	dnf_zmean_sof ! = −9999

2013 to February 2016) in the wide area. The Gold catalog is an internal release with better photometry than previous releases since problematic objects have been removed or better flagged, and photometric redshift codes have been rerun using the improved photometry. The procedures used to determine the photometric redshift are similar to the ones used in the Year 1 Gold internal release, which had value-added products and detailed characterizations of survey performance designed to support cosmological analyses (Drlica-Wagner et al., 2018). In the first public data release, at the end of year 3, the final median coadded images depth for a 1.95" diameter aperture at signal to noise 10σ is $g = 24.33$, $r = 24.08$, $i = 23.44$, $z = 22.69$, and $Y = 21.44$ magnitudes. The catalog contains 400 million galaxy and stellar objects (Dark Energy Survey Collaboration: Abbott et al., 2018).

2.1.2 Data

In this section we briefly give the details of the sample we use to remap the photometry from simulations to resemble DES, as explained in chapter 3. To test the remap method, we take a subsample from the Year 3 Gold catalog of the wide field. The area of sky considered is within $20° < \alpha < 40°$ and $-28° < \delta < 13°$ which corresponds to an area of about 280 deg^2. The Gold catalog contains photometry processed in different ways. To reproduce the photometry of DES we choose the sof (single object fitting for PSF) magnitudes in *griz* and their corresponding errors. The sof magnitudes are computed by the DES Data Management (DESDM; Morganson et al. 2018) with a simplified pipeline of the one used to compute the mof (multi object fitting for PSF) magnitudes, see Drlica-Wagner et al. (2018) for further details. The sof pipeline was only run on the *griz* bands but not on the Y band.

To remap the photometry between simulations and DES, we want to avoid having problematic photometry and objects. We apply some quality selection criteria, shown in Table 2.1. We select objects that have a high confidence of being galaxies, using a confidence flag from the catalog shown in Table 2.1. We remove objects with extreme colors (according to Crocce et al. (2019a)) that may be either nonphysical or from specific samples such as high redshift quasars, to avoid odd photometry and photometric redshift values in the sample. Galaxies with extreme magnitudes that are not possible to be measured can be removed. Finally, we select galaxies that have their photometric redshift successfully determined.

2.2 Physics of the Accelerating Universe Survey

The Physics of the Accelerating Universe[6] (PAU) is a photometric survey that uses the PAUS camera (PAUCam; Padilla et al. 2019) mounted at the William Herschel Telescope (WHT) at the Observatory of the Roque de los Muchachos in the Canary Islands. The camera was successfully mounted at the telescope in June 2015 and it have observed periodically since then. The camera has a field of view of 1 degree of diameter and a focal plane composed of 18 CCDs. The main feature of PAUS is that it observes with 40 narrow band (NB) filters that have a full width at half maximum (FWHM) of 13 nm and cover the optical wavelength range from 450 nm to 850 nm in intervals of 10 nm (Fig. 2.3). The narrow band observations are complemented with 6 broad band filters, *ugrizY*.

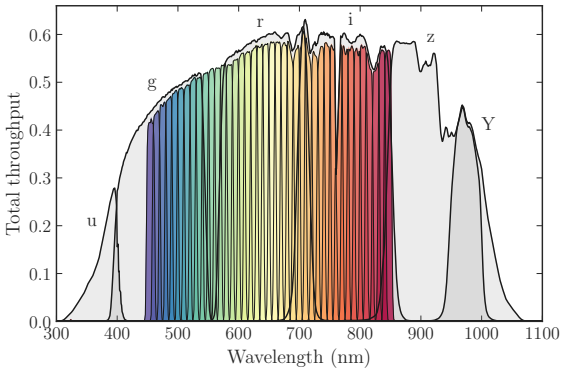

FIGURE 2.3: Effective response throughput of PAUS filters, once the transmissions curves have been taken into account. Including the effects of atmospheric transmission corresponding to the Apache Point Observatory, the quantum efficiency of Hamamatsu CCD, and the optics of the William Herschel Telescope (Casas et al., 2016). Broad-bands are shown in shaded grey and narrow-bands in spectral palette.

The use of narrow bands in PAUS offers a better resolution and detection of features of the spectral energy distribution than photometric surveys with few broad

[6] https://www.pausurvey.org

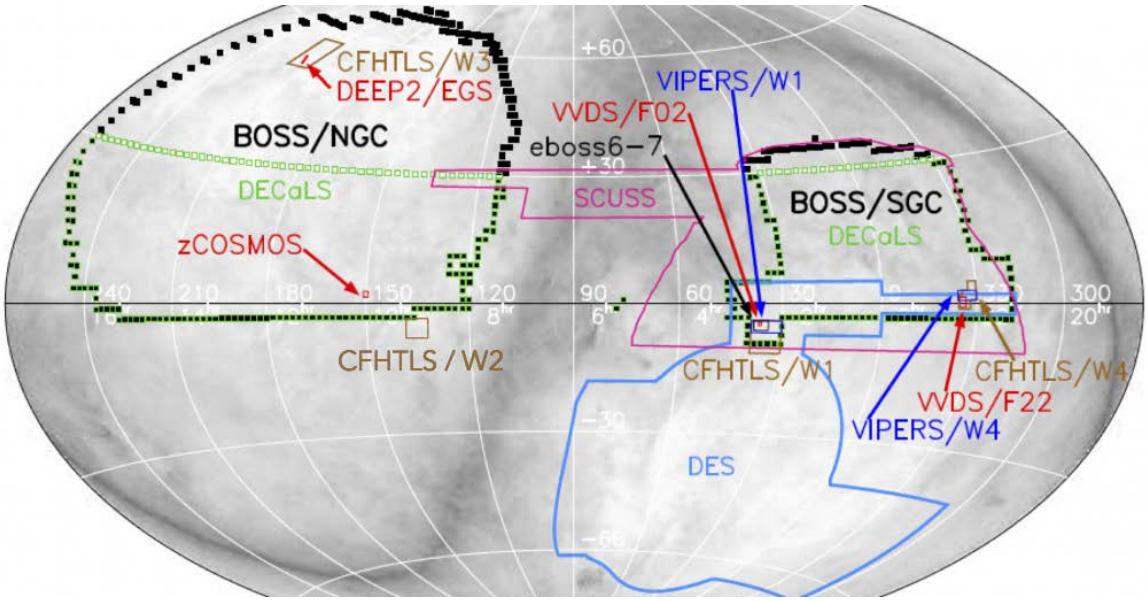

FIGURE 2.4: The Physics of the Accelerating Universe observational fields in comparison to other surveys footprints. PAUS targets COSMOS and W1, W2, W3 and W4 fields from CFHTLS. *Source:* PAUS webpage.

band filters, thus improving the photometric redshift precision by an order of magnitude but at the expenses of covering a smaller area than broad band surveys for the same observational time. While the typical broad band photometric redshift error is of $\sigma_{68}/(1+z) \simeq 0.05$ (Hildebrandt et al., 2012), PAUS aims to reach sub-percent photometric precision, of about $\sigma_{68}/(1+z) \simeq 0.0035$ (Martí et al., 2014), for about 3 million galaxies down to $i < 22.5$ and to detect galaxies until $i \sim 24$ with lower precision photometric redshifts.

PAUS targets the COSMOS[7] (Scoville et al., 2007) field and the W1, W2, W3 and W4 fields from the Canada-France-Hawaii Telescope Lensing Survey[8] (CFHTLenS; Heymans et al. 2012) covering approximately 100 square degrees. The fields observed with PAUS (see Fig. 2.4) were chosen to overlap with other surveys in order to have galaxy shapes, sizes, positions and broad-band photometry available from other photometric surveys such as CFHTLens, DES and KiDS, and spectroscopic redshift for calibration from the Sloan Digital Sky Survey[9] (SDSS; York et al. 2000), the VIMOS Public Extragalactic Redshift Survey[10] (VIPERS; Guzzo et al. 2014) and the Cosmic Evolution Survey (zCOSMOS; Lilly et al. 2009).

[7] http://cosmos.astro.caltech.edu
[8] http://cfhtlens.org
[9] http://www.sdss.org
[10] http://vipers.inaf.it

The area, depth and precision of PAUS increase the number density of galaxies with known redshift by almost two orders of magnitudes to tens of thousands of redshifts per square degree (Eriksen et al., 2019), which is ideal to study intrinsic alignments, redshift space distortions and galaxy clustering (Gaztañaga et al., 2012; Eriksen, and Gaztañaga, 2015). In addition, the photometric redshift precision is good enough to detect the baryon acoustic oscillations signal (Benítez et al., 2009).

2.2.1 Data

In this section, we briefly discuss the PAUS data that we use in chapter 4.

Narrow bands

We quickly summarize the production of narrow band fluxes. The observed PAUS data are transferred from the WHT to the Port d'Informació Científica (PIC; Tonello et al. 2019). There, the data go through the image process, reduction and analysis pipeline, called nightly pipeline (Serrano et al. in prep.). After the nightly pipeline, the MEMBA pipeline performs photometric calibration and measures the galaxy fluxes (Serrano et al. in prep.). The photometry is calibrated by fitting stellar templates to the *ugriz* broad bands (Smith et al., 2002) of the Sloan Digital Sky Survey (SDSS). Then, the spectral energy distribution of the corresponding best fit template is used to predict the expected fluxes in the PAUS narrow bands. The zero-points for each image are determined by comparing the predicted and observed fluxes. The narrow band fluxes are measured on individual exposures and determined using forced photometry by placing an aperture on each galaxy in the image. The aperture is centered at the position of the galaxy given by the already known position in the COSMOS and CFHTLens fields as measured from these surveys. The radius of the aperture for each galaxy is set to measure 62.5% of fraction of light, considering that the galaxy follows a Sersic profile and knowing the radius of the galaxy and the Point Spread Function (PSF) at its position. Every galaxy is observed a few times in different exposures and the resulting flux is obtained by combining the different measurements.

The MEMBA pipeline is run periodically, usually after each observation period. Each PAUS field has been reduced with different versions of the pipeline as it has been evolving with time. In this work we use the fluxes corresponding to the production identified as 866, available at PIC.

Broad bands

In addition to the narrow bands, six broad bands are used to compute the photometric redshift of PAUS (Eriksen et al., 2019). The broad bands come from the COSMOS catalog as available in Laigle et al. (2016) (COSMOS2015). The bands used are u^* from CFHTLenS and B, V, r, i^+, z^{++} from Subaru (Taniguchi et al., 2015). In this COSMOS2015 catalog, photometric redshifts determined with 30 bands are also available. We use those redshifts in this thesis. We denote them as $zp_{\rm gal}$.

Spectroscopic data

To validate the photometric redshifts of PAUS and determine their accuracy, in Eriksen et al. (2019) and Eriksen et al. (2020), the photometric redshift estimation of PAUS is compared to the zCOSMOS spectroscopic redshift data (Lilly et al., 2007). These data cover an area of about 1.7 deg^2 centered at $\alpha = 150.119°$ and $\delta = 2.206°$, span the redshift range between 0.1 and 1.2, and contain galaxies in the magnitude range $15 < i < 22.5$. The COSMOS field is chosen to validate the photometric redshifts because it has been widely observed in the wavelength range from ultraviolet to infrared with multiple bands, and it has extensive spectroscopy available. In addition to spectroscopic reshifts, the zCOSMOS catalog contains a confidence class to indicate the reliability of the redshift determination (Lilly et al., 2009). A confidence class of $3 \leq$ conf. ≤ 5 denotes a highly secure redshift.

In this thesis, we use the spectroscopic data set described above. We denote the spectroscopic redshift of zCOSMOS as $z_{\rm spec}$. We consider the highly secure spectroscopic redshifts but we also use the full redshift sample without confidence cuts for comparison.

Photometric redshifts

In Martí et al. (2014) it was assessed the photometric redshift performance in PAUS using simulated data. They obtained a root mean square (RMS) scatter in the redshfit determination of $\sigma_{68}/(1+z) \simeq 0.0035$. To calculate the expected precision, the photometric redshifts were determined using the Bayesian photometric redshifts (BPZ; Benítez 2000) code, which is a template-based algorithm that uses priors within a Bayesian framework. Recently, a preliminary performance of the photometric redshift of PAUS in the COSMOS field was carried out obtaining a precision of $\sigma_{68}/(1+z) \simeq 0.0037$ (Eriksen et al., 2019). In Eriksen et al. (2019), instead of

using BPZ to determine the photometric redshifts, they design and use a new algorithm called BCNZ2. BCNZ2 is a template fitting based code specifically developed to determine the photometric redshifts of PAU. BCNZ2 compares the observed flux in all bands with a set of galaxy flux models that evolve with redshift. To find the best model, the code performs a linear combination of templates and it estimates the redshift probability distribution including the use of priors. BCNZ2 also takes into account a different zero point for each band since PAUS relies on broad bands observations from external surveys that might have different photometry and apertures. In addition, the template sets include emission lines, which are a critical feature to extract spectral energy distribution information to determine precise photometric redshift from PAUS fluxes.

In this work, we use the photometric redshifts of PAUS determined with BCNZ2, which we denote as $z_{\rm b}$.

2.3 Euclid

Euclid is a medium class astronomy and astrophysics space mission of the European Space Agency (ESA) due for launch in 2022. The main goal of Euclid is to understand the origin of the acceleration of the Universe and to investigate the nature of dark energy, dark matter and the validity of general relativity by analyzing their observational traces in the geometry of the Universe and the growth of cosmic structures. In order to achieve that, Euclid will measure precise shapes and photometric redshifts for more than 1.5x10^9 galaxies, will take over 50 million spectra and will sample the distribution of clusters of galaxies up to redshift $z \sim 2$ in a large area of the sky.

The payload module of Euclid is equipped with a 1.2 m diameter telescope that delivers a 1.25x0.727 deg^2 field of view. The telescope contains two instruments: the Near-Infrared Spectro-Photometric instrument (NISP; Costille et al. 2018), and the Visible imager at visible wavelengths (VIS; Cropper et al. 2018). Both instruments will observe a field of view of 0.53 deg^2. The VIS instrument will observe galaxies through a single optical broad band which is similar to the sum of the riz standard broad bands, covering a wavelength range between 540 and 900 nm with a magnitude limit depth of 24.5 at a signal to noise of 10σ for extended sources. The VIS instrument is equipped with 36 CCDs and it will measure the shapes of galaxies with a resolution better than 0.2 arcsec given by its narrow PSF. The other instrument

TABLE 2.2: Scientific requirements for galaxy clustering and weak lensing observations in Euclid (Laureijs et al., 2011).

Spectroscopic Galaxy Clustering	
Spectroscopic redshift accuracy	$\sigma_z < 0.001(1+z)$
Redsdhift range	$0.7 < z < 2.05$

Photometric Galaxy Clustering and Weak Lensing	
Photometric redshift accuracy	$\sigma_z < 0.05(1+z)$
Redsdhift range	$0 < z < 2$
Error in mean redshift in bin	$\Delta\langle z\rangle < 0.002(1+\langle z\rangle)$
Catastrophic failures	10%
Density of galaxies	> 30 galaxies/arcmin^2

of Euclid, NISP, is composed of a near-infrared photometric channel and a spectroscopic channel. The photometric channel will observe through three near-infrared bands, YJH, covering a wavelength range between 920 and 2000 nm, and reaching a magnitude limit depth of 24 in each band at a signal to noise of 5σ for point sources. The spectroscopic channel of NISP will measure the spectral energy distributions in the wavelength range between 1100 and 2000 nm of more than 50 million galaxies up to a redshift of 2. The NISP focal plane will be composed of a grid of 16 detectors with a resolution of 0.3 arsec for pixel (Racca et al., 2016; Racca et al., 2018).

To achieve the Euclid scientific goals, the telescope is designed to carry out a survey optimized to extract the maximum cosmological information using two main probes: weak gravitational lensing and galaxy clustering (described in Secs. 1.6.1 and 1.6.2, respectively). With its two instruments, Euclid will perform both a spectroscopic and a photometric galaxy survey that will allow us to determine cosmological parameters using its main cosmological probes: galaxy clustering with the spectroscopic sample (GC_s), galaxy clustering with the photometric sample (GC_{ph}), and weak gravitational lensing (WL). The combination of cosmological probes will allow us to measure the expansion rate of the universe and the growth of cosmic structures.

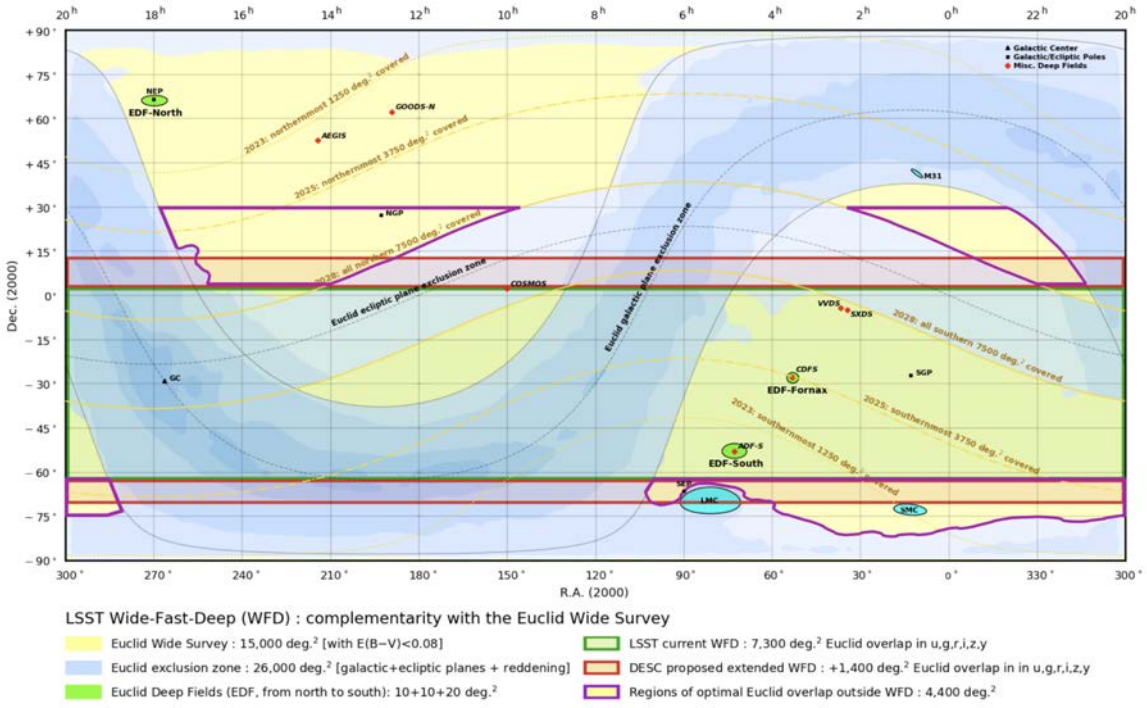

FIGURE 2.5: Euclid observational footprint in comparison to Rubin-LSST survey footprint (Capak et al., 2019).

2.3.1 Combination with ground based surveys

The WL analysis of Euclid data requires an accurate knowledge of the redshift distributions of the samples used for the analysis, as shown in Table 2.2. Euclid photometric data alone cannot reach the necessary photometric redshift performance. Given the specifications of the satellite, the Euclid data should be complemented with ground-based observations to derive precise and accurate photometric redshifts. The satellite would need *griz* bands to achieve an accuracy of $\sigma_z < 0.05(1+z)$, while the addition of the u band would allow one to reach an accuracy of $\sigma_z < 0.03(1+z)$. These bands can be provided by additional ground-based surveys. The combination of Euclid and ground-based surveys can enrich the science exploitation of both. Euclid will provide additional information to ground-based surveys such as very precise shape measurements – thanks to the high spatial resolution of its wide optical band – and near-infrared spectroscopy. Euclid's data will help ground based surveys improve their deblending of faint objects and improve their photometric redshift estimates, which will definitely boost their scientific outcome.

During 6 years, Euclid will cover over 15 000 deg^2 of the extra-galactic sky (wide survey), as well as three patches of the sky of 40 deg^2 in total (deep survey). We show

their footprint in Fig. 2.5. The satellite will spend 10% of the observing time on the deep survey, which will be two magnitudes deeper than the wide survey. The area observed by Euclid will also be covered by ongoing and future ground based imaging surveys.

Several ground based surveys will be needed to cover all the observed area of Euclid, as Euclid covers both celestial hemispheres and those cannot be reached from a single observatory on Earth. The coverage of the sky will be formed of inhomogeneous complementary data. It is very likely that there will be at least three distinct areas in terms of photometric data available. The Southern hemisphere is expected to be covered with the Vera C. Rubin Observatory Legacy Survey of Space and Time (Rubin-LSST; Ivezić et al. 2019) and the already described DES. The Northern hemisphere will be covered by a combination of surveys: the Panoramic Survey Telescope and Rapid Response System[11] (PanSTARRS; Chambers et al. 2016), the Canada-France Imaging Survey[12] (CFIS; Ibata et al. 2017), the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP; Aihara et al. 2018), and the Javalambre-Euclid Deep Imaging Survey (JEDIS). In addition, some area North of the equator may also be covered by Rubin-LSST with shallower limiting magnitude than the one in the Southern hemisphere (see Fig. 2.5).

Rubin-LSST is one of the best complementary surveys to Euclid since it greatly overlaps in area, covers two Euclid Deep Fields and reaches a deeper photometric depth that will lead to better photometric redshift estimation (Rhodes et al., 2017; Capak et al., 2019). Rubin-LSST is expected to start operations in 2022 and during 10 years it will observe over 20 000 deg^2 in the Southern hemisphere with 6 optical bands, *ugrizy*, covering a wavelength range from 320 to 1050 nm. The predicted final magnitude depth for coadded images for point sources detected at the 5σ limit are $u = 26.1$, $g = 27.4$, $r = 27.5$, $i = 26.8$, $z = 26.1$ and $Y = 24.9$ according to the Rubin-LSST design specifications (Ivezić et al., 2019). Among other scientific themes, Rubin-LSST has been designed to study dark matter and dark energy using WL, $\mathrm{GC_{ph}}$, and supernovae as cosmological probes. The Rubin-LSST survey will be the stage-IV ground based experiment providing the best photometry for Euclid-detected galaxies when Euclid data become available.

Another suitable ground-based candidate to cover the optical and near-infrared

[11] `https://panstarrs.stsci.edu`
[12] `http://www.cfht.hawaii.edu/Science/CFIS/`

range in the Southern sky is the DES photometric survey. DES completed observations in 2019 after a 6-years program, so part of the data are already available. As we described in Sec. 2.1, DES has covered 5 000 deg^2 around the Southern galactic cap with 5 broad band filters, *grizy*, with wavelengths ranging from 400 to 1065 nm (Dark Energy Survey Collaboration: Abbott et al., 2016). DES has observed about 300 million galaxies up to redshift 1.4. The coadded magnitude limit depths for 10σ detections in 1.95" diameter apertures are $g = 24.6$, $r = 24.4$, $i = 23.7$, $z = 23.0$. These depths correspond to the published values at the first data release (Dark Energy Survey Collaboration: Abbott et al., 2018) plus 0.3 magnitudes. These values are expected to improve as the analysis of the survey progresses.

Chapter 3

Remapping photometry from real data to simulations

Simulations are an essential tool to work with real data. They are used to create a realistic environment to design surveys, study systematic and selection effects, determine expected errors and biases of observations, and test methodologies of analysis. Therefore, simulations need to properly reproduce the observed properties of galaxies since they are the tracers of the underlying matter distribution of the universe and what we mainly observe in surveys. In addition, simulations also need to be representative of real data in terms of survey area, density and depth. In practice, differences in observable properties arise between simulated and real data due to the complexity of data and the limitations and simplifications of simulations. For example, the flux of each galaxy that is transmitted by the instrument is difficult to determine as it is usually deduced from a combination of laboratory and on-sky measurements that only approximate the real transmission. Discrepancies in observables such as magnitude or flux can lead to a wrong calibration of photometric redshifts (photo-zs) since they are derived directly from galaxy colors.

In this work we focus on the transformation of existing simulations to model real observables that allow us to reproduce photo-z distributions as the ones from real data, since photo-zs are used in a wide range of analyses such as weak lensing or galaxy clustering. Our work aims to reproduce real photometry distributions in simulations by transferring statistical properties of the photometry from real observations to simulations (we call 'remap' this transfer process of properties) while keeping the correlation between photometric bands. We choose to faithfully reproduce photometry in order to recover realistic photo-zs in simulations.

Cosmological surveys use selection criteria to define the optimum galaxy sample to extract the maximum information from observations in order to do science. So the

definition of the galaxy sample is a fundamental step that affects the cosmological analyses. For example, an adequate choice of the galaxy samples to be studied increases the constraining power on cosmological parameters determined from galaxy clustering or weak lensing measurements. A precise and accurate knowledge of the redshift distribution of galaxies is required to perform tomographic analysis in which galaxies are divided into redshift bins. Therefore, in order to ensure that our transformed photometry from the simulation is useful to perform cosmological analyses, we determine the photo-z from the remapped simulated photometry and check if the remapped photometry and resulting redshifts allow us to reliably define the same galaxy samples that have been defined in real data such as the redMaGiC sample (a sample of luminous red galaxies used for example as a lens galaxy sample in several cosmological analyses in DES) and a magnitude-limited sample (a lens sample optimized for galaxy clustering and galaxy-galaxy lensing DES analyses).

We want a simulation with photometry and redshifts like the ones in DES (introduced in Sec. 2.1). Our chosen simulation to resemble DES is MICE (described in Sec. 3.1). Unfortunately, the photometric properties in MICE are slightly different than those measured in DES. So we want to modify the photometry of MICE by transferring real photometric properties from observations of DES to MICE (the specific DES data used in this work is described in Sec. 2.1.2 and corresponds to the data of the first three years of observations denoted as Y3 Data). However, a proper transfer of photometric properties is a complex process.

In Secs. 3.2 and 3.7, we present the two methods we have used to make the simulated photometry resemble the real one. The lens galaxy samples we are trying to reproduce are described in Sec. 3.3. The evaluation of the lens sample with the first remap method is presented in Sec. 3.4. A modification of the first method and the limits of the algorithm are explained in Secs. 3.5 and 3.6, respectively. Finally, we summarize the chapter in Sec. 3.8.

3.1 Simulation data

To develop and asses the methodology to transfer photometric properties from real to simulated data, we use the simulated data from the second version of MICE[1]. In this section we briefly describe the simulation.

[1] `http://maia.ice.cat/mice/`

MICE is a suite of dark matter N-body simulations (Fosalba et al., 2015a) produced in the Marenostrum facilities by the Institut de Ciències de l'Espai team. The MICE Grand Challenge simulation was generated using 4096^3 dark matter particles with a mass of $m_{\rm p} = 2.93 \times 10^{10}\ h^{-1}{\rm M}_{\odot}$ gravitationally interacting in a comoving volume of side 3072 h^{-1}Mpc from an initial redshift of $z = 100$, producing a light cone that covers an octant of the sky. The simulation assumes that the cosmic expansion is given by a standard flat ΛCDM model with input cosmological parameters:

$$\{\Omega_{\rm m},\ \Omega_{\Lambda},\ \Omega_{\rm b},\ h,\ n_{\rm s},\ \sigma_8\} = \{0.25, 0.75, 0.044, 0.7, 0.95, 0.8\}\,. \tag{3.1}$$

MICE identifies dark matter halos using the Friends of Friends (FoF) algorithm and populate them with galaxies using a hybrid of Halo Occupation Distribution (HOD) and Halo Abundance Matching (HAM) technique. In addition, the simulation was built to follow local luminosity function as Blanton et al. (2003) for the bright galaxies, Blanton et al. (2005a) for the faintest galaxies, and the galaxy clustering as a function of luminosity and color as in Zehavi et al. (2011). The final catalog includes a broad range of galaxy, halo, clustering and lensing properties for approximately 200 million galaxies in an area of 5000 deg^2 and reaching redshift $z = 1.4$. The clustering, halo and lensing properties of this catalog have been validated and discussed in Fosalba et al. (2015a), Crocce et al. (2015) and Fosalba et al. (2015b), respectively. For further details on the creation of the mock catalogs we refer the reader to Carretero et al. (2015).

3.2 Remap of photometry using abundance matching

The first process we use to remap photometry consist in sequentially applying a series of abundance matching transformations. First we need to explain in more detail the abundance matching method before describing the full process used to correctly remap photometry. Briefly, the abundance matching method consist in determining a relation of abundances of a certain variable between two data sets. The matching ensures that for each values of the variable, the same abundance is found in both data sets. We are going to use magnitudes g, r, i and z as variables for the abundance matching process. The data sets for which we want to match abundances are MICE and DES. We apply this method over each of the four magnitudes separately to

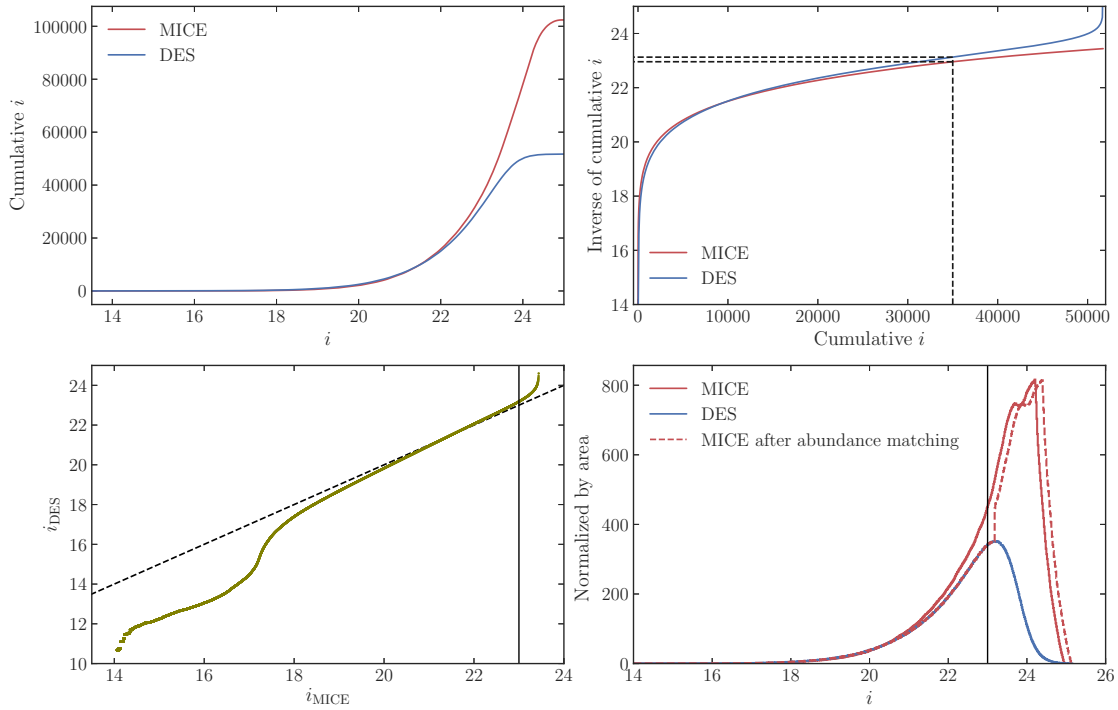

FIGURE 3.1: Steps of the abundance matching technique. *Top left:* Cumulative distribution of magnitude i of MICE (red) and DES (blue). *Top right:* Inverse of the cumulative distribution of i as a function of the cumulative. Black line indicates that for every cumulative (abundance) value, the corresponding values of i (inverse of the cumulative) of MICE and DES are obtained. *Bottom left:* Magnitude i of DES as a function of i of MICE. Every point correspond to the respectively values of i obtained for different values of abundances. This is the function that must be apply over the magnitude of MICE to resemble DES. *Bottom right:* Initial i distribution of MICE (red), DES (blue) and the resulting i distribution of MICE after the abundance matching transformation is applied (dashed red).

transform the magnitudes of MICE to resemble the same distribution and abundance of DES magnitudes.

3.2.1 Abundance matching technique

To understand this technique, we show some of the steps of the remap through abundance matching of magnitude i in Fig. 3.1. In the first step of abundance matching, the cumulative distribution of the variable we want to match is computed for the two data sets. In this case, we compute the cumulative distribution of magnitude i for MICE and DES (top left panel of Fig. 3.1). In the second step, the cumulative distribution is inverted (top right panel of Fig. 3.1). So for each cumulative value of

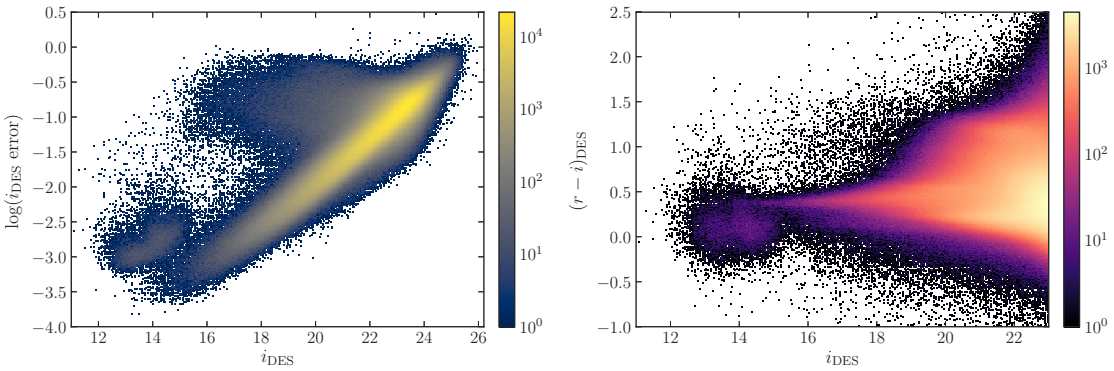

FIGURE 3.2: Distributions we try to reproduce with MICE. *Left:* Error of i as a function of i of DES. *Right:* Color $r-i$ as a function of i of DES.

magnitude i, i.e. abundance of i, we find the corresponding values of the inverse of the cumulative for MICE and DES (this step is represented in the dashed black lines of the top right panel of Fig. 3.1). Since the inverse of the cumulative of magnitude i corresponds to the magnitude i itself, we find a relation between the magnitude i of both data sets so that for each value of magnitude i both data sets have the same abundance. The relation between values of magnitude i that have the same abundance is shown in the bottom left panel of Fig. 3.1. This relation indicates the transformation that must be applied to the magnitude i of MICE to resemble the DES one. Once the magnitude i of MICE is transformed, the distribution of magnitude i becomes the same as in DES as shown in the bottom right panel of Fig. 3.1. We only apply the abundance matching until $i < 23$ (indicated with a black line) because after the transformation we cut objects fainter than that since the i magnitude limit at which DES detects objects of five signal-to-noise ratio is close to 23 (this is the limiting magnitude for observations of the first three years of DES. The limiting magnitude from the whole six year observations is expected to be deeper) and we want to avoid incompleteness.

3.2.2 Abundance matching applied to the remap

The complete process to transform all magnitudes of MICE to resemble DES consist of several steps. For clarity we itemize and explain them in detail. The list is split in four main items that correspond to the steps of the transformation of each individual magnitude. To make the notation clear, we indicate if a magnitude comes from MICE or DES with the respective name as a subscript label. The transformation of each

magnitude consist on a first abundance matching process, then a noise realization over the transformed magnitude, and finally a second abundance matching process. To indicate that a magnitude or color is the result of a first abundance matching applied to it, it will have an apostrophe mark. If a magnitude is the product of a noisy realization, it will have an asterisk as superscript. The label 'remap' without any punctuation mark denotes that a magnitude is the result of the second abundance matching process and thus the final transformed magnitude.

1. In the first step, the i magnitude of MICE ($i_{\rm MICE}$) is modified to have the same distribution as the DES one ($i_{\rm DES}$). The transformation is composed of three stages.

   (a) An abundance matching between $i_{\rm MICE}$ and $i_{\rm DES}$ is performed as described in Sec. 3.2.1, thus transforming $i_{\rm MICE}$ to match the analogous distribution of DES. The transformed magnitude of MICE is denoted as $i'_{\rm remap}$.

   (b) We want the remapped photometry of MICE to carry the same intrinsic uncertainty that DES observations have. So we generate noisy realizations of the remapped i magnitude:

   $$i^*_{\rm remap} = i'_{\rm remap} + N\,(\mu = 0,\, \sigma = \text{error}) \tag{3.2}$$

   where N is a random number from a normal distribution. For each value of $i'_{\rm remap}$ we look at the distribution of $i_{\rm DES}$ and its error (shown in the left panel of Fig. 3.2) and take the distribution of all the values of error that correspond to a magnitude value of $i'_{\rm remap}$. To determine the σ of the normal distribution for the $i'_{\rm remap}$ magnitude, we take the corresponding distribution of errors and perform an accept-reject algorithm to obtain a sampling of the error distribution that will be assigned to σ. The accept-reject method is a sampling technique that allows to obtain samples from a given distribution (for a detailed description see e.g. Neal 2003).

   (c) Given the addition of noise to $i'_{\rm remap}$, the distribution of the magnitude has slightly changed. To ensure the abundance for each value of i remains the same between the remapped magnitude and DES, we perform a second abundance matching between $i^*_{\rm remap}$ and $i_{\rm DES}$. The resulting magnitude from the second transformation with abundance matching is the one we keep. We call it $i_{\rm remap}$.

2. In the second step, we modify the r magnitude. We apply the same sequence of abundance matching as with i but instead of transforming r directly we transform $r-i$. Transforming $r-i$ forces the correlation between both magnitudes to remain. When we tried to remap each magnitude separately and without taking into account their correlation, the transformed distribution of the magnitudes were similar to DES but the correlations between the magnitudes were lost. Remember that magnitudes are discrete points of the spectral energy distribution, hence the relation between magnitudes is used to determine the photo-z. If the relation between magnitudes is lost for every objects, this will lead to a wrong photo-z. To further keep the link between r and i, the abundance matching is applied over bins of i magnitude. We take the relation $(r-i)_{\rm DES}$ as a function of $i_{\rm DES}$ (shown in the right panel of Fig. 3.2) and $r_{\rm MICE}-i_{\rm remap}$ as a function of $i_{\rm remap}$. Note that we use the already remapped i magnitude since we want to modify $r_{\rm MICE}$ keeping the correlation with $i_{\rm remap}$. We follow the same a-c stages as in step 1.

   (a) The distributions $(r-i)_{\rm DES}$ and $r_{\rm MICE}-i_{\rm remap}$ are split in bins of $i_{\rm DES}$ and $i_{\rm remap}$, respectively. For every bin of i an abundance matching is performed between both data sets, transforming $r_{\rm MICE}-i_{\rm remap}$ into $(r-i)'_{\rm remap}$ which resembles the equivalent distribution of DES.

   (b) We add realistic photometric error to r. First, we compute the remapped r from the remapped color: $r'_{\rm remap}=i_{\rm remap}+(r-i)'_{\rm remap}$. Then we generate a noisy realization of $r'_{\rm remap}$ by adding a random number from a normal distribution, as done in step 1b for i, with the σ of the normal distribution depending the error distribution of $r_{\rm DES}$. Then we calculate the noisy magnitude $r^*_{\rm remap}$.

   (c) To ensure that noisy $r^*_{\rm remap}$ follows the same distribution as $r_{\rm DES}$, we perform a second abundance matching, this time between $r^*_{\rm remap}-i_{\rm remap}$ and $(r-i)_{\rm DES}$. Obtaining a $(r-i)_{\rm remap}$ distribution similar to the homologous one in DES. Finally, we compute the remapped r magnitude: $r_{\rm remap}=i_{\rm remap}+(r-i)_{\rm remap}$. We keep the resulting magnitude $r_{\rm remap}$.

3. Next we remap the g magnitude. As explained in the previous step, we need that correlations between magnitudes to remain. So instead of directly transforming the g magnitude, we perform an abundance matching over the $g-r$ color splitting the data sets in bins of $r-i$ and i. We use $(r-i)_{\rm remap}$ and $i_{\rm remap}$ to

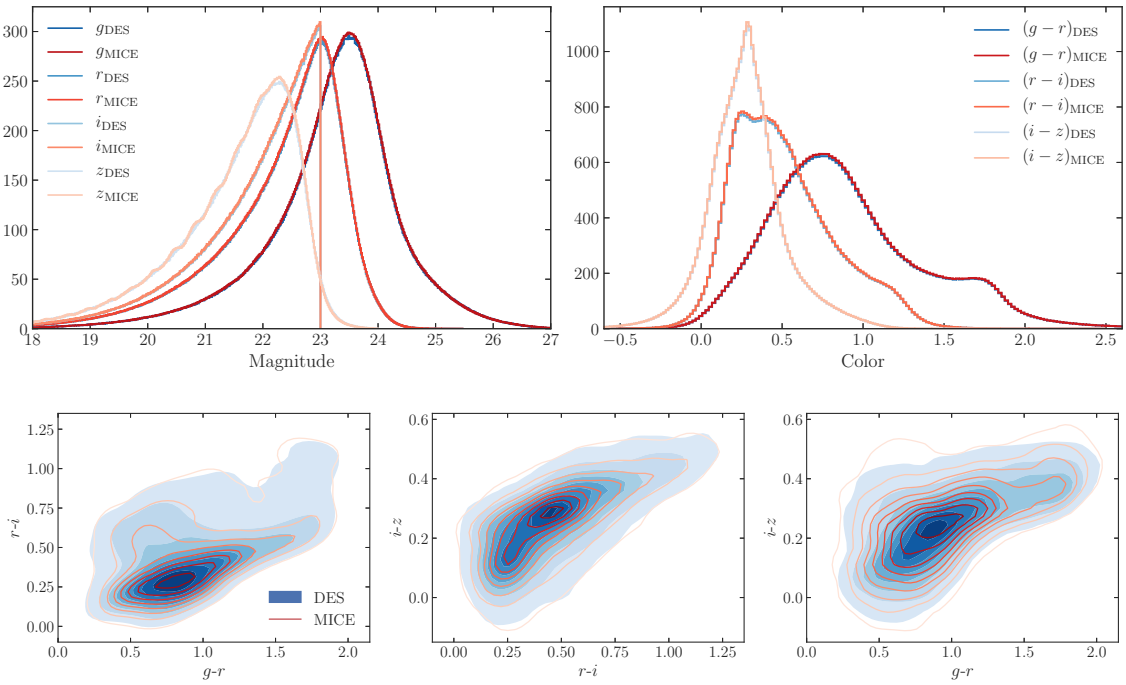

FIGURE 3.3: *Top:* Comparison of g, r, i, z magnitudes (left) and $g-r$, $r-i$, $i-z$ colors (right) distributions of DES (blue color scheme) and the remapped magnitudes and colors of MICE (red color scheme). Both sets of real and simulated distributions are practically identical and overlap in the plot. *Bottom:* Color-color diagram of DES (blue shades) and remapped MICE (red lines). The remapped magnitudes and colors of MICE have been obtained through the abundance matching process.

split the MICE data, $(r-i)_{\rm DES}$ and $i_{\rm DES}$ to split the DES data. We apply the abundance matching over $g_{\rm MICE}-r_{\rm remap}$ to resemble the homologous distribution of DES. To obtain $g_{\rm remap}$, we follow the same a-c steps as we have done for r. We do not explain them in detail to avoid repetition.

4. Finally, we remap the z magnitude. Once again, to keep correlations between magnitudes, the abundance matching is applied over the $i-z$ color splitting the data into bins of $r-i$ and i. The MICE data are split using $(r-i)_{\rm remap}$ and $i_{\rm remap}$. The abundance matching is perform over $i_{\rm remap}-z_{\rm MICE}$ to transform the distribution to resemble the equivalent one in DES. The same a-c steps to get $r_{\rm remap}$ are followed to obtain $z_{\rm remap}$.

The remap of photometry using abundance matching to transfer properties of magnitudes and colors from real data to simulations allows us to recover the overall distribution of the mentioned variables in simulations as shown in the top panels of

Fig. 3.3. This process of remapping not only reproduces the distribution of magnitudes but also the correlation between colors. As seen in the bottom panels of Fig. 3.3, the color-color diagrams of DES and remapped MICE are quite similar but not equal. The differences in the color-color diagram are specially noticeable in the $i-z$ versus $g-r$ given that the abundance matching for both distributions did not consider the correlation with each other. For example, we applied the abundance matching over $i-z$ taking into account the relation with $r-i$ and i but not with $g-r$. The computational time was already large when we applied the abundance matching in bins of $r-i$ and i, so adding an extra dimension by also binning $g-r$ would be computationally expensive. Overall the remapped properties follow the same distributions as the ones in DES. However, the correlations not taken into account in the remapping technique are not guarantee to work.

3.2.3 Photometric redshift determination

Magnitudes are a measure of the brightness of an object in a given band, and thus they describe the spectral energy distribution at a certain wavelength. The information of the spectral energy distribution is used to determine the photo-z. Therefore, magnitudes must keep the correlation between them in order to correctly define the spectral energy distribution of an object. We have reproduced the overall magnitude and color distributions of DES in MICE through the remapping process using abundance matching. We need to asses if the correlation between magnitudes are physical, in other words, if their relation correctly defines the spectral energy distribution of galaxies and thus photo-zs can be obtained. To that end, we use the remapped magnitudes as input variables to determine the photo-z.

We use the Directional Neighborhood Fitting (DNF; De Vicente et al. 2016) training-based algorithm to estimate photo-zs. DNF is one of the algorithms used to compute photo-z in DES. DNF photo-zs are available in the catalog of objects of the Y3 Data of DES, which are the same objects we use to remap the photometry in MICE. Using DNF will allow us to compare the redshifts of DES with the one obtained with the remapped MICE photometry.

For completeness, we briefly summarize the algorithm here. DNF estimates the photo-z of a galaxy based on its closeness in the space defined by magnitudes, to a set of training galaxies whose redshifts are known. The key feature of DNF is that the metric that defines the distance between objects is given by a directional

neighborhood metric which is the product of an Euclidean and angular neighborhood metrics. The algorithm returns two photo-z estimators: $z_{\rm mean}$ that is the result of fitting a linear adjustment to the directional neighborhood of a galaxy and $z_{\rm mc}$ that is an estimate of the photo-z probability distribution function from the nearest neighbor. When working with tomographic bins, galaxies are classified into different bins using their $z_{\rm mean}$ and the photometric distribution, $n(z)$, within each bin is obtained by stacking their $z_{\rm mc}$. This is an approach used by DES in analyses of their First Year Data results providing redshift distributions that were validated with other independent assessment methods. Within the First Year Data results, DNF has been used in analyses such as the derivation and validation of redshift distribution estimates of galaxies used as weak lensing sources (Hoyle et al., 2018), the calibration of photo-z biases of photo-z methods in cross-correlation clustering-based methods (Gatti et al., 2018), and to select a sample dominated by luminous red galaxies and optimized to measure baryon acoustic oscillations (Crocce et al., 2019b).

As a training sample for DNF we consider the same sample used to run DNF in the Y3 Data. This sample is a compilation of spectra from spectroscopic surveys that overlap with the DES footprint (see Gschwend et al. (2018) for a detailed list) and whose objects were matched to DES using a matching radius of 1 arcsec. From the spectra compilation, VIPERS was removed alongside half of objects randomly selected to be used as validation sample.

Due to a different color selection cuts and magnitude limit depth of some of the surveys, there are discontinuities in the magnitude-redshift space and the combined spectroscopic redshift distribution have two peaks that cause biases towards those overrepresented redshift when estimating the photo-z (Zhou et al., 2020).

The spectroscopic redshift distribution for the training sample is shown in black in the right panel of Fig. 3.4. Running DNF we observed that the only way to recover a similar redshift distribution as DES was using the same training sample. We tested several samples, with different $n(z)$, photometric quality and completeness in the color-redshift space. We found that the training sample has a great impact in the photo-z determination and using the same as in DES is key to recover the same $n(z)$.

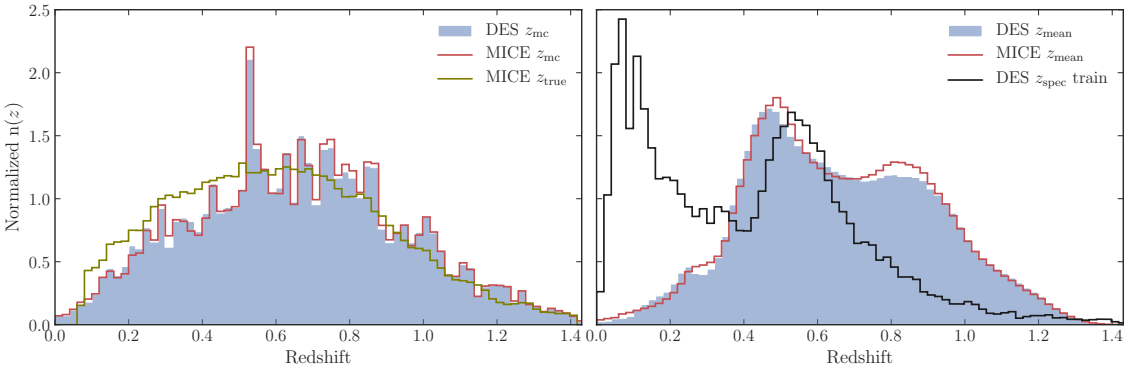

FIGURE 3.4: Photo-z distribution obtained through DNF using the remapped photometry of MICE (red line) compared to the distribution of photo-z of DES (shaded blue) from the Y3 Data catalog. *Left:* $z_{\rm mc}$ distribution of MICE and DES, and true redshift distribution of MICE (green line). *Right:* $z_{\rm mean}$ distribution of MICE and DES, and spectroscopic redshift distribution of the DES training sample used to train DNF (black line).

3.2.4 Photometric redshifts for remapped objects with abundance matching

Photo-z for the remapped photometry is determined using DNF as described above. In Fig. 3.4 we present the resulting photo-z distributions of the remapped MICE objects and compare them with the ones from the Y3 Data catalog of DES obtained with the same code. In the left panel, we see a good agreement between the MICE and DES $z_{\rm mc}$ distributions. However, there are noticeable differences with respect to the true redshift of MICE. This discrepancy will have an adverse effect in the definition of galaxy samples as we will show later. In addition, in the right panel, there are small differences between the MICE and DES $z_{\rm mean}$ distributions that will also translate into mismatches when trying to reproduce the same galaxy sample selection in MICE and DES.

3.3 Sample selection

As we mentioned in the introduction of this chapter, the precise definition and selection of a sample (number of galaxies, errors in the photo-z determination, etc.) is the main aspect that affects the performance of the analyses from cosmological probes such as weak lensing and galaxy clustering. For example, efforts have been done to optimize the sample selection to extract the maximum information from clustering

measurements to constrain cosmological parameters (Tanoglidis et al., 2020), to optimize the measurement of baryon acoustic oscillations (Crocce et al., 2019b) and to optimize the lens sample for galaxy clustering and galaxy-galaxy lensing analyses (Porredon et al., 2020).

Performing the sample selection in simulations allows us to define, test and improve the sample selection, as well as to asses the error and bias of the selection. We want the remapped photometry of MICE to be as similar as possible to DES to be able to reproduce the same sample selections that will be performed in the survey. We check whether the remapped photometry is useful for sample selection by applying the same selection criteria on MICE. We focus on the selection of two samples: RedMaGiC and a magnitude-limited sample.

3.3.1 RedMaGiC selection

Luminous red galaxies (LRG) are frequently selected as a type of galaxy sample. These galaxies have a bright intrinsic luminosity, occupy a narrow range of colors, are the most massive galaxies in $z < 1$ and have a strong 4000Å break in their spectral energy distribution (Eisenstein et al., 2001). Their characteristics make them easy to detect and the sharp break allows to accurately infer their photo-zs. However, at high redshift, the 4000 Å break is redshifted out of the optical bands due to the cosmic expansion, so LRG are more difficult to detect (Prakash et al., 2015). In addition, there are fewer LRG at high redshift due to galaxy evolution. Therefore, LRG form a sample with high photo-z accuracy but only at low redshift, becoming a low density sample at higher redshift.

LRG reside in massive dark matter halos and cluster strongly. So, they are commonly used as clustering sample to measure, for example, redshift space distortions and the scale imprinted by baryon acoustic oscillations. Several projects of DES have used a LRG subsample of the survey for clustering measurements defined by the red sequence Matched filter Galaxy Catalog (redMaGiC; Rozo et al. 2016) algorithm. This LRG subsample is called redMaGiC sample. In DES, for example, the redMaGiC sample has been used as a tracer sample of the underlying mass for the joint analysis of galaxy clustering, galaxy lensing and cosmic microwave background lensing two-point functions (Abbott et al., 2019). The redMaGiC sample has also been used to measure the clustering of DES galaxies that will be combined with weak lensing samples to obtain accurate cosmological constraints from the large-scale

structure and lensing correlations (Elvin-Poole et al., 2018). It has also been considered as a tracer sample to be cross-correlated with the weak lensing source sample to estimate the redshift distribution and photometric bias of the source sample (Gatti et al., 2018).

Since redMaGiC is one of the samples widely used in DES, it is important that the redMaGiC algorithm can extract a LRG sample from the remapped photometry of MICE similar to the one obtained with DES. In order to understand the behavior of a redMaGiC sample selected from the MICE remapped photometry, we briefly explain how the algorithm works.

RedMaGiC selects a subsample of LRG with constant comoving density and lower photo-z uncertainty. As explained in more detail in Rozo et al. (2016), the algorithm first fits a red sequence template to each galaxy of the sample. The templates are defined by modeling the colors as a function of redshift and magnitude. The modeling of templates is produced by redMaPPer as described in Rykoff et al. (2016). Then, redMaGiC determines the goodness of the template fit and the corresponding best fitting photo-z. Given the photo-z, the algorithm computes the corresponding galaxy luminosity. If the resulting luminosity is bright and the goodness of fit is good enough, the galaxy is considered a LRG. The goodness of the fit should be smaller than a certain value that varies with redshift in order to keep constant the sample comoving density. For a galaxy to be considered a LRG, its luminosity should be brighter than a luminosity threshold that the algorithm determines from the colors of the input data. Therefore, the remapped colors must be coherent. In addition, redMaGiC removes the biases in its photo-z estimation using spectroscopic redshifts if available. Thus, the true redshift of our simulation should be properly correlated with the remapped colors. To sum up, if redMaGiC is able to get a subsample of LRG from our remapped sample, it would imply that the resulting photometry and redshift are well correlated and defined.

3.3.2 Magnitude-limited sample

Another option to define a sample is to select galaxies according to a flux or magnitude limit that depends linearly with the photo-z. Magnitude limited samples have a larger number density and reach higher redshift than the redMaGiC sample, but at the expenses of having slightly larger photometric uncertainties.

TABLE 3.1: Selection criteria and redshift bins for the magnitude-limited sample defined in Porredon et al. (2020). There are two redshift bins versions since the definition of the redshift range of the bins changed along the development of this work.

Concept		Selection
Flux selection		$i < 4 * z_{\mathrm{mean}} + 18$
Remove most luminous objects		$i > 17.5$
Bin	Redshift range v2.1	Redshift range v2.2
1st	0.20 - 0.35	0.20 - 0.40
2nd	0.35 - 0.50	0.40 - 0.55
3rd	0.50 - 0.65	0.55 - 0.70
4th	0.65 - 0.80	0.70 - 0.85
5th	0.80 - 0.95	0.85 - 0.95
6th	0.95 - 1.05	0.95 - 1.05

In analyses of DES, a magnitude-limited sample have been used to optimize the clustering measurements (Crocce et al., 2016). To reduce the uncertainty in the measurements, a large area and a high number of objects across all the redshift range are ideally needed. However, the same magnitude depth is not achieved in the whole survey area. To ensure the completeness of the galaxy sample a selection based in the magnitude-limit is applied and only areas that achieve that magnitude limit are kept for clustering measurement. Another magnitude-limited sample selection has been performed in DES to obtain a sample optimized to measure the baryon acoustic oscillations signal. It mainly contained red galaxies with a good photo-z accuracy but was expanded with more objects to have a higher number density than the redMaGiC sample (Crocce et al., 2019b).

Another magnitude-limited sample, defined in Porredon et al. (2020), was designed to optimize the lens sample to have better cosmological constrains in the combination of galaxy clustering and galaxy-galaxy lensing analysis for the Y3 Data of DES. In their defined optimum selection they apply a magnitude cut in the i band that depends linearly with the photo-z. The sample is then divided into six predefined tomographic bins. They remove all possible stars and mask regions in which the coaddition process produces a discontinuous point spread function (PSF) or fails to recover observations in one band, remaining a total area of 4203.5 deg^2.

The selection criteria for the magnitude-limited sample and the tomographic bins are summarized in Table 3.1. We will apply the same selection criteria on the

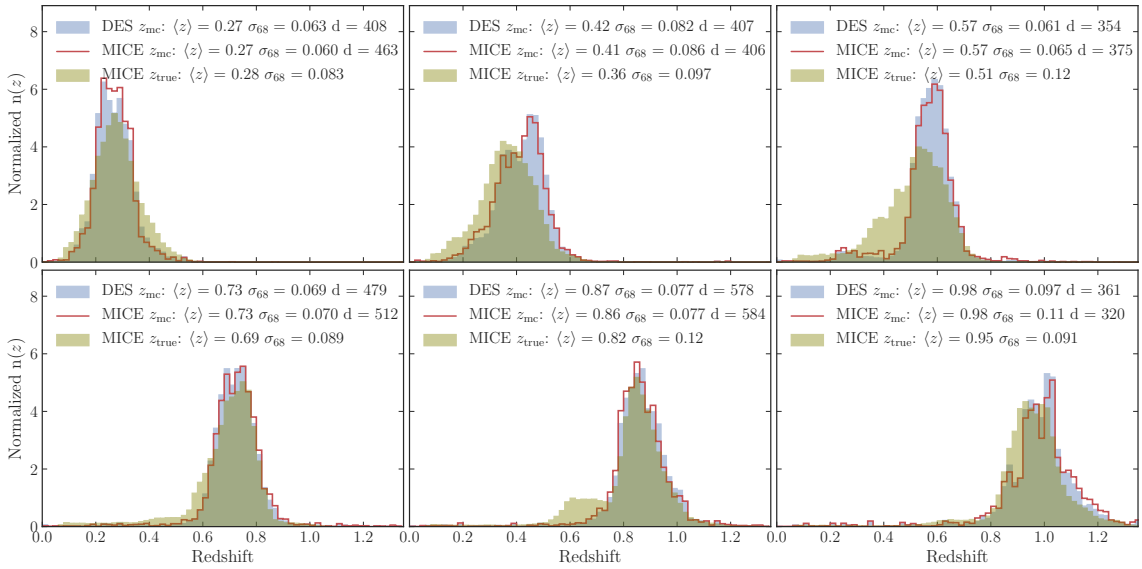

FIGURE 3.5: Photo-z distributions of DES (blue), remapped MICE (red) and true redshift of MICE (green) resulting from applying the magnitude-limited sample selection criteria and splitting the sample in redshift bins (version 2.1) as defined in Table 3.1. $z_{\rm mean}$ is used to split the magnitude-limited sample in redshift bins. On each plot, the mean $\langle z\rangle$ and the 68% confidence widths σ_{68} of every distribution are shown. In addition to the galaxy density d of the remapped MICE and DES.

remapped photometry and photo-z of MICE. Note that they use the photo-zs obtained with DNF, the same code we have run over the remapped photometry. Thus the comparison between their magnitude-limited sample and the one we will obtain with the remapped MICE will be easy to compare and will allow us to check if the remapped process preserves the spectral energy distributions of the galaxies.

3.4 Evaluating the magnitude-limited sample selection

The photometry of MICE has been remapped to resemble DES through a series of abundance matching processes over the magnitudes and colors as described in Sec. 3.2. The photo-zs of the remapped photometry have been determined using DNF. Now we apply the same selection criteria as in DES to obtain a limited-magnitude sample using the remapped photometry and redshift of MICE. We will asses whether the resulting sample has the same properties as the DES Y3 Data. To compare both magnitude-limited samples, the resulting redshift distributions for every redshift

bin defined in Table 3.1 are shown in Fig. 3.5. There we see that the overall $z_{\rm mc}$ distribution of DES and the remapped MICE have small differences. However, the true redshift distribution of MICE differs considerably from the photo-z distributions. The discrepancy is specially worse in the second, third and fifth bins where redshift tails are distinguishable. On top of the mismatch between distributions, there is another issue. Taking a look at the galaxy surface density d (number of galaxies divided by area) of DES and the remapped MICE at every redshift bin, there are big differences in density between both data sets. The two mentioned issues indicate that the redshift distribution of DES and the remapped MICE are different. Therefore, it is necessary to modify the methodology of abundance matching to remap MICE to resemble DES.

3.5 Inclusion of redshift binning in the remap process

We have transformed the MICE magnitudes to resemble the DES ones by establishing a relation between the magnitudes and colors of DES and MICE with the abundance matching method described in Sec. 3.2. As explained in the previous section, the abundance matching method to transform magnitudes does not fully keep the correlations between magnitudes and redshift. Magnitudes and colors describe the spectral energy distribution of galaxies and are used to determine their photo-zs. The remap process does not properly maintain the color correlations and that produces photo-z distributions different from the ones from observations from DES. Causing that the redshift distribution of the magnitude-limited sample of the remapped MICE differs from the DES one and that galaxies are placed at the wrong redshift bin.

In order to correct this issue, we modify the remapping method. We check if including the redshift as a variable during the remap process helps to improve the transformation. In the remap, in order to keep correlations among colors, we performed the abundance matching binning the data as a function of the color we wanted to remap, the adjacent color in the spectral energy distribution, and the i magnitude. To include the redshift as a variable, we split the data into redshift bins before performing the abundance matching. That way we split beforehand the colors according to their redshift, and thus the transformation of magnitudes is performed in groups

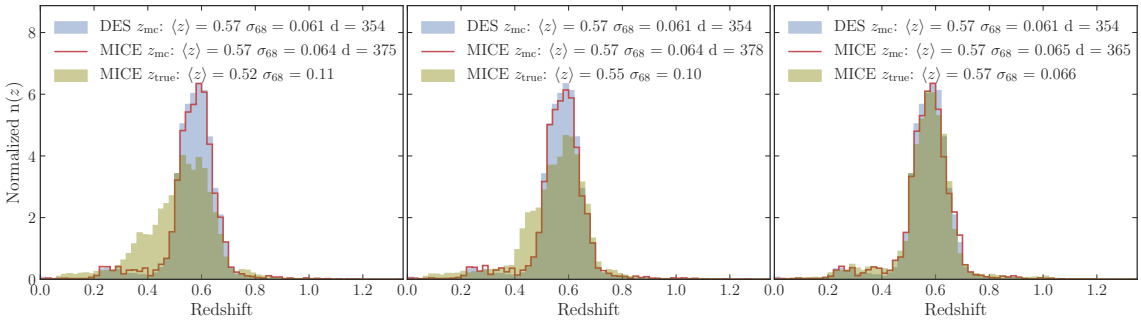

FIGURE 3.6: Photo-z distribution of DES (blue), remapped MICE (red) and true redshift of MICE (green) product of the magnitude-limited sample selection. Distributions correspond to the redshift range of the third bin defined in Table 3.1 (version 2.1). The number of redshift bins used to split the data in the remap of MICE varies in each plot. Two $(0.07-0.75,\ 0.75-1.42)$ and four $(0.07-0.41,\ 0.41-0.75,\ 0.75-1.08,\ 1.08-1.42)$ bins have been used in the left and center panel, respectively. In the right panel, a binning of 0.05 redshift width has been used.

of colors related to a certain redshift range. To split the data in redshift bins, the true redshift is used in MICE and the photo-z $z_{\rm mean}$ is used in DES.

Since the whole remap process is computationally expensive, we try to split the data in two and four redshift bins in order to see if the redshift distribution of the magnitude-limited sample improves. The full remap algorithm is not coded in parallel, so for every redshift bin all the remap process is computed, making the code time consuming. In the left and central panels of Fig. 3.6, we present the resulting redshift distributions when using two and four redshift bins to split the data to remap MICE (we will talk about the last panel later). It can be appreciated that the true redshift distribution of MICE becomes closer to the photo-z distributions of DES when increasing the number of redshift bins.

We consider that binning the data sets in four redshift bins produces redshift distributions of the magnitude-limited sample of remapped MICE that resemble enough the redshift distributions of DES while keeping the computational time of the remap process reasonable. So we run the redMaGiC algorithm over the remapped MICE catalog to see if we are able to recover the redMaGiC sample in the remapped photometry.

3.5.1 Evaluating the redMaGiC sample selection

The photometry of MICE has been transformed to resemble DES using the abundance matching method by previously splitting both datasets in four redshift bins to

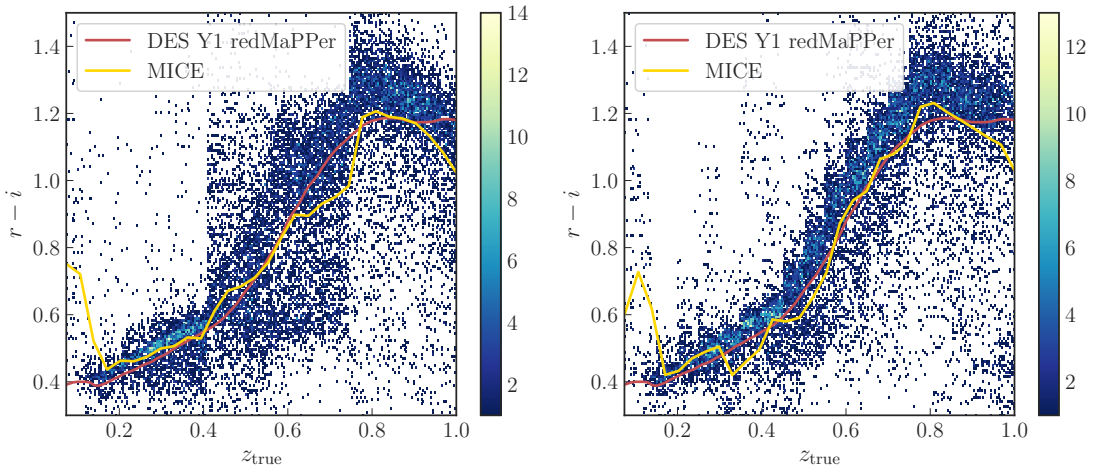

FIGURE 3.7: Density histogram of $r-i$ color of the remapped photometry as a function of the true redshift of MICE, only considering massive halos. Four redshift bins (left panel) and a binning of 0.05 redshift width (right panel) have been used to split MICE and DES data before performing abundance matching over photometry. The mean color value across redshift of remapped MICE is shown in yellow. In red, we show the mean color across redshift of the redMaPPer sample from the First Year Data of DES as presented in DeRose et al. (2019).

establish a remap of colors with dependence on the redshift. Now, we want the run the redMaGiC algorithm on the remapped photometry to asses whether the remapped colors and photo-z allow to recover a similar redMaGiC sample as the DES one.

As we mentioned in Sec. 3.3.1, the redMaGiC algorithm selects a sample of LRG with constant comoving density and lower photo-z uncertainty by establishing a selection threshold of luminosity that depends on the input colors and takes into account redshift. Therefore, the remapped colors and redshift must be well correlated for this selection to be effective. However, when we tried to run the redMaGiC algorithm, the scatter of colors as a function of redshift was too large to successfully detect a red-sequence galaxy sample with low photo-z uncertainty. Inside the redMaGiC algorithm there is another algorithm called redMaPPer algorithm, which optimizes the sample selection to detect clusters with accurate photo-z (also known as redMaPPer sample). The problem to successfully run the redMaGiC algorithm was in the part of the redMaPPer algorithm. Therefore, we take a closer look at the scatter of colors as a function of redshift of the remapped MICE only considering massive halos. As seen in the left panel of Fig. 3.7, the scatter of color is, indeed, too large and the redshift bins used to split the data to perform the abundance matching are too noticeable.

To visually detect the amount of scatter, we compute the average value of color as a function of redshift of massive halos and plot it in the same figure (yellow line). We compare the scatter with the one obtained from the redMaPPer cluster sample of DES (red line) as shown in DeRose et al. (2019). We see there are differences between both scatters. Therefore we must improve our remap method.

To improve the remap and thus reduce the color scatter, we increase the number of redshift bins we use to split MICE and DES data and then perform the abundance matching process. Recall that to perform the abundance matching, we also split the data in bins of colors and magnitude to keep the correlation between these variables. We need data in every bin of redshift, magnitude and color to use the abundance matching method. So bins should not be too narrow to have enough objects in each bin. At the same time, if the binning is too sparse, a single bin will represent a larger range of redshift and color producing a remap with too much scatter in these variables. A binning of 0.05 in redshift width is the narrower redshift bin we are able to achieve while having enough data in each bin.

We redo the remap of MICE magnitudes as explained in Sec. 3.2 but now using a redshift binning of 0.05 width. We determine the photo-z through DNF and check if we are able to recover the magnitude-limited and redMaGiC samples. In the right panel of Fig. 3.6, we show the photometric and true redshift distributions of the remapped MICE in comparison to the photo-z distribution of DES for the third redshift bin of the magnitude-limited sample. We see that the three distributions are really close. The tails and mismatches of the true redshift distribution have disappeared when using more redshift bins in comparison to only using two or four bins. In the right panel of Fig. 3.7, we show the density of the $r - i$ color as a function of the true redshift of remapped MICE for massive halos when using more redshift bins in the remap process. In comparison to the case with only four bins, we see the color scatter has decreases but the redshift binning used for the remap is still perceptible. The mean value of the color along redshift is closer to the mean value obtained from the redMaPPer cluster sample of DES. So, overall, the resulting photometry and samples for this method are satisfactory enough.

3.6 Limitations of the algorithm implementation

Up until now, the process of remapping has been tested and applied on a MICE patch of about 900 deg^2. Now we are satisfied with the resulting photometry and photo-zs and so we want to transform the full MICE area, about 5000 deg^2, to resemble the DES data. Recall that we split the data in redshift bins and then the abundance matching code is run for every redshift bin following the process described in Sec. 3.2. The code is not optimized to be run over each subsample of redshift bin in parallel because running the code over a single subsample of redshift bin already takes on the full memory capacity (16 GB) of the computer, even when using data from just a small patch of MICE. In addition, to remap a single redshift bin already takes a minimum of thirty minutes. Considering that we split the data into redshift bins of 0.05 width and the redshift of MICE goes up to 1.4, it would take minimum thirty hours to remap the full area of MICE. If we wanted to increase the number of redshift bins to reduce even more the scatter between the colors and redshift, it would take even longer. So we tried to optimize the code by transferring it to a data platform at the Port d'Informació Científic[2] (PIC), a computer data center that provides technical support to develop projects that require strong computing resources for storage and analysis of large amounts of data. We team up with PIC to improve the code and to be able to use it with larger amounts of data.

After several implementations, we realized that the code can not be applied directly to other pairs of simulations and real data. We had to manually determine the bins of colors and magnitudes to have enough objects in each bin while trying to have narrow bins in the parts of colors and magnitudes distributions with more objects in order to be able to accurately reproduce the variables distributions. The binning needs a lot of manual tweaking and we have implemented it specifically to remap MICE and DES. We though a general and fast method to transform simulated data to resemble real one would be more useful that the method we have implemented here. That is why we explore another method presented in the following section.

[2]https://www.pic.es/

FIGURE 3.8: Source: Pitié et al. (2005). Example of color palette transfer. Original picture (left) to be recolored to match the palette of the source picture (center). The original picture with the transferred color palette from the source picture is the picture on the right.

3.7 Remap of photometry with N-Dimensional pdf Transfer Function

We were looking for an alternative method, to transform the observables of simulations to resemble real data, that was computationally faster than the abundance matching process and easily applicable to any simulation and real data sets. We found the method called N-dimensional pdf transfer function.

In this section, we present and use the N-dimensional pdf transfer function, which determines a continuous transformation that maps the N-dimensional probability density function distributions of observables from real data to simulations. The method was originally proposed by Pitié et al. (2005) as a process to map the N-dimensional pdf to transfer the color palette from a target picture to another picture. We show an example of color palette transfer between pictures in Fig. 3.8. In this work, we apply the pdf transfer function in a novel way to reproduce the statistics of magnitudes and colors from surveys such as DES in the MICE simulation. We use the pdf transfer function to map simultaneously the set of probability density distribution of magnitudes in order to keep correlations among them. Simultaneously mapping all magnitudes allows to preserve the correlations between them. As we have seen in this chapter, this is very important since magnitudes define the spectral energy distribution of objects and thus are used to estimate the photo-z. Then, if the correlations between magnitudes are lost for every objects, this will lead to a wrong photo-z estimate.

We explain the steps of the N-dimensional pdf transfer function algorithm as described in Pitié et al. (2005). To remap the simulated data to resemble the real one, the pdf transfer function algorithm takes two data sets: the simulation or original

data set, denoted as X, and the real or target data set, Y. Each object of the data set, x_j, has an associated vector with the magnitudes to be remapped as components, $x_j = (g_j, r_j, i_j, z_j)$. The dimension of the vector (four in our case) defines the number of dimensions, N, of the remap problem. The goal of the algorithm is to find a continuous mapping function, t, that transform the pdf of X, denoted by $f(x)$, in the pdf of Y, $g(y)$. To find the mapping function t, the algorithm iteratively and randomly rotates both samples X and Y, and projects them onto all axes n, in order to reduce the dimension of the problem to one dimension. Then for every axis, the algorithm matches both cumulative distributions to find the one-dimensional pdf transformation t_n:

$$t(x) = C_Y^{-1}(C_X(x)) \tag{3.3}$$

where C_X and C_Y are the cumulative pdfs of X and Y, respectively. This remap is similar to the abundance matching process described in Sec. 3.2. The mapping function t_n is applied over the projected simulated data in the axis n and transforms the simulated data. The transformed data now has magnitude distributions closer to the real data. To completely reproduce the real distributions in the simulated data, the algorithms iteratively rotates both samples X and Y, repeats the projection of the distributions in every axis and applies the pdf transformations t obtained in each projection over the simulated data. The iterations continue until both the transformed simulated distribution and the real data distributions are identical. The mathematical proof of converge of the algorithm is explained in detail in Pitié et al. (2005). In Fig. 3.9, we show a two dimensional example of how the original distributions change at each iteration of the algorithm until they resemble the target distributions.

Here, we describe the steps of the full process we use to remap the magnitude distributions of MICE into DES data.

1. When mapping magnitudes from real to simulated data is important to keep the correlations between magnitudes in order to obtain a correct photo-z. In imaging surveys, the photo-z is inferred from observations of the spectral energy distribution integrated over a few filter bands. But degeneracies may arise between objects that have similar colors and magnitudes but different redshift. Since the redshifts of DES are photometric, in order to reduce degeneracies, we split the MICE and DES data in bins of redshift before applying the pdf transfer function, as in Sec. 3.5. Since this method is far more efficient and

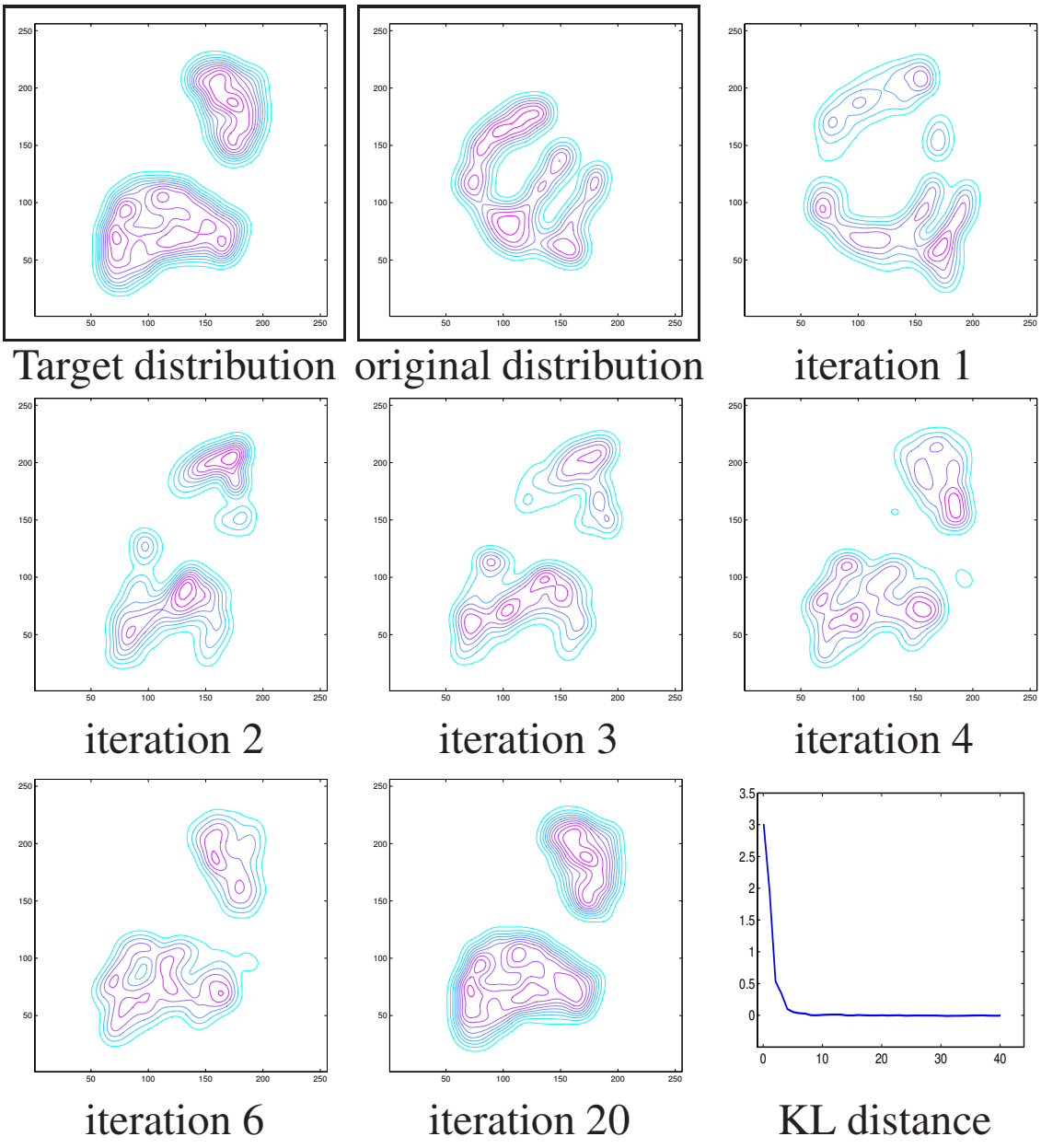

FIGURE 3.9: Source: Pitié et al. (2005). Example of how the pdf transfer function works in a 2-dimensional case. The original distributions change at each iteration of the algorithm until they resemble the target distributions. In the last plot, the Kullback-Leibler distance (an indicator of distance or difference between two distributions) as a function of the number of iterations is shown.

faster than the abundance matching, we can split the data in redshift bins of width 0.01. The redshift of MICE goes up to 1.4, but due to an issue during the generation of MICE, there are no objects with redshift lower than 0.07. Then, when splitting the data into redshift bins, the first one is a larger bin of width 0.1 instead of 0.01 to avoid having too few objects.

2. To ensure that the number density of galaxies is the same in the simulated data and the real data, for every redshift bin we perform an abundance matching over the i magnitude between both samples. Using the abundance matching relation we transform the i magnitude of the simulation to have the same density as a function of i as the survey. Then we remove objects fainter than $i = 23$ to ensure that objects that enter in the transfer process do not carry large photometric uncertainty since $i \sim 23$ is approximately the limit of the 10σ signal to noise detection threshold of DES.

3. Finally, we apply the N-dimensional pdf transfer function for every redshift bin. We describe the algorithm we have implemented as detailed in Pitié et al. (2005). For every iteration k of the algorithm:

   (a) A random rotation R is generated and applied over the simulated, $x_r \leftarrow Rx^{(k)}$, and real data, $y_r \leftarrow Ry^{(k)}$.

   (b) Both samples are projected over all axes n (of the N dimensions given by the number of magnitudes in our case) and the marginals of the distributions f_n and g_n are computed for each axis.

   (c) For each axis n, the one-dimensional pdf transformation t_n is found by matching both marginals f_n and g_n following equation 3.3.

   (d) Every dimension of the rotated simulated sample x_r is transformed to be closer to the real sample by applying the corresponding pdf transformation to every element:

   $$(x_1, x_2, x_3, x_4)_r \leftarrow (t_1(x_1), t_2(x_2), t_3(x_3), t_4(x_4)).$$

   Recall that (x_1, x_2, x_3, x_4) corresponds to the magnitudes (g, r, i, z) in our case. To avoid confusion in the notation we denote them using the former notation.

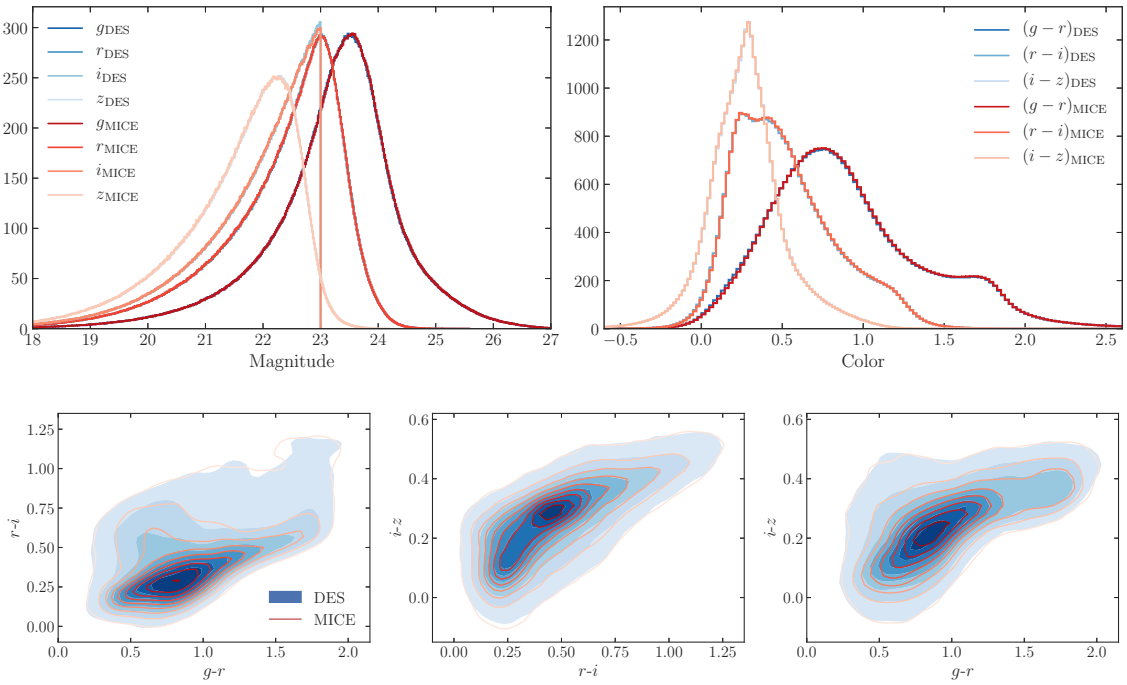

FIGURE 3.10: *Top:* Comparison of magnitudes (left) and colors (right) distributions of DES (blue color scheme) and remapped MICE (red color scheme). The N-dimensional pdf transfer function method has been used to transform MICE magnitudes. *Bottom:* Color-color diagram of DES (blue shades) and remapped MICE (red lines). The full octant of MICE has been remapped but only a subsample of 30 000 galaxies is shown in the plot.

(e) Both simulated, $x^{k+1} \leftarrow R^{-1}x_r$, and real, $y^{k+1} \leftarrow R^{-1}y_r$, samples are rotated back to return to the original coordinate system.

The algorithm iterates these steps, using a random rotation in each iteration, until all distributions converge. Producing a final mapping t that transforms every component n of every object of the simulated sample x_j to resemble real data:

$$t(x) = (t_1(x_1),\, t_2(x_2),\, t_3(x_3),\, t_4(x_4)) = (y_1,\, y_2,\, y_3,\, y_4) = y.$$

Through experimentation we have found that 2000 iterations are enough to completely transfer the pdf properties of magnitudes from DES to MICE in a reasonable amount of time.

A difference between the abundance matching method described in Sec. 3.5 and the N-dimensional pdf transfer function is that in the later method the correlations between magnitudes remain without the need to specifically remap the colors. In Fig.

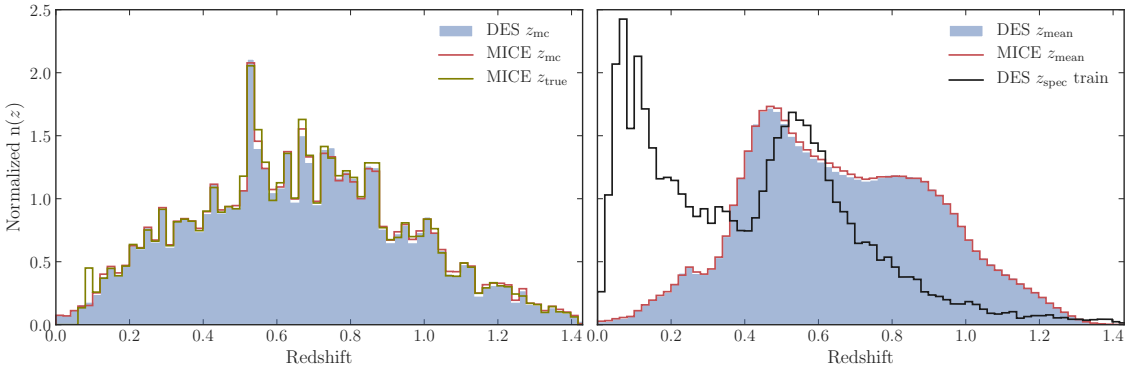

FIGURE 3.11: Photo-z distribution obtained through DNF using the remapped photometry of MICE (red line) compared to the distribution of photo-z of DES (shaded blue) from the Y3 Data catalog. MICE photometry has been transformed using the pdf transfer function. *Left:* $z_{\rm mc}$ distribution of MICE and DES, and true redshift distribution of MICE (green line). *Right:* $z_{\rm mean}$ distribution of MICE and DES, and spectroscopic redshift distribution of the DES training sample used to train DNF (black line).

3.10 we show the magnitudes and colors from DES and remapped MICE using the pdf transfer function. We see that the distributions of both data sets are identical and, more important, the color distributions are the same even without directly remapping them. In addition, the color-color diagrams of the real data are recovered in the simulated data.

An advantage of this method with respect to the abundance matching is that there is no need to define a binning of magnitudes to correctly remap them. The binning of magnitudes was done manually, making more difficult to quickly apply the abundance matching method to other sets of real data or simulations. The N-dimensional pdf transfer algorithm is not only faster and more efficient from the computational point of view but it also makes the remap simpler. Since it is faster, we can split the data into more redshift bins than in the previous method in order to establish a better correlation between the redshift and magnitudes. The memory efficiency of the algorithm allows to remap all the area of MICE at once without splitting the catalog by area. With the method described in this section we have fulfilled the need of having a fast and general method to remap observables.

3.7.1 Photometric redshifts of remapped objects with the pdf transfer function

We have remapped the magnitudes of MICE to resemble those of DES using the N-dimensional pdf transfer function method. We need to check if the resulting photometry allows us to recover the same redshift distribution as in DES. We determine the photo-zs of the remapped MICE using the same algorithm and training sample as in DES as described in Sec. 3.2.3. In Fig. 3.11, we present the resulting photo-z distributions of the remapped MICE galaxies and compare them with the ones from the Y3 Data catalog of DES. We see a good agreement between the simulated and real photo-z distributions, better than when using the abundance matching methodology to remap MICE. In addition, the true redshift distribution of MICE perfectly agrees with the photo-z distributions of DES and remapped MICE, as opposed to the mismatches that happened when using the abundance matching method as seen in Fig. 3.4. So far the remapped photometry with the pdf transfer method gives better results than the previous method.

3.7.2 Sample selection of remapped objects with the pdf transfer function

Selected galaxy samples allow to improve the amount of information that can be extracted from observations in order to study large-scale structure probes such as galaxy clustering, clusters of galaxies and weak lensing. Having simulated samples that resembles real galaxy samples would allow to study the same large-scale structure probes, asses the performance of the methods used to analyze them and determine possible bias among other purposes. We focus on recovering two lens samples used in DES analyses and described in Sec. 3.3: the redMaGiC and the magnitude-limited samples. Previously we tried to recover them with the photometry remapped with abundance matching and we saw there were issues and mismatches. Now we check if the remap with the pdf transfer function gives better results.

In Fig. 3.12, we present the redshift distributions of the remapped MICE and DES obtained from the magnitude-limited sample selection in each of the six redshift bins defined in Table 3.1. The photo-z distribution of the remapped MICE and DES and the true redshift distribution of the simulation are nearly identical in all redshift bins. There is a better resemblance between redshift distributions when the pdf transfer function is used to remap photometry that when using the abundance

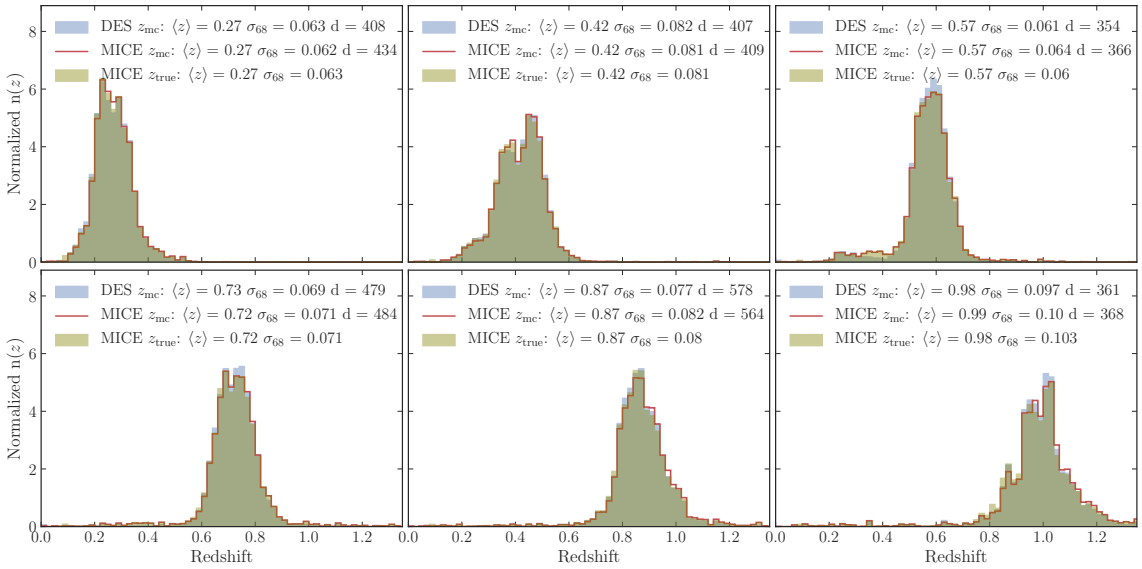

FIGURE 3.12: Photo-z distributions of DES (blue), remapped MICE (red) and true redshift of MICE (green) obtained from the selection of magnitude-limited sample in the six redshift bins defined in Table 3.1. The split of the sample in redshift bins corresponds to version 2.2 of the aforementioned table. To split the magnitude-limited sample in redshift bins, the $z_{\rm mean}$ statistic is used. On each plot, the mean $\langle z\rangle$ and the 68% confidence width, σ_{68}, of every distribution are shown, in addition to the galaxy density d of the remapped MICE and DES.

matching method with previous redshift binning, as it can be seen by comparing Figs. 3.12 and 3.6. There is also a nice agreement of the mean and width of the redshift distributions of simulated and real data. In addition, the galaxy number density of MICE and DES at each bin are more similar than the ones obtained using the others methods of remapping described in this chapter.

After successfully recovering the magnitude-limited sample in the remapped photometry, we want to asses if we are able to select a LRG sample as redMaGiC. The redMaGiC algorithm, as explained in Sec. 3.3.1, selects a sample of LRG with constant comoving density by establishing a selection threshold of luminosity that depends on the input colors. Recall that when the remap of the photometry was done using the abundance matching method in subsamples of redshift bins, as described in Sec. 3.5, the scatter of the remapped colors was too large for the redMaGiC algorithm to successfully detect a red-sequence galaxy sample. Specifically, the redMaGiC algorithm failed in the subpart of the algorithm called redMaPPer algorithm, which optimizes the sample selection to detect clusters with accurate photo-z using colors and redshift. To see the color scatter, in Fig. 3.13, we plot the density map

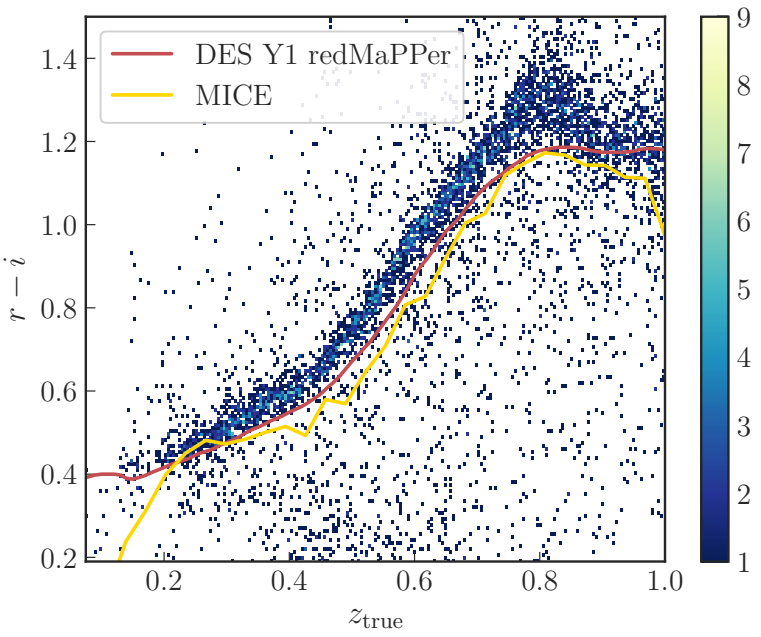

FIGURE 3.13: Density histogram of $r-i$ color of the remapped photometry as a function of the true redshift of MICE. Only massive halos have been considered. The N-dimensional pdf transfer function was used to remap the photometry of MICE. The mean color value across redshift of the remapped MICE is shown in yellow. In red, we show the mean color across redshift of the redMaPPer sample from the First Year Data of DES as presented in DeRose et al. (2019).

of the remapped MICE $r-i$ color as a function of redshift. In the plot, we only consider massive halos in order to compare the mean color of the remapped MICE with the sample of clusters successfully determined by redMaPPer from the DES data. We observe that the scatter of the $r-i$ color has been considerably reduced in comparison to the color dispersion when the remap is done with the abundance matching method, shown in Fig. 3.7. The mean value of the color from the photometry remapped with the pdf transfer function method (yellow line) is closer to the mean color value across redshift of the redMaPPer sample from the DES data (red line) than in the previous remap methods. Taking into account the low scatter of the color, we send the full octant of the remapped MICE for redMaGiC algorithm to be run over the catalog. This time, redMaGiC successfully obtains both a LRG and a clusters of galaxies samples from the remapped MICE. Both samples are available at the CosmoHub[3] data portal.

[3]https://cosmohub.pic.es/

3.8 Summary and conclusions

Accurate representations in simulations of real observations from surveys are crucial to perform reliable cosmological analyses. However, normally there are discrepancies between real data and simulations due to the complexity of the data and the limitations and simplifications that simulations need to assume. In this work we aimed to reduce the differences between the observed magnitudes of surveys and simulations. To reduce these differences, we transformed the photometry of the existing simulations to model real observations by determining a relation that mapped the statistical properties of the real observables to simulations. Since photo-zs are used in a wide range of analyses, in this work we focused on the recovery of realistic photo-zs in simulations. Photo-zs are determined from the photometry in several bands, thus we wanted to reproduce the observed photometry distributions in simulations. In this chapter, we explored two methods to transfer the statistical properties of the photometry from real observations to simulations while keeping the correlation between photometric bands.

Cosmological surveys use several selection criteria to define the optimum galaxy sample to extract the maximum information from observations in order to do science. To select the galaxy sample and perform precise science, accurate photo-z of galaxies are required. Therefore, in order to ensure that our transformed photometry from the simulation was useful to perform cosmological analyses, we determined the photo-z from the remapped simulated photometry and checked if the remapped photometry and resulting redshift allowed us to reliably define the same galaxy samples that have been defined in real data. In this work, we wanted a simulation with photometry and redshifts like the ones in DES. So once the statistical properties of the photometry distribution of DES were transferred to the MICE simulations, we determined the photo-z of the transformed simulated galaxies using the same machine learning algorithm and training sample as the ones used in DES. After the photo-z were determined, we used the transformed photometry and the obtained photo-z to selected the same lens galaxies samples as the ones defined in DES, which are the redMaGiC and the magnitude-limited samples.

The first method we developed to remap the photometry from real data to simulations was an abundance matching method. This process, that transfers the distributions of the photometry from DES to MICE, consisted in sequentially applying

a series of abundance matching transformations. The abundance matching transformation consisted in determining a relation for the g, r, i and z magnitudes that have the same abundance in the MICE and DES datasets. The matching ensured that for each values of each magnitude, the same abundance was found in both data sets. The obtained abundance matching relation was applied over each of the four magnitudes separately to transform the magnitudes of MICE to resemble the same distribution and abundance of DES magnitudes.

In order to ensure that the correlation between magnitudes remained, the data was split into bins of color and magnitude, and the abundance matching method was applied over each subset. For example, to transform the g magnitude, the data sets were split into bins of $g-r$, $r-i$, and i. This process of remapping managed to reproduce the distribution of magnitudes and colors. However, there were small differences between the color-color diagrams of DES and the remapped MICE. The differences in the color-color diagram were due to the fact that the abundance matching for each magnitude did not consider the correlation with each of the other colors. For example, the $i-z$ color was not considered in the remap of the g magnitude. But adding an extra dimension by also binning $i-z$ was computationally expensive. Despite the small differences, we computed the photo-z using these transformed magnitudes and the same approach as in DES. The difference between the color-color relations led to differences between the obtained photo-z of MICE and the photo-z of DES. This discrepancy had an adverse effect in the definition of the galaxy samples. The photo-z and true redshift distributions of MICE obtained with the magnitude-limited sample selection were different to the photo-z distribution of DES for the same sample selection. In addition, the dispersion between the color and redshift relation was too high to successfully run the redMaGiC algorithm.

This problems led to a modification in the remap process where the data sets were also split into redshift bins in order to help to better remap the correlation between magnitudes and redshift. Then the abundance matching was applied into subsets of redshift, magnitude and color bins. The addition of the redshift binning improved the resemblance of the the color-color diagram, the photo-z distribution, and the magnitude-limited sample of the remapped MICE and DES. There was still scatter between the color and redshift relation of the remapped data set but it was low enough to run the redMaGiC algorithm.

We realized that we needed to include more redshift bins to improve the remap but due to computational limitation of the algorithm it was not efficient. In addition,

the binning of the magnitudes and colors needed a lot of manual tweaking to have enough objects in each bin. So we spent a lot of time to specifically implement it to remap MICE into DES. Then, we explored another method that was easier to apply to any pair of real and simulated data, and that was computationally more efficient.

The second method we used to remap the data was the N-dimensional pdf transfer function, which determines a continuous transformation that maps the N-dimensional probability density function distributions of the magnitudes from two datasets. This method was originally used to transfer the color palettes between pictures and here we used it in a novel way to transfer statistical properties of observables between data sets. An advantage of this method is that there is no need to define a binning of magnitudes to correctly remap them. Just providing the magnitudes as input variables, this method keeps the correlation between them and thus allows to determine photo-z resembling real data. The transformed magnitudes and the determined photo-z distributions were almost identical to the ones in DES. The magnitude-limited sample was also very similar to the one in DES. Since this algorithm is computationally faster and more efficient, the number of redshift bins could be increased, which reduced the scatter between the color and redshift relation and thus allowed us to obtain a redMaGiC sample similar to the one in DES. With the N-dimensional pdf transfer algorithm there is no need to specify magnitude or color bins to keep the correlations between them. This process was done manually in the previous method. Hence this method can be quickly and easily applied to any pair of simulation and real data. In addition, the memory efficiency of the algorithm allowed us to remap a larger amount of data. With this method we fulfilled the need of having a fast and general method to remap observables.

To sum up, we have seen that reducing the differences between observables from simulations and real data is not a trivial task. This is especially true regarding magnitudes whose physical relation should be kept in order to correctly describe the spectral energy distribution of galaxies and thus be able to determine their photo-zs. Both methods presented in this chapter can be used to remap the magnitudes from simulated data to resemble observations. But the N-dimensional pdf transfer algorithm is faster, more general and gives better results. We would like to highlight that we have used the N-dimensional pdf transfer function to remap magnitudes but we think the algorithm can be easily used to remap any other observable property over any pair of real and simulated data. This method was originally proposed to transfer the color palettes between pictures, its applications in cosmology could be

advantageous and worth exploring.

Chapter 4

Self-Organizing Map

Photometric surveys observe millions of galaxies with filters spanning from the optical to near-infrared wavelengths to estimate their photo-zs, which are used for cosmological purposes such as weak lensing and galaxy clustering analyses. The performance of the cosmological analyses depends on the accuracy and precision of the photo-zs. The photo-z precision requirement of the Euclid mission is of $\sigma \leq 0.05(1+z)$ for photometric galaxy clustering and weak lensing analyses, with a mean error for each redshift bin smaller than $\Delta\langle z\rangle \leq 0.002(1+\langle z\rangle)$. This level of accuracy can only be achieved using spectroscopic samples to calibrate the $n(z)$ that are fully representative of the entire range of galaxy types and redshifts that conform the photometric sample.

The Complete Calibration of the Color-Redshift Relation survey (C3R2; Masters et al. 2017) is an ongoing spectroscopic effort performed at the VLT and Keck facilities aimed to identify and observe the galaxies that are needed to have a fully representative spectroscopic sample, specifically for the Euclid survey. To define a representative spectroscopic sample, the galaxy color–redshift relation is empirically calibrated using the machine learning algorithm called Self-Organizing Map (SOM). The SOM maps and projects the high-dimensional galaxy color space given by the photometric sample onto a two dimensional grid. Each cell of the grid has a set of colors assigned that correspond to the mean spectral energy distribution of galaxies from the photometric sample that occupy that cell. Studying the distribution of available spectroscopic redshifts in the SOM map, this method allows us to detect which regions of the galaxy color space are not represented in the current available spectroscopic samples.

In the first part of this chapter, we will use the SOM to identify and target galaxies that are needed to fill the cells without spectroscopy and thus complete the color-redshift mapping as part of the C3R2 effort. In order to build and populate a SOM

map for C3R2 with enough data, we use different fields and surveys. The mix of different surveys makes the photometry heterogeneous. So we need to ensure that it is consistent throughout the different fields in order to properly define and populate the SOM. After introducing the SOM algorithm in Sec. 4.1, we will describe in Sec. 4.2 the conversions we use to homogenize the C3R2 data and the process followed to select the observational targets.

In the second part of this chapter, given the usefulness of the SOM to classify galaxies in the high dimensional color space and map them into a two dimensional grid, we will also use the SOM technique to explore the color space of the photometry of galaxies from the Physics of the Accelerating Universe (PAUS) Survey and study the relation between colors and photo-zs in the survey. Since PAUS observes using forty narrow bands, the numbers of input variables, i.e. colors, is high. Then the definition of the SOM requires special attention. In Sec. 4.3, the thorough process used to decide the input features and the configuration of the SOM is described. Once the SOM is defined and trained, we study the color-redshift coverage of galaxies in PAUS. We also detect the cells without spectroscopic counterpart, which can be a source of bias when comparing spectroscopic and photometric redshifts when assessing the accuracy of the latter. We will also see that the classification of the SOM could be useful to detect anomalous objects in the survey.

4.1 Self-Organizing Map

A Self-Organizing Map (SOM) is a type of neural network algorithm first described in Kohonen (1982). This unsupervised machine learning technique projects the high dimensional data given by the input variables into a two dimensional space preserving the topology. The conservation of the topology means that objects with similar properties in the high dimensional space will be grouped together in the map. In astronomy, this becomes an interesting tool to identify the relation between observed colors and redshift in a very visual way. For example, if the input variables to train the SOM are colors, galaxies with similar observed colors are grouped together in cells. Each cell has assigned a unique value of colors determined by the SOM and these colors can be related to the average redshift of all the galaxies that fall within each cell.

In recent years, the SOM has been used, for example, to estimate galaxy photo-z full probability density functions (Carrasco Kind, and Brunner, 2014), to model

galaxy phenotypes as a function of redshift to calibrate the uncertainties coming from shot-noise and sample variance using Deep Field observations (Sánchez et al., 2020), to group galaxies by phenotype when combining Deep and Wide Field observations in order to break degeneracies in photo-z (Buchs et al., 2019), to estimate the uncertainty on the SOM photo-z direct calibration using absolute magnitudes instead of colors in order to calibrate photo-zs (Wright et al., 2019), and to optimally select objects whose spectroscopy will contribute to completely calibrate the color-redshift relation (Masters et al., 2017). Later in the chapter we will come back to the last application since we also have used SOMs to explore the calibration of the color-redshift relation.

Throughout this chapter we will use the SOM to map and study the galaxy color-redshift space. We have implemented the algorithm in C for computational efficiency following the methodology of the SOM described in Masters et al. (2015) which we explain next. The input of the SOM, in our case, is the colors from the considered galaxy sample. The structure of the SOM is a two dimensional grid of pixels or cells, where each cell has an associated weight vector, $\vec{w}$, with the same dimensions, m, as the number of parameters of the input data. The weight vector establishes the relation between the high dimensional data and the two dimensional map. The number of cells within the map, in other words, the dimension of the map, is one of the parameters one can adapt to the required precision of each problem. In this work we choose the map to have a rectangular form. At the beginning of the algorithm, the weight vectors are initialized following a random normal distribution. The input parameters of the training data will modify them. Usually we choose the observed colors of galaxies as training parameters since they are the quantities more related to redshift, although not exclusively. To train the map, the algorithm starts an iterative process, of $N_{\rm iter}$ iterations, where it randomly selects a galaxy, with properties $\vec{x}$, from the training sample and assigns it to the cell with the weight vector more similar to the input values of the galaxy. This cell is called best matching unit (BMU). To identify which cell is the closest one, a reduced χ^2 distance between each weight vector, $\vec{w}_k$, and the training object is used:

$$d_k^2\left(\vec{x},\vec{w}_k\right) = \frac{1}{m}\sum_{i=1}^{m}\frac{\left(x_i - w_{k,i}\right)^2}{\sigma_{x_i}^2} \tag{4.1}$$

where k indexes the number of cells in the map. This metric wants to take into account the uncertainties, σ_{x_i}, of each color of the galaxy. The cell with the smallest

χ^2 becomes the BMU. Once the BMU, b, is determined, the weight vectors of the map are updated:

$$\vec{w}_k\,(t+1) = \vec{w}_k\,(t) + a(t) H_{b,k}(t)\,[\vec{x}(t) - \vec{w}_k(t)] \tag{4.2}$$

where t is the current iteration. The learning rate function $a(t)$ decreases with time, reducing the changes on the map at each iteration. It is defined as:

$$a(t) = 0.5^{(t/N_{\rm iter})} \tag{4.3}$$

The element $H_{b,k}(t)$ is the neighborhood function evaluated for a cell, k, and the current BMU, b, which depends on the euclidean distance, $D_{b,k}$, between them:

$$H_{b,k}(t) = e^{-D^2_{b,k}/\sigma^2(t)} \tag{4.4}$$

For our framework we choose that the SOM has periodic boundary conditions to avoid boundary effects, which has to be taken into account when computing $D_{b,k}$. The width of the neighborhood function $\sigma(t)$ is defined as:

$$\sigma(t) = \sigma_s \left(\frac{1}{\sigma_s}\right)^{(t/N_{\rm iter})} \tag{4.5}$$

The term σ_s is chosen to be the value of the smaller dimension of the map. In order to reduce computational time and given the decrease of the learning rate function with each iteration, not all weight vectors are updated at each iteration. Only the weight vectors within $D_{b,k} < 3\sigma(t)$ are updated. The final weight vectors establish the relation between the high dimensional data and the projected two dimensional map.

The most important choice of parameters to train the SOM are the variables of the input data (colors, magnitudes, any observable property) and the number of cells of the map. Input variables may vary the smoothness of the map. The dimension of the map determines the dispersion of the input properties of the galaxies within each cell. If there are too few cells, each one will represent a larger group of galaxies. While too many cells may produce an overrepresentation of less common galaxies.

4.2 Calibration of the color-redshift relation

Current spectroscopic surveys cannot sample all the color and magnitude space covered by modern imaging surveys. To meet the precision required for weak lensing cosmology, an accurate photo-zs should be derived. To achieve that, a complete coverage of spectroscopic samples is necessary, specially when using machine learning photo-z estimation.

An ongoing spectroscopic effort to solve this problem, focused on the Euclid survey, is the Complete Calibration of the Color-Redshift Relation (C3R2; Masters et al. 2017). The calibration can also be used for other wide field surveys such as the upcoming Rubin-LSST and WFIRST. The C3R2 observations are being carried out at the Keck telescopes using the DEIMOS (Faber et al., 2003), MOSFIRE (McLean et al., 2012) and LRIS (Oke et al., 1995) instruments, and with a Large Programme at the European Southern Observatory (ESO) with the Very Large Telescope (VLT) using the optical and near-infrared multi-object spectrographs FORS2 [1] (Appenzeller et al., 1998) and KMOS [2] (Sharples et al., 2013).

The C3R2 project uses the SOM technique to characterize the full empirical color-redshift relation, $P(z|C)$, to the Euclid depth ($i_{\rm AB} = 24.5$). To define the map, they use photometric data from the Cosmological Evolution Survey (COSMOS) field (Scoville et al. 2007; Capak et al. 2007; Lilly et al. 2007; Laigle et al. 2016), the VIMOS-VLT Deep Survey (VVDS; Le Fèvre et al. 2004; McCracken et al. 2003) and the Extended Groth Strip field (EGS; Davis et al. 2007). These fields were chosen because they have uniform, well calibrated photometry with similar $ugrizYJH$ bands and depth as those expected to be used in Euclid. Moreover, their different locations allow us to mitigate cosmic variance effect and to complete the color-redshift space in the minimum period of time since it is easier to schedule observations for different locations. However, the photometry of all fields needs to be incorporated onto the Euclid color system in a consistent way to be able to train and reliably place galaxies on the color map. This is not straightforward since the surveys have different instruments, filter calibrations and reduction pipelines to produce their photometric catalogs. Therefore to derive conversions between the photometry in the COSMOS, VVDS and EGS fields, in C3R2 they use observations of the fields performed with the same instruments such as the Canada-France-Hawaii Telescope Legacy Survey [3]

[1] https://www.eso.org/sci/facilities/paranal/instruments/fors/overview.html
[2] https://www.eso.org/sci/facilities/develop/instruments/kmos.html
[3] http://www.cfht.hawaii.edu/Science/CFHTLS/

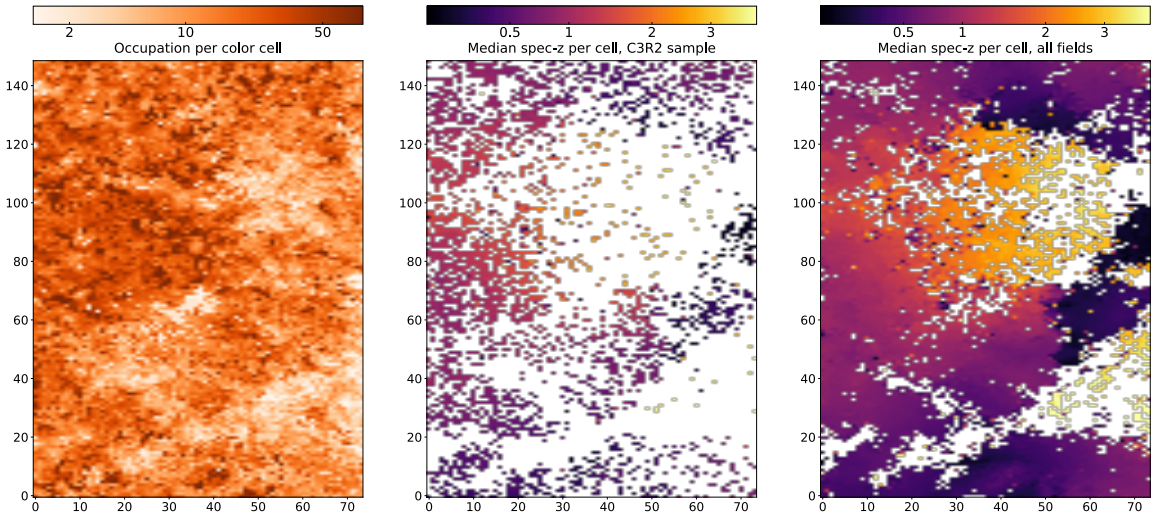

FIGURE 4.1: *Source:* Masters et al. 2019. SOM of 11 250 cells trained and populated with COSMOS, VVDS and EGS galaxies. *Left:* Occupation density of galaxies that fall within each cell. *Center:* Median spectroscopic and high confidence redshift of objects observed by C3R2. The survey has fulfilled 35% of the cells. *Right:* Median spectroscopic redshift of objects that fall within each cell for all considered fields. White cells represent the lack of spectroscopic objects and thus spectroscopic information, which correlates with regions with low occupation. Currently, 76% of cells are covered.

(CFHTLS) Deep Fields for the *ugriz* optical bands, the WIRCAM Deep Survey (WIRDS; Bielby et al. 2012) for JHK_s and UltraVISTA (McCracken et al., 2012) for the near-infrared including the Y band, see Masters et al. 2019 for further details.

To train the map, C3R2 uses as input variables the eight colors $u-g$, $g-r$, $r-i$, $i-z$, $z-Y$, $Y-J$, $J-H$ and $H-K_s$, that correspond to the expected bands that Euclid will use to compute photo-z. Colors are chosen as input variables because they describe the spectral energy distribution and thus the information needed for redshift determination. Once the training is done, galaxies are matched to the final weight vectors to find their corresponding cell in the color space. The final values of the weight vectors establish a partition of the color space. The color values of the weight vectors associated to each cell are related to the redshift by averaging the redshift of all the galaxies that fall within each cell. This way a map between colors and redshift is obtained. Galaxies populate almost all the SOM cells, so there is photo-z information for almost all cells but there are some regions of the map that lack spectroscopic and high confidence redshift, as can be seen in Fig. 4.1. The aim of C3R2 is to target and observe spectroscopic galaxies of missing regions of color space in order to complete it.

TABLE 4.1: Center of the targeted fields in C3R2 survey as appointed in Guglielmo et al. (2020).

Field	Ra	Dec
COSMOS	$10^{h}\,0^{m}$	$2^{\circ}\,12'$
VVDS	$2^{h}\,26^{m}$	$-4^{\circ}\,30'$
EGS	$14^{h}\,19^{m}$	$52^{\circ}\,41'$
SXDF	$2^{h}\,18^{m}$	-5°
ECDFS	$3^{h}\,33^{m}\,5.61^{s}$	$-27^{\circ}\,41'\,8''$
	$3^{h}\,31^{m}\,51.43^{s}$	$-27^{\circ}\,41'\,38.80''$
	$3^{h}\,31^{m}\,49.94^{s}$	$-27^{\circ}\,57'\,14.56''$
	$3^{h}\,33^{m}\,2.93^{s}$	$-27^{\circ}\,57'\,16.08''$

To determine which targets in the unsampled regions have higher priority, C3R2 establishes a system of prioritization based on the amount of new information a galaxy contributes to the calibration of the $P(z|C)$, how common are the unsampled colors according to the population on the empty cells, and the probability of successfully obtain spectroscopic redshift given the instrument, exposure time, and expected galaxy properties.

In addition to the COSMOS, VVDS and EGS, the Subaru/XMM-Newton Deep Survey field (SXDF; Furusawa et al. 2008) and the Extended Chandra Deep Field-South Survey field (ECDFS; Lehmer et al. 2005) have also been targeted since they provide additional spectroscopic redshifts. They have been merged to the Euclid color system after applying color correction conversion as the former three fields (Guglielmo et al., 2020). The center positions of the fields are shown in Table 4.1.

In Masters et al. (2017) they estimate that about 5000 new spectra would be necessary to fill the cells without spectroscopic redshifts that establish the color-redshift mapping. The galaxies in these unmapped cells usually correspond to objects with faint magnitudes ($i_{AB} > 23$) that surveys usually do not target due to their magnitude limit of observations. Efforts to observe them are being performed for objects with $z > 1$ (Guglielmo et al., 2020) and for $z < 1$ (Castander et al., in prep.). The C3R2 programme is designed to sample these faint galaxies in both the optical and near-infrared wavelength ranges to explore a wide redshift range. Up to now, the Keck observations (Masters et al. 2017; Masters et al. 2019) have obtained 4454 high quality galaxy redshifts that have increase the redshift coverage of the SOM cells to 76%. The KMOS VLT observations have yielded 424 high-quality spectroscopic

redshifts that have filled 269 previously empty cells.

4.2.1 Conversions to a homogeneous color system

As mentioned above, to build a single SOM map from photometry coming from different fields we need to ensure that their photometry is consistent. In particular, the VLT observations have included the ECDFS field that was not originally included in the Keck programme to the C3R2 project (the C3R2 project trained the SOM using the photometry from COSMOS, VVDS and EGS fields, and later included the spectroscopy from SXDF and ECDFS). However, photometric data in the common C3R2 system was only available for a small area of the ECDFS field (see Fig. 4.4). In order to properly take advantage of this field we needed enough data with photometry in both systems to compute the transformation between the C3R2 color system and the ECDFS photometry. We use the VVDS field to provide the photometry transformations between systems since it has a larger area with data in the C3R2 system. The best photometric data available in the VVDS field comes from the DES survey, whose photometry is also available in the ECDFS field (see Fig. 4.2). As both DES and C3R2 have data in common in a larger area of the VVDS field, we use that field to calibrate the relations to transform DES photometry into C3R2 photometry. This transformations will be used to convert the optical bands of the ECDFS data. Besides the optical bands, the C3R2 SOM map also includes near-infrared photometry. We use the MUltiwavelength Survey by Yale - Chile (MUSYC; Gawiser et al. 2006) survey to provide the transformations for the near-infrared bands. To transform the MUSYC photometry we use the small area in common with C3R2 in the ECDFS (see Fig. 4.4). We describe below the process we have followed to bring these photometric data to the same system as the other C3R2 fields. With these transformations we could correctly place all the galaxies of the DES-MUSYC ECDFS area in the SOM and select targets for spectroscopy from the empty cells with the same prioritization scheme as in the rest of the C3R2 fields.

VVDS

We transform the photometry in the VVDS field into the C3R2 system that was used to build the C3R2 SOM map. The remapped photometry is used to select the objects for spectroscopy at the VLT. To determine the photometry transformations, we take the observations performed by Keck in the VVDS field in the

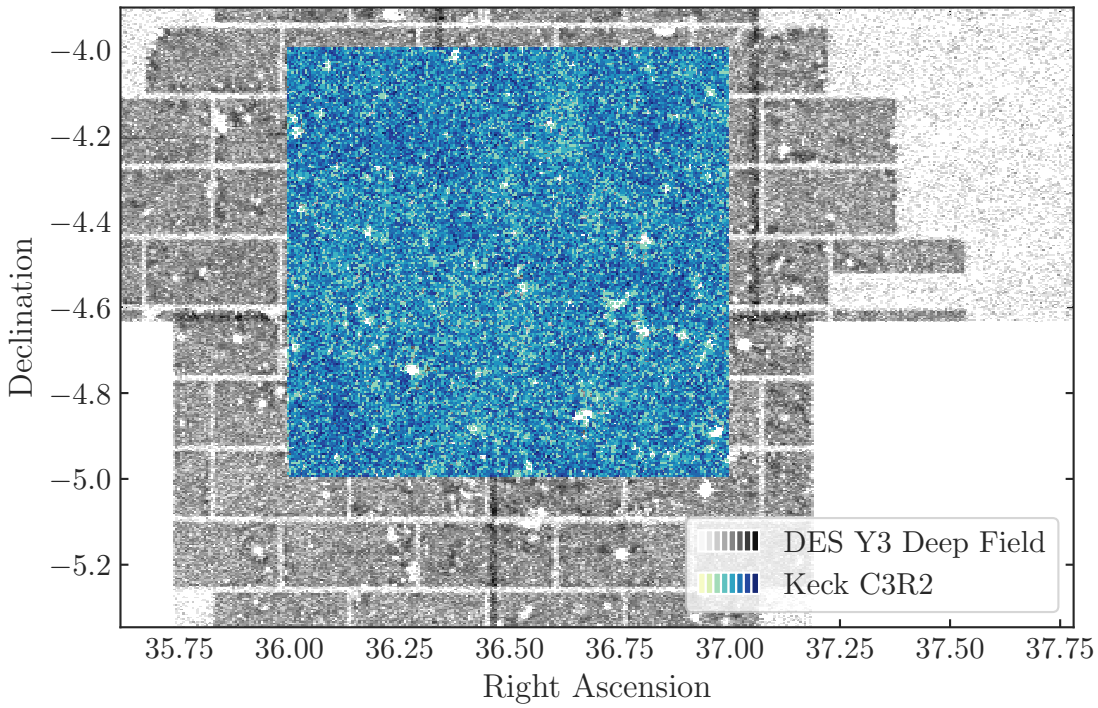

FIGURE 4.2: VVDS field observed by Keck and DES Y3 Deep Field. Objects within 1″ distance are matched.

CFHTLS+WIRDS/UltraVISTA photometric system (C3R2 photometry to simplify) and match those objects to the ones observed by the DES Y3 VVDS Deep Field [4] in the *grizY* bands. We match objects that are withing 1″ distance in both catalogs in the same field (see objects in VVDS in Fig. 4.2). Before determining the conversion between both photometric systems, we want to avoid galaxies with unphysical colors or excessively large errors that will only bring uncertainty to the relation. Therefore, we keep objects according to the following selection criteria

$$\begin{aligned} i < 25, \quad \sigma_{grizY} < 0.5, \quad -0.5 < g - r < 2.5 \\ -0.75 < r - i < 3, \quad -0.5 < i - z < 1.25, \quad -1 < z - Y < 2 \end{aligned} \tag{4.6}$$

to remove colors outliers and galaxies without reliable photometry. The σ_{grizY} is the magnitude error in each band. After the cuts, the remaining number of galaxies to compute the conversions is 42 023. A polynomial fit to the data (Fig. 4.3) gives the

[4] `https://cdcvs.fnal.gov/redmine/projects/des-photoz/wiki/Y3_deep_coadd`

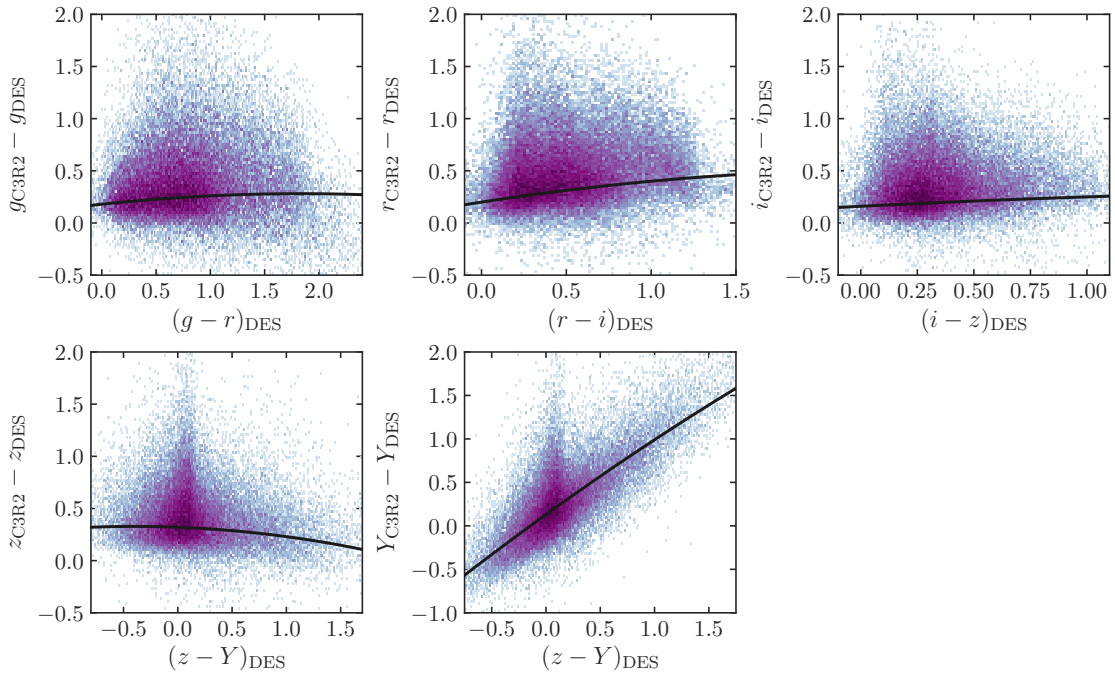

FIGURE 4.3: Difference of C3R2 and DES bands for objects matched in VVDS field as a function of DES colors. A polynomial fit (black line) gives the relation to transform the photometry between color systems.

following transformations:

$$\begin{aligned}
g_{\mathrm{C3R2}} &= g_{\mathrm{DES}} + 0.18 + 0.11(g-r)_{\mathrm{DES}} - 0.03(g-r)^2_{\mathrm{DES}} \\
r_{\mathrm{C3R2}} &= r_{\mathrm{DES}} + 0.20 + 0.25(r-i)_{\mathrm{DES}} - 0.05(r-i)^2_{\mathrm{DES}} \\
i_{\mathrm{C3R2}} &= i_{\mathrm{DES}} + 0.16 + 0.11(i-z)_{\mathrm{DES}} - 0.02(i-z)^2_{\mathrm{DES}} \\
z_{\mathrm{C3R2}} &= z_{\mathrm{DES}} + 0.32 - 0.04(z-Y)_{\mathrm{DES}} - 0.05(z-Y)^2_{\mathrm{DES}} \\
Y_{\mathrm{C3R2}} &= Y_{\mathrm{DES}} + 0.13 + 0.90(z-Y)_{\mathrm{DES}} - 0.04(z-Y)^2_{\mathrm{DES}}
\end{aligned} \tag{4.7}$$

These relations will be used to transform the photometry of DES into the C3R2 color system.

ECDFS

The same process used to obtain the photometric transformations for the optical bands between the DES and C3R2 photometric systems in the VVDS field is applied to obtain the photometric transformations for the near-infrared bands between the MUSYC and C3R2 photometric systems in the ECDFS field. The near infrared JH

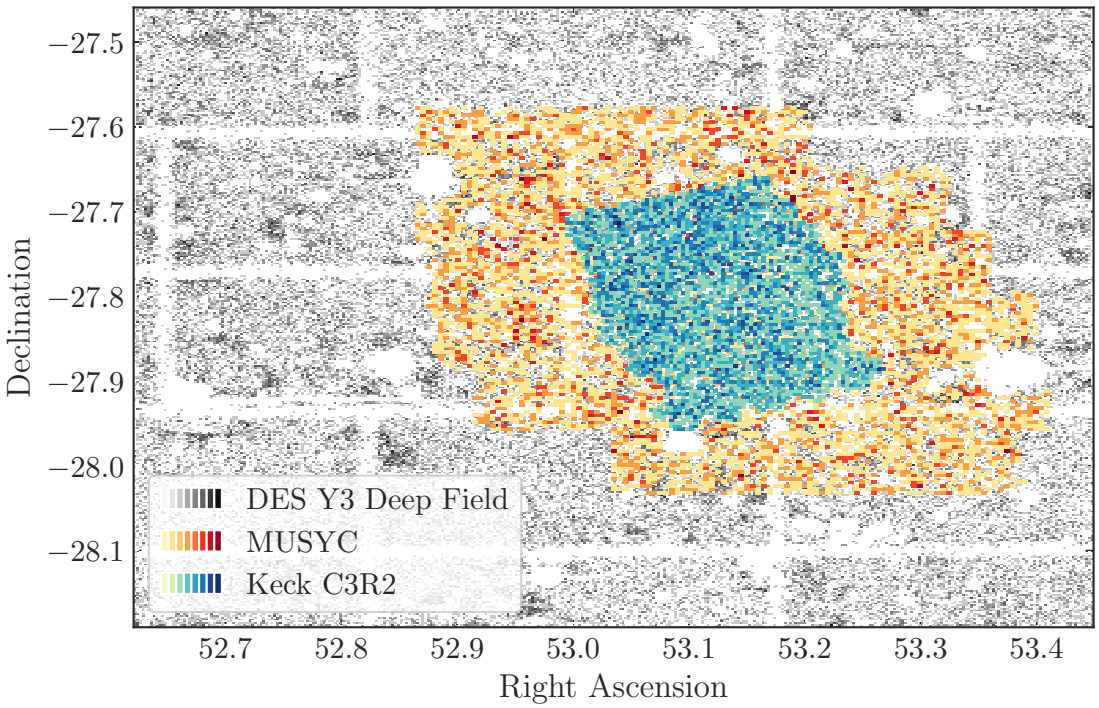

FIGURE 4.4: ECDFS field observed by Keck, MUSYC and DES Y3 Deep Field. Objects within 1″ distance are matched.

bands are added into the photometry system conversion. Despite K_s being a band used to define the final SOM in C3R2, here it is not included, since at the moment these procedures were being done the SOM was just defined with the $ugrizYJH$ colors.

The ECDFS field lacks uniform C3R2 photometry in the optical and near-infrared filters in a sufficiently large area. Therefore to define a VLT system to find the transformations, objects in the ECDFS field observed by Keck are matched to the DES Y3 Deep Field and the MUSYC objects, see galaxies in the ECDFS field in Fig. 4.4. The public data release of MUSYC [5] (Taylor et al., 2009) allows the addition of J and H bands that have a detection depth at 5σ of 23.0 and 21.6, respectively. Which is not optimum in terms of C3R2 depth but can help to break degeneracies when placing objects in the SOM after the conversion between color system is done.

The transformations in the VVDS field for optical bands (Eq. 4.7) are used in the ECDFS field. For the near infrared bands, the following relations between the

[5] http://www.astro.yale.edu/MUSYC/

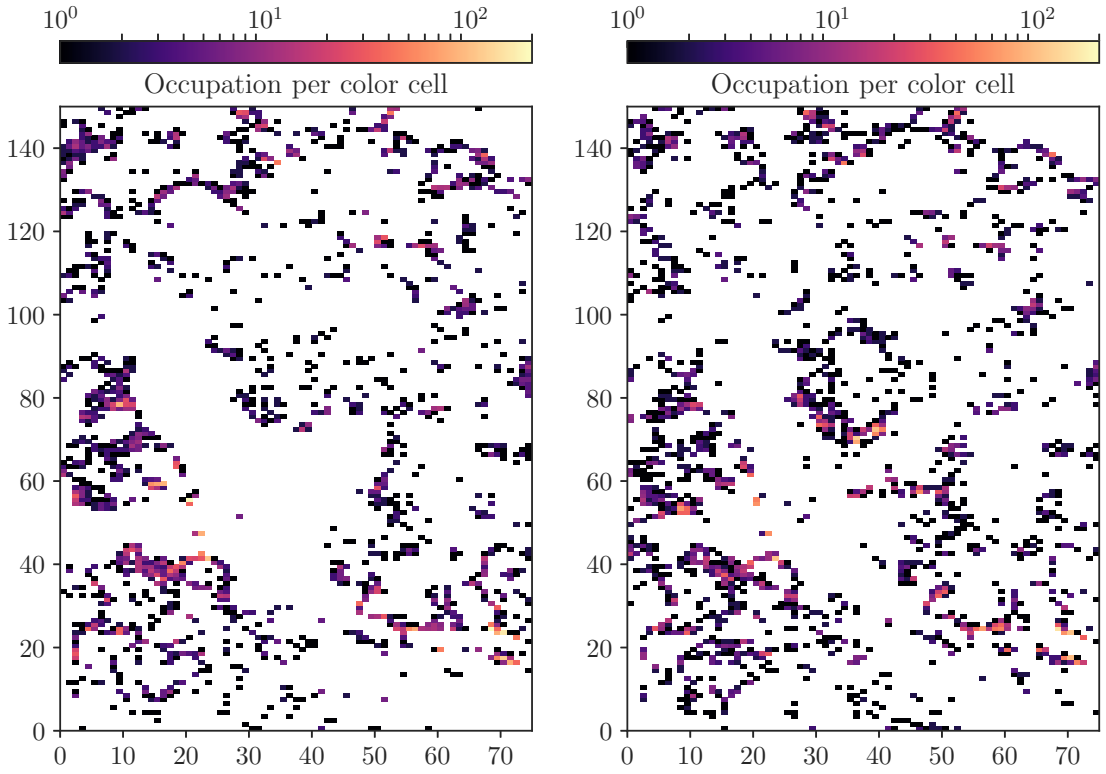

FIGURE 4.5: SOM of 11 250 cells trained with COSMOS objects with C3R2 colors and generated for the first Data Release of C3R2. The SOM is populated with galaxies from ECDFS field observed with all DES Y3 and MUSYC bands and converted to C3R2 photometry system. These galaxies are assigned to a cell through a best color match with the SOM weight vectors. *Left:* Occupation density of galaxies allocated using only $g-r$, $r-i$, $i-z$ and $z-Y$ DES colors. Galaxies fall in 1577 different cells. *Right:* Galaxy placement is done using DES colors and $Y-J, J-H$ MUSYC colors. The number of occupied cells increases up to 1780.

systems are obtained with a polynomial fit as we did previously in the VVDS field:

$$\begin{aligned} J_{\mathrm{C3R2}} &= J_{\mathrm{MUSYC}} - 0.22 - 0.11(J-H)_{\mathrm{MUSYC}} - 0.35(J-H)^2_{\mathrm{MUSYC}} \\ H_{\mathrm{C3R2}} &= H_{\mathrm{MUSYC}} - 0.40 + 0.45(J-H)_{\mathrm{MUSYC}} - 0.25(J-H)^2_{\mathrm{MUSYC}} \end{aligned} \tag{4.8}$$

We apply the relations 4.7 and 4.8 to place the galaxies of the ECDFS field with DES and MUSYC photometry in the SOM determined with C3R2 photometry.

ECDFS infrared bands inclusion

As stated above, conversion of ECDFS objects to C3R2 color system includes DES optical bands and MUSYC infrared bands. In this brief section, we want to emphasize the importance of adding the same bands used to define the SOM color cells when we want to place galaxies within it. Since the cell assignation of galaxies is done through a best match with the weight vectors of the cells following the relation 4.1. Therefore, if the number of bands similar to the color vectors of the SOM increases, the degeneracy of colors decreases, thus objects are better placed in the SOM. As evidenced in Fig. 4.5, when galaxies are placed in the SOM using only the $g-r$, $r-i$, $i-z$ and $z-Y$ DES colors, they tend to group more and occupy fewer cells. In this case, a total of 1577 different cells are filled. Whereas the addition of $Y-J, J-H$ MUSYC colors to the optical ones spreads galaxies across the map, increasing the number of occupied cells to 1780.

4.2.2 Target selection

Once we bring DES *grizY* (for VVDS and ECDFS) and MUSYC *JH* (only for ECDFS) bands into C3R2 color system, we assign galaxies to their corresponding position in the SOM defined by the C3R2 photometry by matching galaxies to the weight vectors related to each cell through relation 4.1. We do not match objects that miss one or more band information to reduce degeneration in the assignation since our photometry already lacks $uJHK_s$ and uK_s observations for VVDS and ECDFS, respectively. The final number of galaxies that can be successfully placed in the SOM is 407 982 for VVDS and 7627 for ECDFS.

To determine which galaxies are useful targets for the C3R2 project, a priority is assigned to them. This priority weights galaxies according to their usefulness for the photo-z calibration, thus prioritizing the galaxies that fall into a well-populated color cell with few or non-existing spectroscopy. The priority decreases as the quality of the spectroscopic coverage of the cell increases. It also decreases if the colors of the galaxies are too different from the color vector of the cell they fall into to avoid calibrating the $n(z)$ with non-representative colors. The priority also sorts galaxies according to the probability of obtaining a secure redshift for a planned observation.

The weight vectors of the SOM defined by C3R2, a list of objects observed in COSMOS through C3R2 photometry and their target priorities have been obtained from the C3R2 team. These COSMOS objects are the ones used to prioritize and

thus select targets for Keck observations for C3R2 in Masters et al. (2019). We take the list of objects provided by the C3R2 team, allocate them into the SOM and compute the priority average of all objects that belong in the same cell as the priority corresponding to that cell. The target selection priority of our galaxies for VLT observations is given by the priority value of the cell they fall into, being the cell priority the average priority computed from COSMOS objects.

When designing VLT observational mask, we try to keep a balance between observing the maximum number of galaxies (placing the maximum number of slitlets per mask) while targeting the ones with higher priority.

4.3 Exploring the color space of PAU

Self-Organizing Maps are very useful to classify galaxies spanning a high dimensional color space and map them into a two dimensional grid. In this section, we want to use the SOM technique to explore the color space of the photometry of galaxies from the Physics of the Accelerating Universe Survey (PAUS) and study the relation between colors and photo-zs in the survey.

As explained in more detail in Sec. 2.2, the main characteristic of PAUS is its coverage of the sky with 40 narrow bands ranging from 450 to 850 nm in steps of 10 nm. These observations are combined with external deep broad band photometry. The combination of narrow and broad bands increases the observed galaxies per area with photo-z sub-percent precision, which can be up to an order of magnitude better than in broad band surveys (Eriksen et al., 2019)

For the analysis of the color space, we use the PAUS observations in the COSMOS field. We choose COSMOS because this field has been extensively observed by spectroscopic surveys as well as photometric surveys covering all the wavelength range from ultraviolet to infrared. This field has been used to determine the photo-z precision of PAUS by comparing it with the highly reliable spectroscopic redshifts of COSMOS. The comparison with COSMOS has been used in Eriksen et al. (2019) where the photo-z of PAUS has been determined through a template fitting technique and in Eriksen et al. (2020) where it has been estimated through a machine learning method. In both cases, the photo-z of PAUS is computed using the 40 narrow bands of the survey in addition to the 6 broad bands from COSMOS catalog as available in Laigle et al. (2016) (COSMOS2015). The bands used are u^* from CFHTLS and B, V, r, i^+, z^{++} from Hyper Suprime-Cam at Subaru. The resulting photo-zs are

compared to the spectroscopic redshifts of zCOSMOS (Lilly et al., 2007) that has a bright magnitude selection of $15 < i_{\rm AB} < 22.5$, which covers approximately a redshift range of $0.1 < z < 1.2$ in 1.7 deg^2 of the COSMOS field.

Our goal is to complement these photo-z analysis of PAUS by exploring the behavior of the color space of the survey.

4.3.1 Implementation of the color map

The color space of the PAUS survey composed of 40 narrow band filters is a highly complex multi-dimensional space. We want to study the color space reducing its dimensionality using a SOM map. The definition of the SOM itself is complex and delicate given the number of input parameters. The map is sensitive to the chosen parameters, specially the number of cells and the input variables. Along this section, we experiment with these parameters by varying the number of cells and training the map with different combination of input variables such as colors or luptitudes (defined in equation 4.10).

The data for the analysis consists of the observations of PAUS in the COSMOS field matched to the COSMOS2015 catalog by their galaxy identification number to obtain a catalog containing the 40 narrow band fluxes of PAUS, the u^* flux from CFHTLS and the B, V, r, i^+, z^{++} fluxes from Subaru. From this match, two photo-zs are available: one from PAUS, $z_{\rm b}$, which is the estimate of the redshift probability distribution determined through the template fitting algorithm described in Eriksen et al. (2019), and a photo-z from COSMOS2015, $zp_{\rm gal}$, determined with 30 bands photometry. In addition, these objects are matched to the zCOSMOS catalog to have spectroscopic redshift, $z_{\rm spec}$, and a confidence class of the redshift reliability. This catalog is very similar to the one used in Eriksen et al. (2019). Since this study was done before this publication, the photo-zs of PAUS are from an earlier version and they may slightly differ from the ones used in the publication.

Choice of input features

Self-organizing maps need an input vector of features whose values are used to train the map and thus determine the relation between the high dimensional space of the features and the two dimensional grid. The grid cells, which are a partition of the color multidimensional space, can be related to other properties such as redshift. We want to choose observable variables related to the spectral energy distribution of

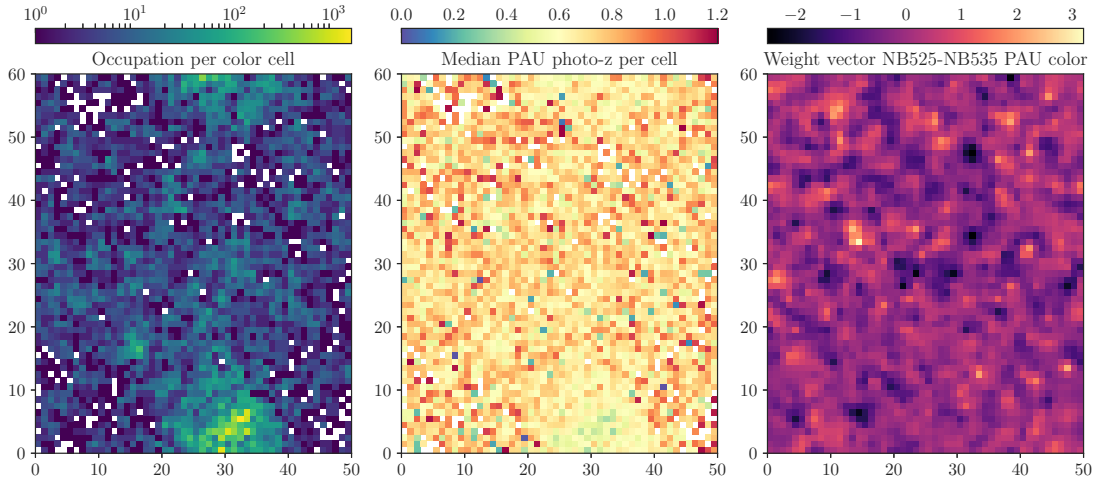

FIGURE 4.6: SOM of 3000 cells trained with colors from 40 narrow bands of PAUS objects in COSMOS field. The SOM is populated with galaxies from PAUS in the mentioned region. *Left:* Occupation density of galaxies. *Center:* Median photo-z of PAUS of galaxies that fall within each cell. White cells lack objects with similar colors as the weigh vectors of these particular cells. Which whether implies absence of observed galaxies with these colors or that the SOM has determined a few unrealistic vectors without objects in real color space. *Right:* Outcome of the training. Weight vector associated to each cell corresponding to color between two narrow bands whose filters are centered at 525 nm and 535 nm.

galaxies as input features since they are used to determine photo-zs. Remember that the function of the SOM is to describe the input variable space and reduce it into a two dimensional grid. It does not determine the redshift itself. During training, the SOM assigns a label of galaxy observable properties to each cell, as described in Sec. 4.1. After the training, we assign a redshift value to each cell of the grid map by computing the average value of the redshifts of all objects that fall in the same cell. That is how we establish a relation between the input features and redshift.

Magnitude colors of the forty narrow bands Intuitively, our initial choice as input features to calibrate the multidimensional space are colors (the difference between magnitudes values) of the 40 narrow bands of PAUS. To ensure the map is trained with the maximum information all objects are added, even the ones not observed in some bands or with measurements not defined in the magnitude system, i.e, with negative fluxes. Missing values are marked with a placeholder number and are not taken into account when training the map. The resulting training sample consist of 38 896 objects from PAUS.

The SOM trained with the 40 narrow bands colors of PAUS is shown in Fig. 4.6.

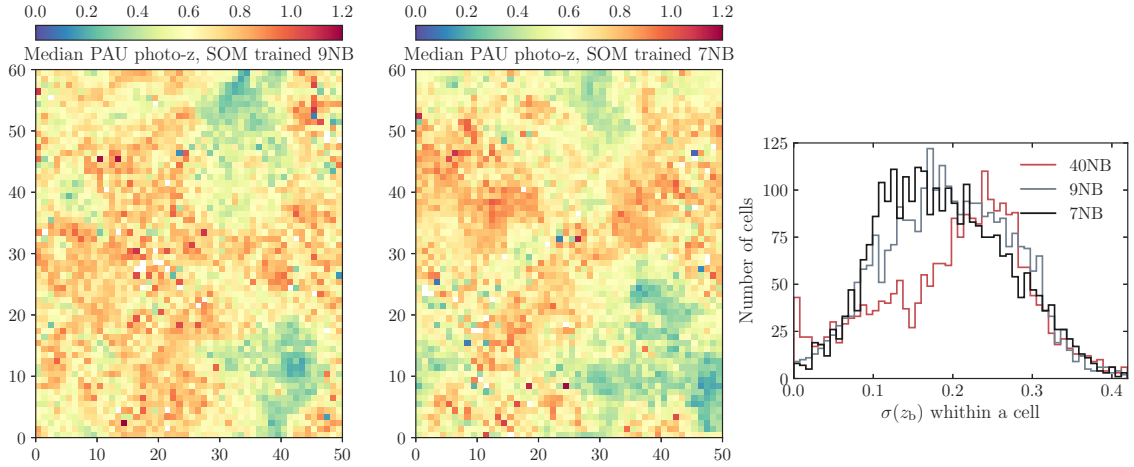

FIGURE 4.7: *Left:* SOM of 3000 cells trained with colors from 9 narrow bands of PAUS centered at 475, 515, 555, 595, 645, 695, 745, 785 and 835 nm. The SOM is then populated with these galaxies and the median photo-z of PAUS of galaxies that fall within each cell is shown. *Center:* Median photo-z of PAUS of galaxies that fall within each cell of a self-organizing map trained with 7 narrow bands centered at 475, 535, 585, 645, 715, 785 and 835 nm. *Right:* The dispersion of photo-z of objects that populate each cell is computed through the standard deviation statistic, $\sigma(z_b)$, when the map is trained with 40, 9 and 7 narrow bands. Taking into account only cells with more than one object, the distribution of cells that have certain dispersion is shown.

The median PAUS photo-z per cell (center panel) follows an irregular pattern, implying that adjacent cells do not correspond to similar objects in the high dimensional color space. The pattern of a randomly selected color from the color vector associated to each cell (right panel), also shows an abrupt change of values between closer cells. This indicates that the 40 narrow bands of PAUS do not allow to define a smooth map. The amount of galaxies we have is not enough to train the SOM and coherently map the 39 dimensional data, since a large number of dimensions in the input data increases the volume of the data space and the sparse of the data.

Fewer narrow bands Then, the second choice of input features is to select a few bands of the 40 narrow bands to reduce the dimensionality of the problem. We test the performance of two sets of bands: 9 and 7 narrow bands with a few bands of separation between them and avoiding the ones at the edge of the PAUS optical range. The resulting SOMs are shown in Fig. 4.7. These SOMs have a median photo-z per cell across the map smoother than when the map is trained with 40 narrow bands. We compute the dispersion of the photo-zs of the objects that populate

each cell when the map is trained with 7, 9 and 40 narrow bands (right panel in Fig. 4.7). The mean of these dispersions are 0.189, 0.196 and 0.204 for 7, 9 and 40 narrow bands respectively. So a reduction of the number of input features reduces the dimensionality of the classification problem and thus helps the map to group galaxies in packs with similar characteristics.

Six broad bands Since we believe that just selecting a subset of bands of the 40 narrow bands seems a bit arbitrary, we try to consider them again as a training features but this time including the u^*, B, V, r, i^+, z^{++} broad bands from COSMOS. The broad bands of COSMOS together with the narrow bands of PAUS are used in Eriksen et al. (2019) to estimate the photo-zs of the PAUS objects. Therefore we think it is interesting to also consider the broad bands to study the color-redshift space. Another aspect of using narrow bands that we suspect is affecting the performance capacity of the SOM to extract information, is that narrow bands might be too correlated between them. So we change the difference of adjacent narrow band magnitudes as input and as another test we use the difference of the narrow band magnitudes with respect to the i magnitude of COSMOS. Since COSMOS bands are less correlated than narrow bands, we think they can help the SOM disentangle information between bands and thus determine a better correlation between colors and the redshift of galaxies.

In order to have the maximum reliable information, we set the requirements that objects in the training sample must have been observed in all the bands considered and their measurements should be defined in the magnitude system, i.e, their fluxes should be positive. Since we think that training the map with objects than have missing bands might lead to biases and a wrong definition of the colors space determined by the SOM. If we only use broad bands to define the map, the number of objects with all bands with positive fluxes is 38 765. However, if we also consider the narrow bands of PAUS, the number reduces to 18 140.

In Fig. 4.8, we show the photo-z estimates for cells computed as the median photo-z of PAUS of the objects associated with each cell. We show the photo-z estimates for the SOM when it has been trained using as input features the color magnitudes of the 6 broad bands of COSMOS without (left panel) and with (central panel) the color magnitudes of the 40 narrow bands of PAUS. The difference with previous choices of input is the inclusion of broad bands that is the only set of features that generates a SOM with smooth photo-z estimates across the map. Implying that

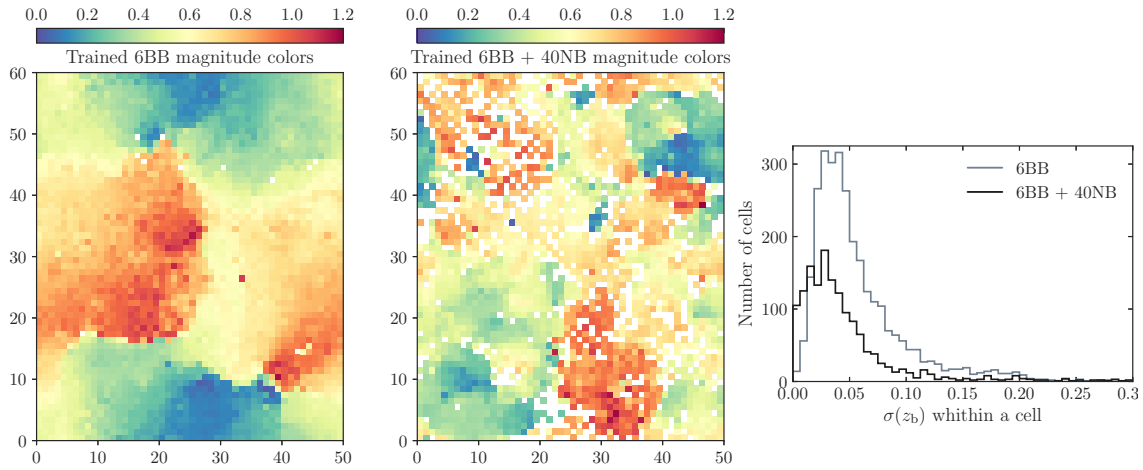

FIGURE 4.8: *Left:* SOM trained with magnitude colors $u^* - B$, $B - V$, $V - r$, $r - i^+$, $i^+ - z^{++}$ from 6 broad bands of COSMOS. The SOM is then populated with PAUS galaxies in COSMOS field. The median photo-z of PAUS of galaxies that fall within each cell is shown in the plot. *Center:* Median photometric redshift of PAUS of galaxies that fall within each cell. The input variables to train the map are the same colors as the ones used in the SOM of the left panel in addition to the difference of magnitude between the 40 narrow bands of PAUS and the i^+ magnitude of COSMOS. Having a total of 45 input features. Only objects with observations defined in the magnitude system, i.e positive fluxes, in all bands have been considered. *Right:* The dispersion of photo-z of objects that populate each cell is computed through the standard deviation statistic, $\sigma(z_b)$, when the map is trained only using the 6 broad bands of COSMOS (grey) and when adding the 40 narrow bands of PAUS (black). The distribution of cells that have certain photo-z dispersion is shown in the plot. Only cells with more than one object have been taken into account.

the SOM is capable to group similar objects in the high dimensional space together in the map based on the information of these magnitudes. Therefore our final choice of training features should include the 6 broad bands of COSMOS. In Fig. 4.8, we might appreciate that the use of narrow bands reduces the smoothness. However, the dispersion of the photo-zs of the objects that fall in each cell (right panel) has a mean value of 0.048 when including magnitude colors of the 40 narrow bands and 0.059 when they are not included. We think that the former case gives a better dispersion because we have a smaller galaxy sample since we only consider objects with observations in the magnitude system, i.e. positive fluxes, in all bands and a lot of galaxies have some narrow band flux missing.

Luptitudes To be able to use all the information of the narrow bands we use another quantity derived from the flux measurement called luptitude. We have avoided

to use directly fluxes as input variables because they span several orders of magnitudes. So we chose magnitude, specifically the difference between magnitudes, i.e colors, as input features since colors span a narrow range of values. We remember that magnitudes, m, are defined as:

$$m_{\rm k} = m_{0,\rm k} - 2.5 \log f_{\rm k} \tag{4.9}$$

where $m_{0,\rm k}$ is the zero pint in the k PAUS narrow bands. According to this definition, non-positive fluxes $f_{\rm k}$ are not defined in the magnitude system, loosing these undefined values. Therefore we try to use luptitudes, μ, as defined in Buchs et al. (2019):

$$\mu_{\rm k} = \mu_0 - a \sinh^{-1}\left(\frac{f_{\rm k}}{2b}\right) \tag{4.10}$$

where $\mu_0 = m_0 - 2.5 \log b$, $a = 2.5 \log e$ and b is a parameter that determines the point at which the behavior of luptitudes changes between logarithmic and linear. The optimal value of b, as defined in Lupton et al. (1999), is $b = 1.042\sigma$ where σ^2 is the variance of the flux. We assume all fluxes have a similar error for simplification and choose $b = 100$. Then in the luptitude system, negative and zero fluxes are defined so that at small fluxes luptitudes behave as a flux and for large fluxes as a magnitude. Specifically, we use the difference of luptitudes as input features, the same way we have used the difference of magnitudes before.

Before we had imposed that objects in the training sample must have information in all the bands used as input features. That resulted in the exclusion of several galaxies that had a non-positive flux in some narrow band, which led to a non-defined magnitude value. The advantage of using luptitudes instead of magnitudes as input features is that no information is lost in the conversion from the fluxes of the narrow bands to luptitudes.

Now, the number of objects with information in all bands available to train the maps in Fig. 4.9 is the same when using only the luptitudes colors from the 6 bands of COSMOS (left panel) and when adding the luptitudes colors from the 40 narrow bands of PAUS (central panel). If we compare the COSMOS and PAUS sample with luptitudes as training features and the sample with magnitudes (whose map is shown in the central plot of Fig. 4.8), we see that the former sample has 38 765 objects with information in all the bands whereas the latter sample only has 18 140 objects.

Despite the enlargement of the training sample provided by the use of luptitudes, there is an increase of the dispersion of photo-z of objects that fall within the same

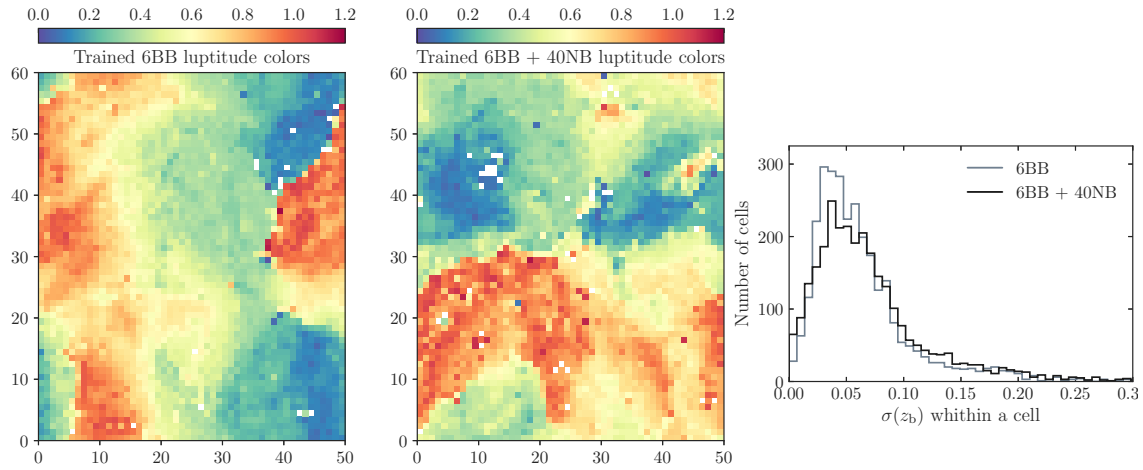

FIGURE 4.9: *Left:* Median photo-z of PAUS of galaxies that fall within each cell. The self-organizing map was trained using as input variables the difference of luptitudes $u^* - B$, $B - V$, $V - r$, $r - i^+$, $i^+ - z^{++}$ from 6 broad bands of COSMOS. The SOM is populated with PAUS galaxies in COSMOS field. *Center:* The input variables to train the map are the same luptitude colors as the ones used in the SOM of the left panel, in addition to the difference of luptitudes between the 40 narrow bands of PAUS and the i^+ luptitude of COSMOS. Having a total of 45 input features. The median photo-z of PAUS of galaxies that fall within each cell is shown in the plot. *Right:* The dispersion of photo-z of objects that populate each cell is computed through the standard deviation statistic, $\sigma(z_b)$, when the map is trained using only the difference of luptitudes of the 6 broad bands of COSMOS (grey) and when adding the difference of luptitudes between the 40 narrow bands of PAUS and the i^+ luptitude of COSMOS (black). The distribution of cells that have certain photo-z dispersion is shown in the plot. Only cells with more than one object have been taken into account.

cell. The increase of the dispersion can be seen by comparing the right panel of Fig. 4.9 with the same panel in Fig. 4.8 where magnitude colors are used as training values. So with luptitudes as training features the SOM groups less similar galaxies together in the same cell. The average value of the dispersion when using the luptitude colors of the 6 broad bands is 0.064 and by adding the luptitude colors of the 40 narrow bands the mean becomes 0.073.

Final choice of input features To make a final decision on the training features, we take into account the visual smoothness of the median photo-z across the map and the average value of photo-z within a cell, which indicates the capability of the map to group similar galaxies within the same cell and in adjacent cells. We also consider that fluxes from all used bands are available in the system we choose to have the maximum training data. Therefore, we think that best choice of input features

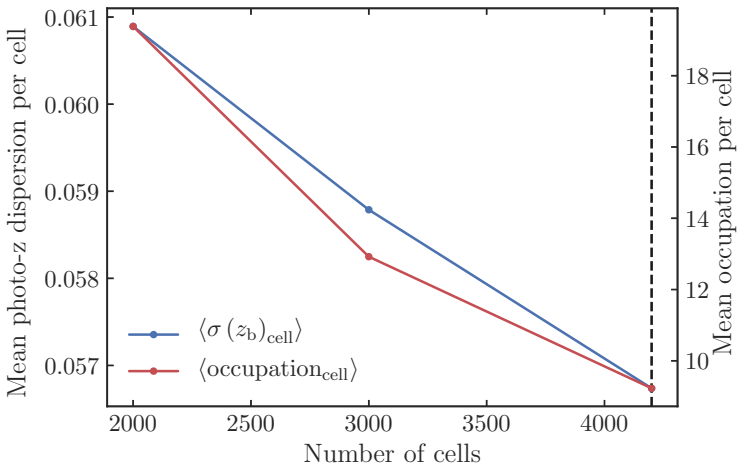

FIGURE 4.10: Variables to determine the best number of cells for a rectangular map with toroidal configuration. In all cases, the input training variables are the magnitude colors of the 6 broad bands of COSMOS. The mean occupation of objects that fall within a cell as a function of the number of cells is shown in red. This measurement loosely asses if individual cells represent a correct number of objects to ensure there is no over or underrepresentation. In blue it is shown the mean dispersion of photo-z of objects that populate each cell as a function of the number of cells. This parameter shows the closeness of the characteristics of objects to weight vector of each cell, i. e, how well the colors of a cell represent the colors of objects, and if the colors of objects that fall within a cell translates to similar photo-z between them.

to create a map to study the color space and photo-z of PAUS are the magnitude colors $u^* - B$, $B - V$, $V - r$, $r - i^+$, $i^+ - z^{++}$ from the 6 broad bands of COSMOS as shown in the left panel of Fig. 4.8.

Choice of number of cells

An influential parameter in the representative capacity of the map is the number of cells. Cells are the smallest unit or class at which we classify galaxies. Similar to other machine learning algorithms, if the number of groups or classes are too high, the cells may not be well characterized because they are populated with few objects and the map may overfit the data, meaning that rare galaxies or objects are overrepresented and the map classification can not be applied to other samples. On the contrary, if the number of cells is too low, the same unit must represent a larger group of objects with more dissimilar characteristics. To compromise and find the best number, we train maps with different sizes and take a look at the mean of the dispersion of the photo-zs in each cell and the average occupation per cell. The dispersion indicates

how close galaxies are within a cell to the characteristics defined by the weight vector of the cell. The average occupation is an indicator of how common the characteristics of the cell are among the galaxies. If the average occupation per cell is too low we may be producing an overfitted map. Other indicators can be used to asses the representative power of the map.

We generate maps with different sizes, all trained with the same input parameters. In Fig. 4.10, we see that the number of objects per cell and the dispersion of the photo-zs within a cell drop as the number of cells increases. Due to time constrains we do not generate more maps, as we can already see that for our sample of 38 765 objects the average occupation per cell rapidly decreases but the dispersion of photo-z reduces more slowly. Our choice of optimum number of cells is 4200 that corresponds to a map with average number per cell of 9 objects, mean photo-z dispersion of 0.0567 and median dispersion of 0.0419. This dispersion of galaxies within a cell is comparable to the typical photo-z error achieved using broad band photometry, which usually is $\sigma_{68} \sim 0.05(1+z)$ (Hildebrandt et al., 2012). If we wanted smaller dispersion we should increase the number of cells at the expense of overfitting the map which we want to avoid.

4.3.2 Analysis of the color space map and redshift relation

In this section, we use the SOM technique to empirically map the high dimensional color space of PAUS into a two dimensional grid so we can take a look at the photo-z - color relation of the survey. The SOM have been trained using as training features the photometric data of the $u^* - B$, $B - V$, $V - r$, $r - i^+$, $i^+ - z^{++}$ colors from the 6 broad bands of COSMOS, which are the same bands that together with the 40 narrow bands of PAUS are used to compute the photo-z of PAUS. Each cell of the map has an associated weight vector with the same dimensions and features used to train the map. The resulting values of the final weight vectors for each component are shown in Fig. 4.11. After the training, the map is populated with galaxies by matching the colors of each galaxy to the colors of weight vectors with more similar values, following Eq. 4.1. The number of objects in each cell is shown in the top-left panel in Fig. 4.11. To verify that the resulting cells of the map are representative of the colors of PAUS, we compare the distribution of colors of PAUS to the distributions of colors of the weight vectors of each cell. As it can be seen in Fig. 4.12, both

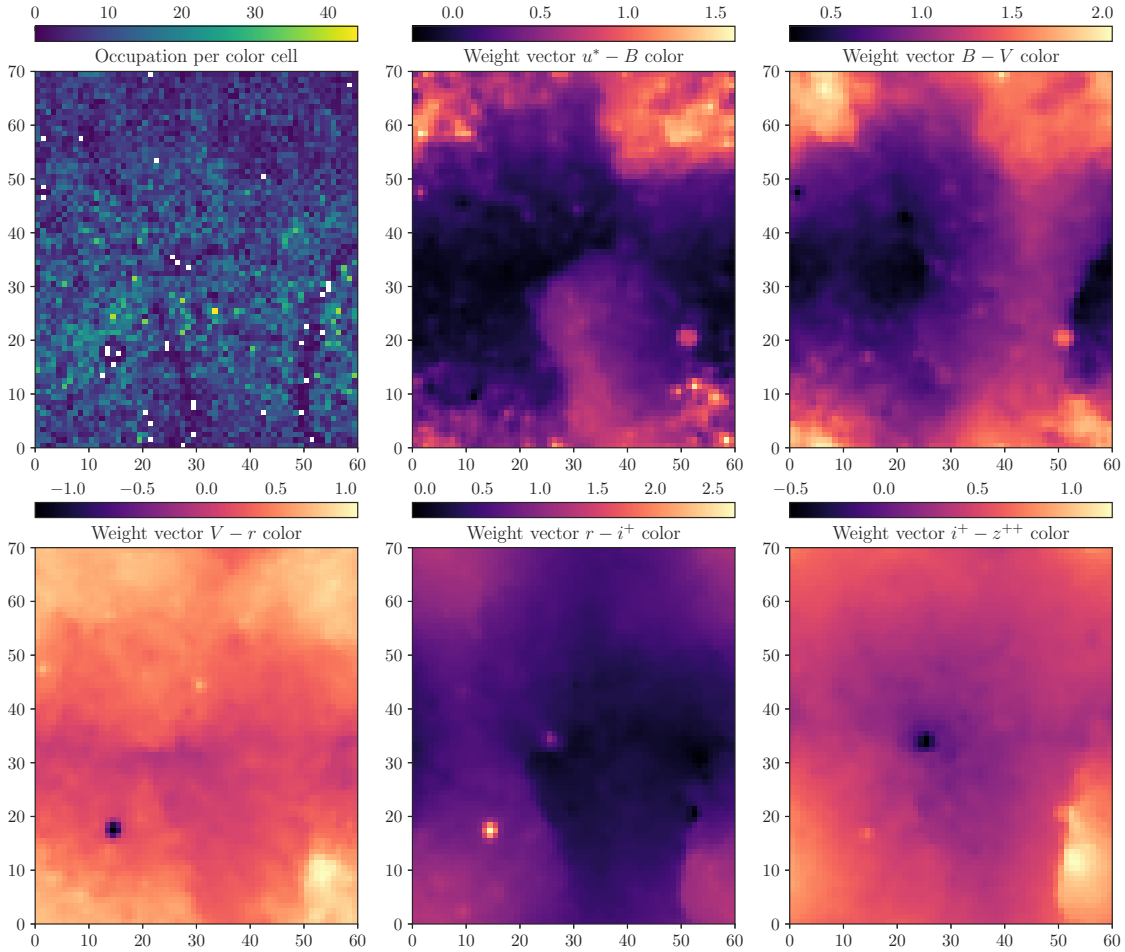

FIGURE 4.11: Self organizing map of 4200 cells trained with colors from the 6 broad bands of COSMOS for PAUS objects in COSMOS field. *Upper left:* Occupation density of galaxies associated to each cell. *Upper center to bottom right:* Map colored by the weigh vector of each cell corresponding to $u^* - B$, $B - V$, $V - r$, $r - i^+$, $i^+ - z^{++}$ colors.

sets of colors cover the same range and have similar shape implying that the SOM is representative enough of the training sample.

Color-redshift coverage

The color space defined in this work is not the full color space of all the PAUS survey but a subsample of objects located in the COSMOS field and whose photo-zs have been determined with a template fitting code specifically designed for objects observed with the narrow bands of PAUS (Eriksen et al., 2019). As we mentioned before, we focus in this subsample because it can be matched to the COSMOS objects with

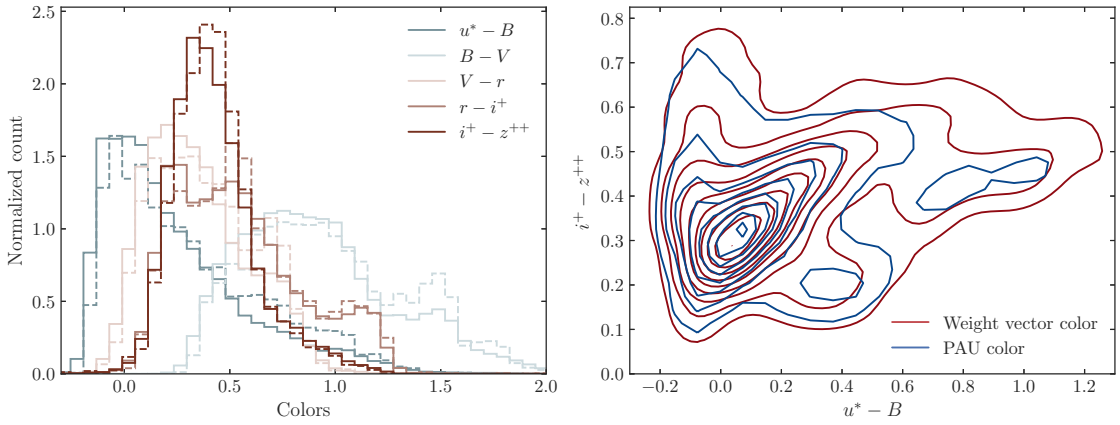

FIGURE 4.12: *Left:* Comparison of color distributions of PAUS galaxies (solid lines) and of the resulting weight vectors of the SOM (dashed lines). *Right:* Color-color diagram of PAUS galaxies (blue) and weight vectors (red).

spectroscopic redshifts that are used to validate the photo-z performance of PAUS in Eriksen et al. (2019) and in Eriksen et al. (2020). As proved in Masters et al. (2017), current available spectroscopic data are systematically incomplete in redshift and thus in color. Therefore photo-z validations lack comparable objects at high redshift. This lack of representativeness in redshift is specially problematic when spectroscopic data are used to train and calibrate algorithms to determine photo-zs since machine learning algorithms are unable to accurately extrapolate information in a space not defined in the training sample.

Here, we show that the incompleteness of spectroscopic objects in color space is also true for PAUS in the COSMOS field. In Fig. 4.13, we show the map colored by the average photo-z of galaxies assigned to each cell and by the average spectroscopic redshift of galaxies with the highest confidence redshift. In the right panel we see that the spectroscopic information does not cover all the parameter space since there are empty regions of color space. There are 1140 cells (27.14%) with missing spectroscopic redshift if only high confidence redshifts are considered according to a confidence class ($3 \leq$ confidence < 5) provided by Lilly et al. (2007). If less reliable redshifts are not removed, 541 cells (12.88%) still remain empty. With a relaxed quality requirement more color space is filled, however the photo-z performance is usually checked with the highest confidence redshifts, otherwise biases are introduced. Therefore it is important to complete the coverage of empty regions to avoid biased calibrations or validations. In Fig. 4.14, we take a closer look at the fraction of objects that fall in

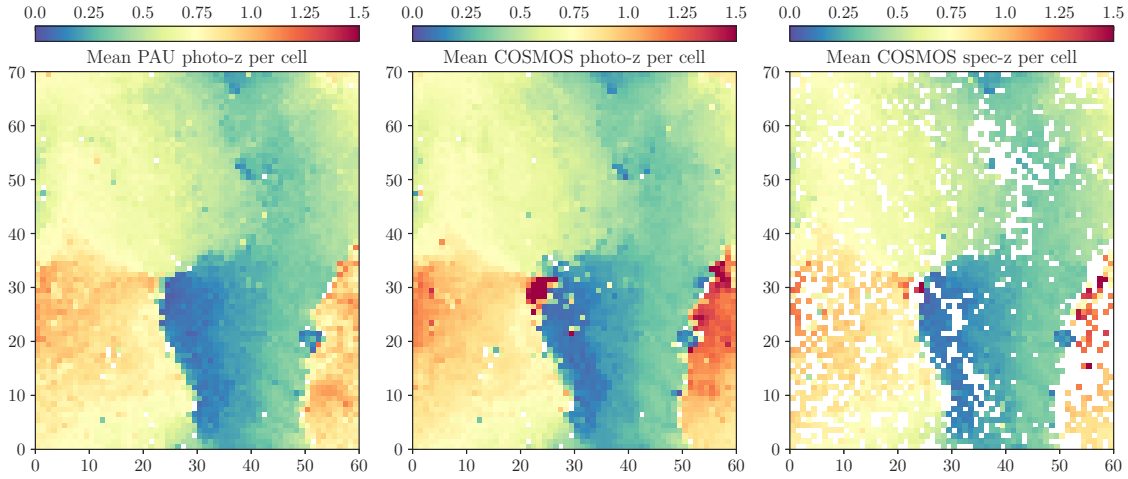

FIGURE 4.13: Different redshift estimates across the map computed as the average value of redshift of objects of PAUS in COSMOS field that fall within each cell. *Left:* Photo-z of PAUS computed through a template fitting method taking into account the 40 narrow bands of the survey and the 6 broad bands of COSMOS. *Center:* Photo-z of COSMOS (Laigle et al., 2016). *Right:* Spectroscopic redshift of COSMOS with quality flag for high confidence redshift $3 \leq$ confidence < 5 (Lilly et al., 2007). White cells means there is no high confidence spectroscopic redshift in that part of the color space.

cells that lack spectroscopic redshift information. We see that the fraction of galaxies without spectroscopic redshift counterpart largely increases for redshift larger than 0.7, which corresponds to fainter galaxies.

Cells without spectroscopic counterpart and anomalous cells

Besides the already known lack of spectroscopic redshift mainly at high redshift, we take a closer look at cells that miss spectroscopic redshift information to try to identify further characteristics or tendencies.

Simultaneously observing cells with missing spectroscopic redshifts and the occupation and resulting colors of the map in Fig. 4.11, we see that some empty cells match areas with lower occupation and correlate with extreme colors of the weight vectors defined by the SOM. We can deduce that missing spectroscopic redshift may correspond to objects that are less likely to be found or observed. Therefore, if we wanted to focus efforts to fill the color space with spectroscopy to compare the photo-z with, the SOM technique can help us identify the regions in color space that potentially result too difficult or expensive to observe.

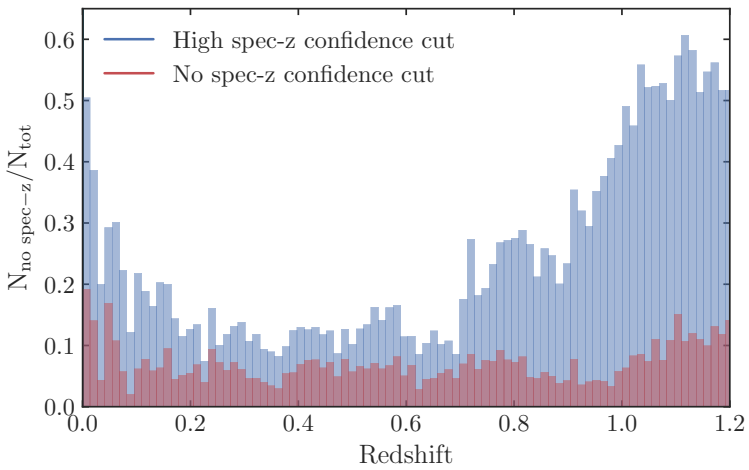

FIGURE 4.14: Fraction of the total number of objects of PAUS in the COSMOS field (N_{tot}) and the number of galaxies without spectroscopic information within the cell they fall ($N_{no\,spec-z}$) as a function of redshift. When only spectroscopic redshift with high confidence level are considered (blue) and when no spectroscopic redshift quality criteria is applied (red).

In Fig. 4.15, we show the redshift dispersion of objects that populate each cell. The dispersion indicates how broad the redshift range described by the cell is. In other words, it identifies cells with large variance in redshift between objects or cells whose objects gather around multiple redshift peaks. In comparison with Fig. 4.13, it is appreciable that cells without spectroscopic redshift correlate with cells that have higher photometric dispersion. Regions with large dispersion overall fall in areas of colors space near the boundary between high and low redshift objects. In the sharpest boundaries separating redshift, the dispersion is higher and the occupation is low indicating that similar colors can correspond to very different redshifts. If more objects in the COSMOS fields are observed with all 40 narrow bands of PAUS (emphasizing that all bands must be observed to be able to incorporate the colors in the training of SOM), the extra color information could help to break the color-redshift degeneracies of the map. To highlight the edge between redshift regimes, we define the smoothness of the map by computing for each cell the average difference of PAUS photo-z between the cell and their adjacent cells. The smoothness is shown in the left panel of Fig. 4.16. Overall, the sharper redshift boundaries have higher dispersion.

To further investigate the map characteristics we asses the representativeness of the map to the data by computing the average difference between the colors of objects

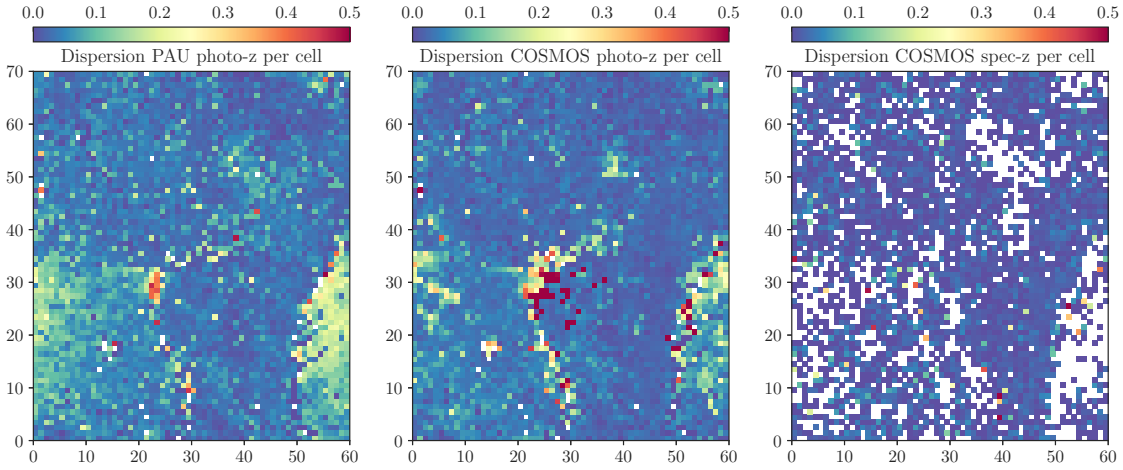

FIGURE 4.15: Standard deviations of redshift of galaxies that fall within each cell. *Left:* Dispersion of PAUS photo-z. *Center:* Dispersion of COSMOS photo-z. *Right:* Dispersion of highly secure spectroscopic redshift of COSMOS.

that fall in a cell and the weight vector colors defined by the SOM divided by the number of colors. This measurement is called quantization error and is defined in Masters et al. (2015). The resulting quantization error across the map is shown in the center panel of Fig. 4.16. The definition is similar to the distance in equation 4.1 but without weighting by the photometric error of the observed colors to avoid that smaller photometric errors produce artificially large quantization errors. As seen in the figure, the quantization error is large in the regions where the redshift changes rapidly from cell to cell and even larger in the corners of the map which correspond to the regions with the most extreme colors (see Fig. 4.11). Objects with large quantization error represent galaxies that are not well represented in the SOM, not even by the extreme colors. The inability to represent these galaxies may be because there is a small amount of objects with strange colors that the SOM can not learn from. Then, if the SOM is trained correctly, it could be used as a tool to detect unusual objects.

The ability of the map to group and represent galaxies can be potentially used to identify anomalous objects in color space. A more exhaustive analysis on this technique is left for future work. However, we take a quick look at the behavior across the map of a parameter that asses photo-z quality called odds (Benítez, 2000). To check if the map can be used to detect a pattern for objects with higher uncertainty in their redshift determination. The parameter is used in Eriksen et al. (2019) to determine the uncertainty in the determination of the photo-z of PAUS. The odds

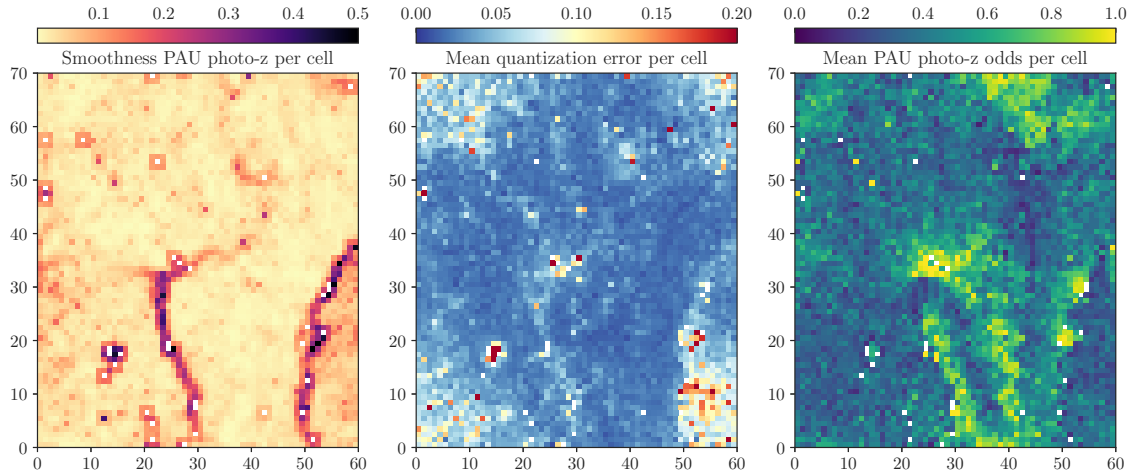

FIGURE 4.16: *Left:* Smoothness of the map defined as the average difference between PAUS photo-z of adjacent cells. *Center:* Mean quantization error per cell defined as the difference between colors of the objects that fall in a cell and the weight vector color of the cell divided by the number of colors. *Right:* Mean photo-z odds of PAUS (see text for further details).

are defined as:

$$\text{odds} = \int_{z_\text{b}-\Delta z}^{z_\text{b}+\Delta z} p(z)\,\mathrm{d}z, \tag{4.11}$$

where z_b is the peak in the redshift probability distribution $p(z)$ determined by the template fitting method used for PAUS and Δz delimits the interval around the peak that in Eriksen et al. (2019) is 0.0035. The odds measure the fraction of the probability distribution around the peak. We compute the average value of the odds parameter of galaxies that fall in each cell and show them in the right panel of Fig. 4.16. The better odds are found at cells with average lower redshift, specially for areas defined as lower redshift by the COSMOS photo-zs rather than the PAUS photo-z. The odds are lower in areas that do not have secure spectroscopic redshift counterparts and in some of the areas with extreme colors. However, we do not find a straight and simple relation between the parameters of the SOM to directly detect objects with high photo-z uncertainty or quality.

We believe that a parameter or set of parameters that better model the redshift and color uncertainties across the map could help to detect and classify strange objects in the PAUS survey. However, this project is left for future work.

4.4 Summary and conclusions

Current spectroscopic surveys cannot sample all the redshift and color space covered by modern imaging surveys. To achieve the scientific goals for weak lensing and galaxy clustering analyses of cosmological surveys such as Euclid, accurate photo-zs should be derived. To achieve that, a complete coverage of spectroscopic samples to calibrate the $n(z)$ is necessary. The C3R2 project aims to identify and observe the galaxies that are needed to have a fully representative spectroscopic sample, specifically for the Euclid survey. They define a representative spectroscopic sample by empirically calibrating the galaxy color–redshift relation with the SOM, a machine learning algorithm. The SOM projects the high-dimensional galaxy color space given by the photometric sample onto a two dimensional grid. Each cell of the grid has assigned a unique value of colors determined by the SOM that corresponds to the mean spectral energy distribution of galaxies from the photometric sample that fall in that cell. Galaxies are grouped in cells by the closeness of their colors to the ones assigned to each cell. A value of redshift is assigned to each cell by averaging the redshift of galaxies that fall within it. This method allows us to detect which regions of the galaxy color space are not represented in the currently available spectroscopic samples.

Spectroscopic observations for the C3R2 are being carried out in the COSMOS, VVDS, EGS, SXDF and ECDFS fields with the VLT and Keck facilities. In the first part of this chapter, our aim was to select galaxies to be observed in the ECDFS field by the VLT telescope that would contribute to fill the cells of color space without spectroscopic redshift. To assess the galaxies worth to be observed, first we needed to determine in which region of the color space they belonged. In other words, the galaxies had to be placed in the SOM defined with the C3R2 photometry. The photometry in the ECDFS was different from the photometry used to define the SOM. Before placing the galaxies in the SOM their photometry had to be converted to the C3R2 photometry system.

The C3R2 project trained the SOM using the photometry from the COSMOS, VVDS and EGS fields. This photometry corresponded to the *ugriz* optical bands from CFHTLS Deep Fields, the $JHK_{\rm s}$ near-infrared bands from WIRDS and the Y band from UltraVISTA. Later on, they included the spectroscopy from the SXDF and ECDFS fields. The photometry of the ECDFS field needed to be incorporated into the C3R2 color system in a consistent way to take advantage of the available spectroscopy

and to be able to reliably place galaxies on the general color map. Therefore, we needed enough data with photometry in both systems to compute the transformation between the C3R2 color system and the ECDFS photometry. However, photometric data in the common C3R2 system was only available for a small area of the ECDFS field. We used the VVDS field to provide the photometry transformations between systems since it had a larger area with data in the C3R2 system. We chose the photometry of the DES survey, whose photometry was also available in the ECDFS field, to calibrate the relations to transform the optical bands of ECDFS galaxies into C3R2 photometry. We used the MUSYC survey to provide the transformations for the near-infrared bands. To transform the MUSYC photometry we used the small area in common with C3R2 in the ECDFS.

Once the photometry of the ECDFS field was transformed to the C3R2 system, galaxies were placed in the SOM and an observational priority was assigned. To determine the priority of each cell, a list of galaxies from COSMOS with already assigned priority by the C3R2 project was used. The priority corresponding to each cell was computed as the average priority of all the COSMOS galaxies that belong to the same cell. The target selection priority of our ECDFS galaxies for the VLT observations was given by the priority value of the cell they fall into. The C3R2 priority criteria prioritized the amount of new information a galaxy contributes to the color-redshift mapping and the probability of successfully obtaining a spectroscopic redshift given the instrument, exposure time, and expected galaxy properties.

In the second part of this chapter, we used the SOM technique to explore the color space of the photometry of galaxies in PAUS and study the relation between colors and photo-zs in the survey. For the study we considered a similar sample as the one used for the analyses of the photo-zs performance in the papers of the collaboration. So we could complement the photo-zs analyses of PAUS. To train the SOM we wanted to choose observable variables related to the spectral energy distribution of galaxies as input features since they are used to determine photo-zs. However, the color space of the PAUS survey composed of 40 narrow band and 6 broad bands filters was a highly complex multi-dimensional space. So the definition of the SOM itself was complex given the number of input parameters. The map is sensitive to the chosen input features used to define it. We experimented with these parameters by training the map with different combination of input variables.

The initial choice of input parameters to calibrate the multidimensional space were the magnitude colors of the 40 narrow bands of PAUS. To properly train the

map, we imposed that all galaxies had information in all bands, i.e that their fluxes were non-negatives. However, a lot of galaxies had some missing or negative narrow band flux, so the training sample was small. Moreover, the high dimensional data required a large amount of data for the SOM to be able to correctly project the high dimensional data into a two-dimensional map. With so few objects available, we tried a subset of the narrow bands filters as input parameters to increase the number of objects with information in all the input parameters and to reduce the dimensionality of the data. With this choice, the dispersion of the galaxy color within the cells was too high. The dispersion of the properties of the cell was an indicator of the ability of the map to correctly partition the input color space without degeneracies. The third choice was to only use the broad band colors with and without the narrow band colors. In both cases, with the presence of the broad band colors, the map was smooth and the dispersion within the cells was low. We also tried using luptitudes instead of magnitudes as input parameters. Luptitudes allowed us to use the negatives fluxes information and thus increased the number of galaxies with information in all bands. However, we saw that the lowest dispersion within a cell and the maximum smoothness of the map was achieved with the use of the six broad bands colors as input features.

Current available spectroscopic data are systematically incomplete in covering the redshift and color space spanned by galaxies. This lack of representativeness in redshift is specially problematic when spectroscopic data are used to train algorithms to determine photo-zs and to validate the photo-zs performance by comparing them to spectroscopic redshifts. Once the SOM was defined, we populated it with PAUS galaxies and computed the average of the photo-z and spectroscopic redshift of the galaxies that fall within each cell. The 27.14% of cells had missing spectroscopic redshifts if only high confidence redshifts were considered. We saw that the fraction of galaxies without spectroscopic redshift counterpart largely increased for redshifts larger than 0.7, which corresponded to fainter galaxies. If less reliable redshifts were not removed, more color space was filled but 12.88% of cells still remained empty. However, the photo-z performance is usually checked with the highest confidence redshifts. Therefore it is important to complete the coverage of empty regions to avoid biased validations.

We tried to identify further characteristics of the cells with missing spectroscopic redshifts. We saw that some empty cells matched areas with lower occupation and correlated with extreme colors of the SOM weight vectors. So missing spectroscopic

redshifts may correspond to objects that are less likely to be observed. Therefore, if we want to focus efforts to fill the color space with spectroscopy, the SOM technique can help us identify the regions in color space that potentially result too difficult or expensive to observe, similar to what was done in the C3R2 project.

The dispersion of redshift indicated how broad the redshift range described by the cell was. Cells without spectroscopic redshifts also correlated with cells that had higher photo-z dispersion. Regions with large dispersion overall fall in areas of colors space near the boundary between high and low redshift objects, indicating that similar colors can correspond to very different redshifts. If more objects in the COSMOS fields were observed with all the 40 narrow bands of PAUS, the extra color information could help to break the color-redshift degeneracies of the map.

We believed that the ability of the SOM to group similar galaxies together could be used to detect galaxies with strange colors or with photo-zs that were miscalculated according to their colors. So we defined a few parameters that model the dispersion of the colors within a cell, the difference of redshift between nearby cells and the average value of uncertainty in the cells photo-z. For example, we found a certain degree of correlation between cells with high dispersion of colors, low occupation and no spectroscopic redshift but still the correlation was not strong enough to use the map as detector of strange objects. We still think that a set of parameters that better characterize the redshift and color uncertainties across the map could help to detect and classify strange objects in the PAUS survey. However, further exploration is left for future work.

Chapter 5

Optimization of the photometric sample of Euclid for GC analyses

The goal of Stage-IV dark energy surveys (Albrecht et al., 2006), such as Euclid and Rubin-LSST is to measure both the expansion rate of the Universe and the growth of structures up to redshift $z \sim 2$ and beyond. These surveys will allow us to constrain a large variety of cosmological models using cosmological probes like weak gravitational lensing (WL) and galaxy clustering. Stage-IV surveys can be classified into spectroscopic and photometric surveys, depending on whether the redshift of objects is estimated observing the full spectral energy distribution, which requires more observational time thus less objects can be observed, or through multi-band filters, which requires less observational time and allows to observe more objects but the precision on the redshift estimates decreases. Galaxy clustering analyses are usually performed with data coming from spectroscopic surveys, while the data obtained from photometric surveys are generally used for WL analyses. However, given the current and future precision of our measurements, the signal we can extract from galaxy clustering analyses using photometric surveys is far from being negligible (see e.g. Abbott et al., 2018a; van Uitert et al., 2018; Euclid Collaboration: Blanchard et al., 2019; Tutusaus et al., 2020). Therefore, upcoming surveys can increase their constraining power if they optimize their photometric samples to include galaxy clustering studies in addition to WL analyses. The main aim in this chapter is to perform such an optimization study for the Euclid photometric sample. The work and results described here will appear in Pocino et al. in prep.

As explained in more detail in Sec. 2.3, the Euclid satellite will observe over a billion galaxies through an optical and three near-infrared broad bands. Given the specifications of the satellite, the combination of Euclid and ground-based surveys can enrich the science exploitation of both. The WL analysis of Euclid data requires an

accurate knowledge of the redshift distributions of the samples used for the analysis. Euclid photometric data alone cannot reach the necessary photo-z performance and additional ground-based data are required. One of the best Stage-IV complementary surveys to Euclid is Rubin-LSST since it greatly overlaps in area, covers two Euclid Deep Fields and reaches a faint photometric depth that will lead to better photo-z estimation. Euclid will perform both a spectroscopic and a photometric galaxy survey that will allow us to determine cosmological parameters using its three main cosmological probes: galaxy clustering with the spectroscopic sample ($\mathrm{GC_s}$), galaxy clustering with the photometric sample ($\mathrm{GC_{ph}}$), and WL. In this work we consider the addition of ground-based optical photometry to Euclid in order to assess the optimal photometric sample to provide the tightest cosmological constraints focusing on the $\mathrm{GC_{ph}}$ analysis and its cross-correlation with WL, called galaxy-galaxy lensing (GGL).

We will optimize the Euclid sample of galaxies detected with photometric techniques by performing realistic forecasts of its cosmological performance and observing the improvement on the cosmological constraining power of different galaxy samples. When performing galaxy clustering analyses with a photometric sample there are several effects that need to be taken into account such as galaxy bias or photo-z uncertainties, among other effects. Here, we try to follow the procedures one would perform in a real data analysis when selecting the samples for the analysis. For that purpose, we use the Euclid Flagship simulation (Euclid Collaboration, in prep). For a given expected limit of the photometric depth, we select the galaxies included within that magnitude limit and use a machine learning photo-z method to study the optimal way to split the catalog into subsamples for the analysis. We generate realistic redshift distributions, $n(z)$, for the chosen subsamples and estimate their galaxy bias, $b(z)$. We study the constraining power of these samples when we modify the number and type of tomographic bins, and when we reduce the sample size by performing a series of cuts in magnitude.

The chapter is organized as follows. In Sec. 5.1 we introduce the Flagship simulation and describe how we create photometric samples with different selection criteria. We define the set of galaxy samples that will be used throughout the chapter, and explain how we estimate the photo-zs. In Sec. 5.2 we describe the cosmological model and we detail the forecast formalism in Sec. 5.3. In Sec. 5.4 we present the results of the optimization when changing the number and type of tomographic bins, and we study the dependency of the cosmological constraints on photo-z quality and sample

size. Finally, we summarize in Sec. 5.5.

5.1 Generating realistic photometric galaxy samples

The cosmological constraining power of Euclid will depend on the external data available as it will dictate the photo-z performance of the samples to be studied. In order to study the impact of the available photometry, we create six samples selected with different photometric depths. For each sample, we compute the photo-z estimates using machine learning techniques taking into account the expected spectroscopic redshift distribution of the training sample. We use these photo-z estimates to split each sample into tomographic bins for which we can compute their photo-z distributions and galaxy bias from the simulation. These $n(z)$ and $b(z)$ are then used to forecast the cosmological performance. In this section we provide a detailed description on how we obtain the realistic photo-z estimates of the Euclid galaxies that are later used in the forecast. We first present the cosmological simulation used to extract the photometry and the galaxy distributions. We then explain how we generate realizations of the photometry for the simulated galaxies taking into account the expected depth of the Euclid and ground-based data. We finally present the method used to estimate the photo-z.

5.1.1 The Flagship simulation

In this work we consider the Flagship galaxy mock catalog of the Euclid Consortium (Euclid Collaboration, in preparation) to create the different samples. The catalog uses the Flagship N-body dark matter simulation (Potter et al., 2017) where galaxies are assigned to dark matter halos using Halo Abundance Matching (HAM) and Halo Occupation Distribution (HOD) techniques. The galaxy mock generated has been calibrated using local observational constraints, such as the luminosity function from Blanton et al. (2003) and Blanton et al. (2005a) for the faintest galaxies, the galaxy clustering measurements as a function of luminosity and color from Zehavi et al. (2011), and the color-magnitude diagram as observed in the New York University Value Added Galaxy Catalog (Blanton et al., 2005b). The spatial distribution of a subsample of Flagship galaxies is shown in Fig. 5.1. The cosmological model assumed in the simulation is a flat ΛCDM model with fiducial values $\Omega_{\rm m} = 0.319$, $\Omega_{\rm b} = 0.049$, $\Omega_\Lambda = 0.681$, $\sigma_8 = 0.83$, $n_{\rm s} = 0.96$, $h = 0.67$. The N-body simulation ran in a 3.78 $\rm h^{-1}$

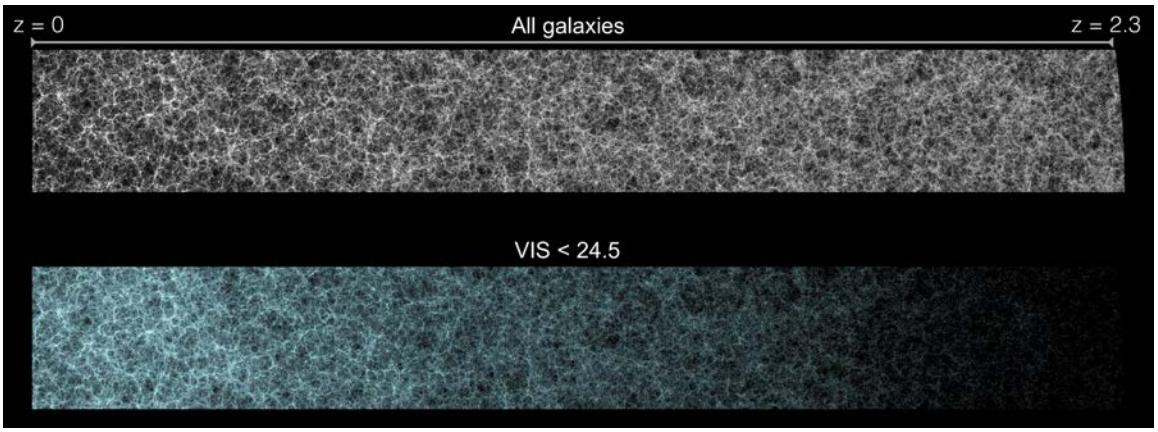

FIGURE 5.1: Two-dimensional projection of a portion of the full light-cone simulation of Flagship mock galaxies in the Euclid survey. The top panel is the full sample of galaxies in the mock. The bottom panel is the sub-samples expected from observations in the VIS band. *Credit:* J. Carretero, P. Tallada and S. Serrano for ICE/PIC/U.Zurich and the Euclid Consortium Cosmological Simulations SWG.

Gpc box with particle mass $m_{\rm p} = 2.398 \times 10^9\ {\rm h}^{-1}\,{\rm M}_\odot$. The catalog contains about 3.4 billion galaxies over 5000 deg^2 and extends up to redshift $z = 2.3$.

For this study we select an area of 402 deg^2, which corresponds to galaxies within the range of right ascension $15^\circ < \alpha < 75^\circ$ and declination $62^\circ < \delta < 90^\circ$. All the photometric galaxy distributions obtained in this patch are extrapolated to the 15 000 deg^2 of sky that Euclid is expected to observe. Note that the selected area is large enough to minimize the impact of sample variance, but small enough to allow for the production of several galaxy samples in a reasonable amount of time. After the photometric uncertainty is added to the photometry of each galaxy, we perform a magnitude cut in $\rm m_{VIS} < 25$ that leads to a number density of about 41.5 galaxies per arcmin^2.

5.1.2 Photometric depth

Each galaxy observation will lead to a measured value of its magnitude and its associated error. The magnitude depth is usually given as the magnitude at which the median relative error has a particular value. In galaxy surveys it is customary to express the depth at a signal-to-noise of 10 for extended objects, that is, when the value of the noise is one tenth of its signal. As explained in detail below, we generate realizations of the photometric errors for a given survey taking into account its magnitude depth and scaling the values of the errors at other magnitudes assuming

background limited observations, that is, that the background signal dominates the contribution to the error.

We simulate four different photometric survey depths. Table 5.1 shows their magnitude limits. The first column corresponds to a combination of Euclid and ground-based photometric depth expected to be achieved in the Southern hemisphere. We label this case as optimistic and it is the deepest case we will study. The magnitude limits for the optical bands are for extended sources at 10σ, similar to those expected from Rubin-LSST. The values for Euclid correspond to a 10σ detection level for extended sources. In addition to the magnitude limits expected in the South, we also want to investigate how the cosmological constraints degrade as the depth is reduced. We investigate three other cases. First, a case were the depth in optical bands are reduced by a factor of two in signal-to-noise ratio. The second column shows the magnitudes limits for this case where the optical bands are reduced by a magnitude value of 0.75. This column represents a possible case where the Rubin-LSST data have a reduced depth in areas outside its main footprint. We also study a case were the limiting fluxes of Euclid are also reduced by 0.75 magnitudes, shown in the third column. Lastly, we explore a case where the ground-based data is degraded by a factor of five in signal-to-noise but the Euclid space data remains at their nominal depth values. This broadly represents the depth that can be achieved from other ground-based data in the Northern hemisphere.

For each survey case, we generate a galaxy catalog drawn from the Flagship simulation. We assign observed magnitudes and errors with the following procedure. First, we compute the expected error for each galaxy, taking into account its magnitude in the Flagship catalog and the magnitude limit of the survey as given in Table 5.1. We assume that the observations are sky limited (the noise is dominated by the shot noise of the sky), and therefore we scale the ratio of the signal to noise between two galaxies i and j as the ratio of their fluxes

$$\left(\frac{S}{N}\right)_i = \left(\frac{S}{N}\right)_j \frac{f_i}{f_j}\,, \tag{5.1}$$

where f_i is the observed flux of galaxy i detected at signal-to-noise ratio $(S/N)_i$. The magnitude (flux) limits in Table 5.1 give us the fluxes corresponding to a signal-to-noise ratio of 10, $f_{10\sigma}$, and therefore we can compute the expected signal-to-noise at

TABLE 5.1: Limiting coadded depth magnitudes for extended sources at 10σ used in each sample.

Band		Optimistic	Ground based degraded -0.75	All degraded -0.75	Ground based degraded -1.75
Ground based	u	25.55	24.8	24.8	23.8
	g	26.75	26.0	26.0	25.0
	r	26.95	26.2	26.2	25.2
	i	26.25	25.5	25.5	24.5
	z	25.45	24.7	24.7	23.7
	y	24.15	23.4	23.4	22.4
Euclid	$\mathrm{m_{VIS}}$	24.6	24.6	23.85	24.6
	Y	23	23	22.25	23
	J	23	23	22.25	23
	H	23	23	22.25	23

which a galaxy of a given magnitude is detected as

$$\left(\frac{S}{N}\right)_i = 10\frac{f_i}{f_{10\sigma}}\,. \tag{5.2}$$

Using the definition of signal-to-noise, $(S/N)_i = f_i/\Delta f_i$, we can compute the expected flux error for each galaxy as

$$\Delta f_i = \frac{f_{10\sigma}}{10}\,. \tag{5.3}$$

The fluxes in the Flagship catalog correspond to the real fluxes of each galaxy. Whenever we observe these galaxies in a given survey, we detect a realization of the real flux. For our study, we generate realizations of the observed fluxes f_i^* for each survey as

$$f_i^* = f_i + N\left(\mu = 0,\, \sigma = f_{10\sigma}/10\right)\,, \tag{5.4}$$

where N is a random number from a normal distribution. We then assign errors to the resulting fluxes according to equation 5.3. Finally, the new fluxes and their assigned errors are converted into magnitudes and their respective magnitude errors.

5.1.3 Samples

In this work we estimate the expected cosmological constraints using the galaxy clustering analysis of tomographic bins defined with photo-z (see Sec. 5.3). The

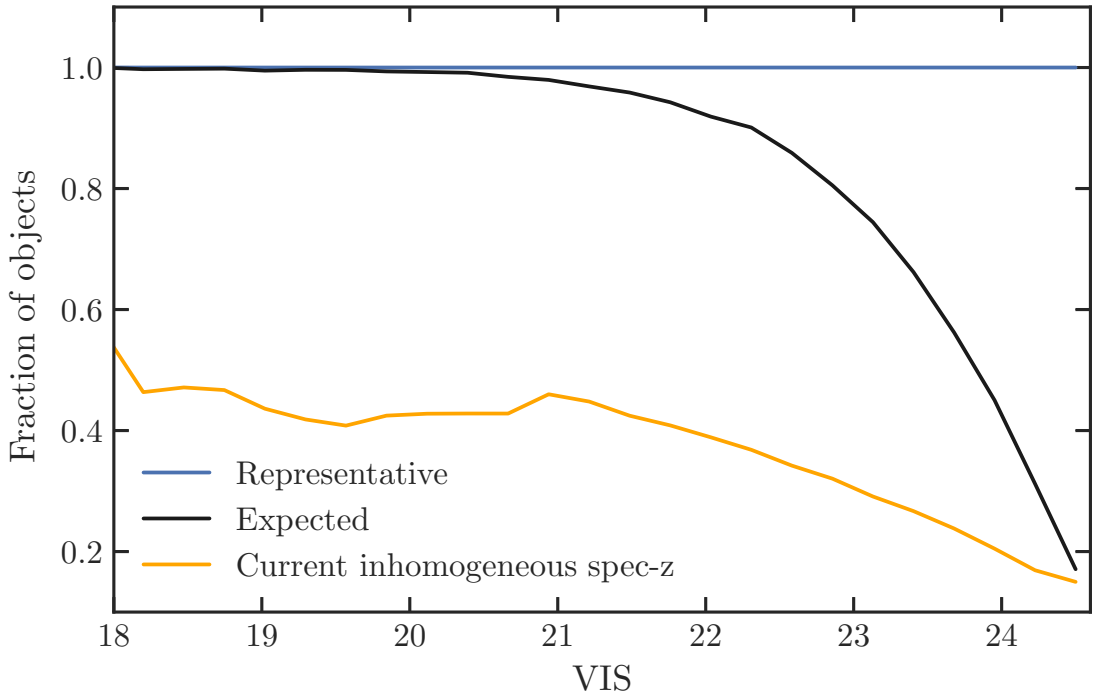

FIGURE 5.2: Fraction of simulated objects with successful spectroscopic redshift as a function of m_{VIS}. The lines represent the completeness fraction of the spectroscopic training samples. The blue line corresponds to the fraction of objects for a random training subsample that is fully representative of the sample under study. In black we show an expectation of the spectroscopic completeness for future ground-based surveys such as Rubin-LSST in m_{VIS} (see Newman et al. 2015). In orange we present the completeness of a training sample with an $n(z)$ similar to the currently available spectroscopic data (see text). Note that the number of objects included in each training set is not represented by the normalization of the different curves in this figure (see Fig. 5.3 top panel for the redshift distributions). Moreover, although our photometric samples go up to $m_{VIS} = 25$, we cut the spectroscopic training samples at $m_{VIS} < 24.5$ because realistic redshifts have not been reliably determined beyond that magnitude limit yet.

magnitude limit of a given sample will give us the galaxies that form the overall sample, while the photo-z algorithm will split that sample into tomographic bins and will provide an estimate of the redshift distributions within these tomographic bins. We can better understand the uncertainties in the method using simulations where we know the true redshift distributions. So far, we have defined four different samples based on the available photometry representing the four cases defined in Table 5.1. The photo-z performance depends on the photometric depth and the spectroscopic data available to train the method. Now, we will generate study cases depending

TABLE 5.2: Cases under study. The photometric limit value corresponds to the column number of Table 5.1 whose magnitude limit depths are used to define each photometric sample. The spectroscopic training sample used to determine the photo-z can be a representative subsample, a sample with a completeness drop in m_{VIS} or a sample with an inhomogeneous spectroscopic redshift distribution as shown in Fig. 5.2.

Sample name	Photometric limit	Spectroscopic training
Case 1: Optimistic	1	Subsample
Case 2: Fiducial	1	Completeness drop
Case 3: Mid-depth	2	Completeness drop
Case 4: Mid-depth Euclid	3	Completeness drop
Case 5: Shallow depth	4	Completeness drop
Case 6: Inhomogeneous spec	4	Inhomogeneous spec-z

on the spectroscopic data available to train the photo-z technique we use. We will use three different spectroscopic samples with different completeness shapes. First, we consider an idealized case where the spectroscopic training sample is a random subsample of the whole sample and thus it is fully representative (blue line in Fig. 5.2). We consider a second case where the spectroscopic sample completeness as a function of magnitude follows the expectations from spectrographs on 8-m class telescopes (Newman et al., 2015). This case is shown in black in Fig. 5.2. It intends to be realistic in the sense that it simulates the spectroscopic incompleteness as a function of magnitude of surveys like zCOSMOS (Lilly et al., 2007), VVDS (Le Fèvre et al., 2013) and DEEP2 (Newman et al., 2013) at least in its shape, although maybe optimistic in its normalization. Finally, we consider a last case where the spectroscopic completeness is similar to the current available spectroscopic surveys, as those listed in Gschwend et al. (2018). We compute how the completeness in spectroscopic data as a function of redshift translates into completeness in m_{VIS} (orange line in Fig. 5.2). These cases are explained in more detail along this section.

We combine the four cases of photometric limits with the three cases of different spectroscopic data available to train the photo-z techniques to generate six galaxy samples for our study. With these six samples we try to encompass a wide range of scenarios to try to understand how the cosmological constraints vary depending on the sample available. We detail these six cases in the following subsections. Table 5.2 summarizes all the cases we consider. All our samples have galaxies down to a magnitude limit of $m_{VIS} = 25$. Note that for our shallower survey (column four

in Table 5.1), galaxies near this m_{VIS} selection limit will have larger errors. It is also important to mention that in all cases we assume the magnitude limit in each band to be isotropic. This will definitely not be the case for Euclid, since ground-based data will consist on a compilation of different surveys pointing at different regions of the sky, with different depths and systematic uncertainties. For instance, Rubin-LSST focuses on the Southern hemisphere, while Euclid will also observe the Northern one. A more detailed analysis taking into account the depth anisotropy of the ground-based data is left for future work.

Case 1: Optimistic

This case uses the deepest magnitude limit and a highly idealized spectroscopic training sample. The sample has magnitudes and errors generated as described in Sec. 5.1.2 with the Euclid and ground-based photometric depth limits shown in the first column of Table 5.1. The photo-z are estimated using a training set that is a complete and representative subsample in both redshift and magnitude of the whole sample.

Case 2: Fiducial

We take this case to be our fiducial sample. We use the deepest photometry as in the optimistic case 1 but the photo-z estimation now makes use of a training sample that has a completeness drop at faint magnitudes that resembles the incompleteness of spectroscopic surveys carried out with spectrographs in 8m-class telescopes. We show the completeness drop in the spectroscopic training sample in Fig. 5.2 (black line). While the completeness as a function of magnitude intends to be realistic of current spectroscopic capabilities, we make the simplifying assumption that this incompleteness does not depend on any galaxy property except its magnitude and therefore we randomly subsample the whole distribution only taking into account the probability of being selected based on the galaxy magnitude.

Case 3: Ground-based mid-depth photometry

We define another sample trained with the same spectroscopic training sample completeness as in the fiducial case but with shallower ground-based magnitude limits in the photometry. The ground-based magnitude limit is a factor of two shallower in signal-to-noise ratio than in cases 1–2. This corresponds to the second column in Table 5.1. This case intends to represent areas on the sky between the celestial

equator and low Northern declinations where Rubin-LSST data at shallower depth may be available.

Case 4: Euclid mid-depth photometry

To explore the possibilities of available photometry, especially the importance of deep near-infrared photometry, we define a case in which both the Euclid and ground-based photometric depth is reduced by 0.75 magnitudes (third column in Table 5.1). The spectroscopic training sample completeness is the same as in cases 2 and 3.

Case 5: Ground-based shallow depth photometry

The complementary ground-based photometry expected to be available in the Northern hemisphere is shallower than the magnitude limits used in our previous cases. We define a sample to roughly represent and cover this option by considering a ground-based flux limit 1.75 magnitudes brighter compared to our optimistic case (fourth column in Table 5.1). To compute the photo-z we use a spectroscopic training set with the same completeness in m_{VIS} as in cases 2, 3, and 4.

Case 6: Inhomogeneous spectroscopic sample

In this last sample, we want to study the case in which the spectroscopic training sample is very heterogeneous and composed by the combination of many surveys targeting galaxies with different selection criteria and with different spectroscopic facilities. We choose a spectroscopic training set that tries to model the $n(z)$ of current available spectroscopic data coming from surveys as those listed in Gschwend et al. (2018). Given that some of these surveys have different color selection cuts and magnitude limit depths, the combined redshift distribution is not homogeneous presenting peaks and troughs, which cause strong biases in the photo-z estimation due to over and under-represented galaxies at different redshift ranges (see e.g. Zhou et al. 2020). We want to remark that we only try to reproduce the $n(z)$ of the overall spectroscopic sample. We do not try to gather this spectroscopic sample applying the same selection criteria of the different surveys used. We consider that this is not necessary for our purposes as we are only interested in the overall trend induced by using an inhomogeneous spectroscopic training sample. We create the spectroscopic training sample by randomly selecting galaxies based on their redshift to reproduce the overall targeted redshift distribution. Given that the Flagship simulation area we

are using (see Sec. 5.1.1) is smaller than the surveys sampling the nearby Universe, our simulated spectroscopic training does not exactly reproduced our overall redshift distribution at low redshifts. The resulting completeness as a function of the m_{VIS} of this spectroscopic redshift sample can be seen in Fig. 5.2 (orange line). The modeled $n(z)$ is shown in the top panel in Fig. 5.3 (orange line). With this case, which intends to represent the currently available data, we can draw a lower bound on the photo-z accuracy that can be expected for Euclid. In this case, we use the same photometric magnitude limits as in case 5.

5.1.4 Photometric redshifts

The cosmological tomographic analysis of a photometric survey divides the whole sample into redshift bins selected with a photo-z technique. In our study, we want to follow as close as possible the methodological steps that one would carry out in real surveys. For that purpose, we compute the photo-zs of all our study cases described in Table 5.2. We use the Directional Neighborhood Fitting (DNF; De Vicente et al. 2016) training-based algorithm to estimate realistic photo-z estimates of our simulated galaxies. The exact choice of the machine learning training set method is not important for our analysis as most methods of this type perform similarly to the precision levels we are interested in (Euclid Collaboration, in preparation).

DNF estimates the photo-z of a galaxy based on its closeness in the observable space (magnitude space in our case) to a set of training galaxies whose redshifts are known. The main feature of DNF is that the metric that defines the distance or closeness between objects is given by a directional neighborhood metric, which is the product of a Euclidean and an angular neighborhood metrics. The algorithm fits a linear adjustment, a hyperplane, to the directional neighborhood of a galaxy to get an estimation of the photo-z. This photo-z estimate is called $z_{\rm mean}$. The residual of the fit is considered as the estimation of the photo-z error. In addition, DNF also produces another photo-z estimate, $z_{\rm mc}$. It is obtained from the nearest neighbor in the DNF metric for each object. Therefore, it can be considered as a one-point sampling of the photo-z probability density distribution. As such, it is not a good individual photo-z estimate of the object, but when all the estimates in a galaxy sample are stacked it can recover the overall probability density distribution of the sample (Rau et al., 2017). When working with tomographic bins, we will classify the galaxies into different bins using their $z_{\rm mean}$ and we will obtain the photometric

distribution, $n(z)$, within each bin by stacking their $z_{\rm mc}$. This is an approach used by DES in analyzing their First Year Data results (e.g., Hoyle et al. 2018, Crocce et al. 2019, Camacho et al. 2019) providing redshift distributions that were validated with other independent assessment methods. Therefore, we define the $n(z)$ by stacking the $z_{\rm mc}$ estimator instead of the true redshift of the simulation to make the photo-z distribution close to what would be obtained in a real data analysis with the assurance that the method has been validated.

We select a patch of sky of 3.35 deg^2 to create the samples to train DNF. These training samples have the magnitudes and errors computed with the same magnitude limits as the sample whose photo-z we want to compute (see Table 5.1). As mentioned before, we generate three types of spectroscopic training samples. For all of them we limit the spectroscopic training sample to galaxies brighter than $\rm m_{VIS}$ = 24.5 as there are few objects whose redshift has been reliably determined beyond that magnitude limit. For our first spectroscopic training sample, we choose a random subsample of the whole distribution up to the $\rm m_{VIS}$ spectroscopic magnitude limit. For the second sample, we apply a selection completeness function similar to the expectation from spectroscopic surveys in 8m-class telescopes. Ignoring any dependency on galaxy properties other than magnitude in the completeness function. For our third spectroscopic sample, we use the redshift distribution described in our case 6 sample.

The true redshift distributions of the spectroscopic training set used to train DNF for each of the sample cases considered in this work are shown in the top panel of Fig. 5.3. In blue, we present the redshift distribution of case 1 with the first spectroscopic training sample that it is fully complete as a function of magnitude. We show in black the resulting $N(z)$ of case 2. Cases 3–5 (olive, red and orange colors in Fig. 5.3) have the same training sample completeness as a function of magnitude. The drop in completeness at faint magnitudes translates into a decrease of objects at high redshift. Last, we present the resulting redshift distribution with the third spectroscopic training set in orange. Gathering multiple selection criteria from different spectroscopic surveys leads to an inhomogeneous redshift distribution for the spectroscopic training sample. In the same Fig. 5.3, we show the overall photo-z distributions of $z_{\rm mean}$ (middle panel) and $z_{\rm mc}$ (bottom panel) values obtained for the full sample for each of the six cases. We see how an inhomogeneous $N(z)$ in the training sample leads to an inhomogeneous distribution of the photo-z.

Fig. 5.4 shows the photo-z obtained with DNF as a function of true redshift for

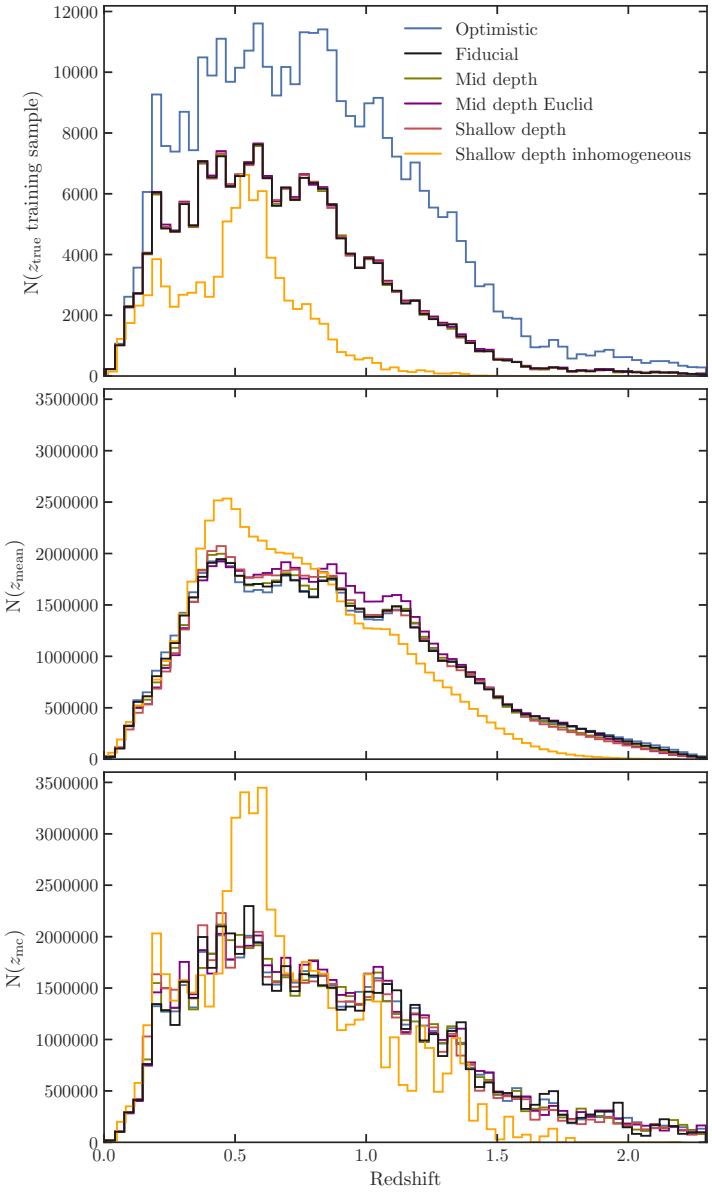

FIGURE 5.3: *Top:* True redshift distributions of the training samples used to run DNF in all 6 cases. The training samples include magnitudes brighter than $m_{VIS} = 24.5$. The true redshift comes from the Flagship simulation. The four training samples with almost identical true redshift distributions have the same completeness drop in m_{VIS} and only differ in the photometric quality. *Middle:* z_{mean} photo-z distributions obtained with DNF for the 6 photometric samples up to $m_{VIS} = 25$. The z_{mean} photo-z estimate returned by DNF is the value resulting from the mean of the nearest neighbors redshifts. *Lower:* Photo-z distributions obtained with DNF for the z_{mc} statistic, which for each galaxy is a one-point sampling of the redshift probability distribution estimated from the nearest neighbor (see text for details).

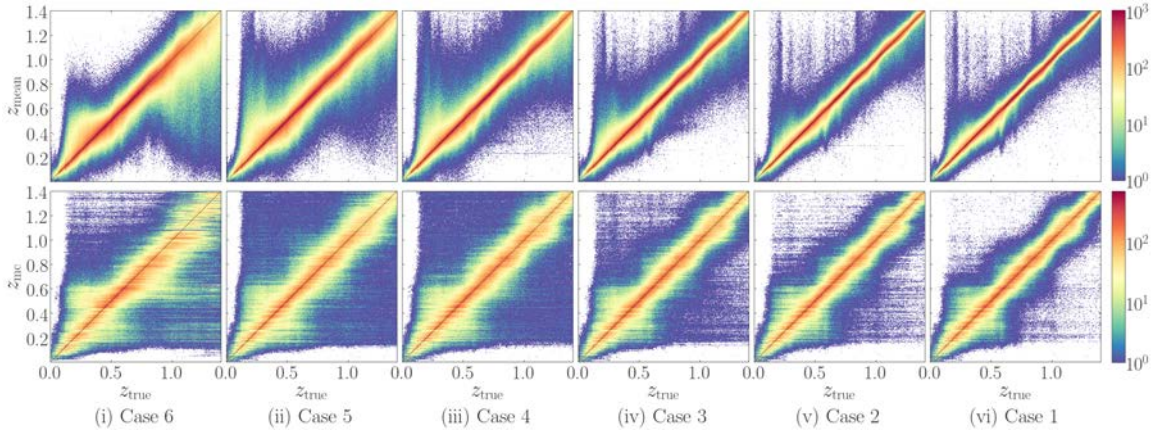

FIGURE 5.4: Scatter plot of both photo-zs given by DNF, $z_{\rm mean}$ (top row) and $z_{\rm mc}$ (bottom row), as a function of true redshift for all the samples considered up to $\mathrm{m_{VIS}} < 24.5$. Left to right: (i) Inhomogeneous spectroscopic sample: Euclid nominal photometry but ground-based photometry 1.75 magnitudes shallower than the optimistic case. The sample is trained using a spectroscopic sample with a $n(z)$ similar to the one obtained with current available spectroscopy; (ii) Ground-based shallow-depth sample: Same photometry as (i) but the training sample has a completeness drop in $\mathrm{m_{VIS}}$; (iii) Euclid mid-depth sample: Both Euclid photometry and complementary surveys are 0.75 magnitudes shallower than the optimistic case; the training sample has a completeness drop in $\mathrm{m_{VIS}}$; (iv) Ground-based mid-depth sample: Ground-based surveys photometry 0.75 magnitudes shallower and trained using a sample with a completeness drop in $\mathrm{m_{VIS}}$; (v) Fiducial sample: Euclid nominal photometry and ground-based surveys photometry optimistic case, trained using a sample with a completeness drop in $\mathrm{m_{VIS}}$; (vi) Optimistic sample: same photometry as in (v) but trained with a random subsample and thus fully representative.

the six samples up to $\mathrm{m_{VIS}} < 24.5$. This figure gives us an indication of how the photo-z scatter decreases with deeper photometry. Photometric samples go up to $\mathrm{m_{VIS}} = 25$. However, we cut the spectroscopic training sample at $\mathrm{m_{VIS}} = 24.5$ to be more realistic. The lack of objects between 24.5 and 25.0 in the training sample forces the algorithm to extrapolate beyond that magnitude, and thus noisier photo-zs are obtained. In Fig. 5.4, we show galaxies only down to $\mathrm{m_{VIS}} < 24.5$ to reduce the noise and make the figure clearer.

5.2 Cosmological model

We optimize the photometric sample of Euclid considering the baseline cosmological model presented in Euclid Collaboration: Blanchard et al. (2019, hereafter EC19): a spatially flat Universe filled with cold dark matter and dark energy. We consider an

extended cosmological model beyond ΛCDM by adding a dark energy fluid parameterized by an equation of state with two parameters, w_0 and w_a, following the CPL (Chevallier, and Polarski, 2001; Linder, 2005) parameterization

$$w(z) = w_0 + w_a \frac{z}{1+z} \,. \tag{5.5}$$

Then, the Hubble parameter as a function of the components of the universe goes as

$$H(z) = H_0 \left[\Omega_{\rm r}(1+z)^4 + \Omega_{\rm m}(1+z)^3 + \Omega_k(1+z)^2 + \Omega_{\rm de} f_{\rm de}(z)\right]^{1/2} \tag{5.6}$$

where the density parameter for the curvature follows $\Omega_k = 1 - (\Omega_{\rm r} + \Omega_{\rm m} + \Omega_{\rm de})$, which for a flat universe is zero since $\Omega_k(z) = -kc^2/\left[a^2(z)H^2(z)\right]$ as seen in equation 1.17, and

$$f_{\rm de}(z) = \begin{cases} 1 & \text{if de} = \Lambda \\ (1+z)^{3(1+\omega_0+\omega_a)\exp\left[-3\omega_a \frac{z}{1+z}\right]} & \text{if de} = \text{dynamic} \end{cases} . \tag{5.7}$$

The cosmological model is fully specified by the dark energy parameters, w_0 and w_a, the total matter and baryon density today, $\Omega_{\rm m}$ and $\Omega_{\rm b}$, the dimensionless Hubble constant, h, the spectral index, $n_{\rm s}$, and the RMS of matter fluctuations on spheres of 8 h^{-1}Mpc radius, σ_8. We assume a dynamically evolving, minimally-coupled scalar field, with sound speed equal to the speed of light and vanishing anisotropic stress as dark energy. Therefore, we neglect any dark energy perturbations in our analysis. We also allow the equation of state of dark energy to cross $w(z) = -1$ using the Hu, and Sawicki (2007) prescription.

A way to relate the cosmological model to the cosmological probes we are interested in is through the matter power spectrum, that for convenience can be written as

$$P_{\rm m}(k,z) = \left(\frac{\sigma_8}{\sigma_{\rm N}}\right)^2 \left[\frac{D(z)}{D(z=0)}\right]^2 T_{\rm m}^2 k^{n_{\rm s}} \tag{5.8}$$

where $T_{\rm m}$ is the scale dependent part of the transfer function (see equation 1.50), $D(z)$ is the growth factor as described in equation 1.60, σ_8 is related to the density fluctuations in spheres of 8 h^{-1} Mpc that correspond to the typical size of clusters

$$\sigma_8^2 = \frac{1}{2\pi^2} \int P_{\rm m}(k, z=0) |\tilde{W}(kR_8)|^2 k^2 {\rm d}k \,, \tag{5.9}$$

and $\sigma_{\rm N}$ is a normalization constant

$$\sigma_{\rm N}^2 = \frac{1}{2\pi^2} \int T_{\rm m}^2(k) |\tilde{W}(kR_8)|^2 k^{n_{\rm s}+2} {\rm d}k \tag{5.10}$$

where $\tilde{W}(x) = 3(\sin(x) - x\cos(x))/x^3$ is the Fourier transform of the filter function.

We will see in detail the exact observables to determine the cosmological parameters from each cosmological probe in Sec. 5.3.

5.2.1 Fiducial values

The fiducial values of the cosmological parameters are given by

$$\begin{aligned} \{\Omega_{\rm m}, \Omega_{\rm b}, w_0, w_a, {\rm h}, n_{\rm s}, \sigma_8\} = \\ = \{0.32, 0.05, -1, 0, 0.67, 0.96, 0.816\} \,. \end{aligned} \tag{5.11}$$

Moreover, we fix the sum of neutrino masses to $\sum m_\nu = 0.06\,{\rm eV}$. Note that the linear growth factor depends on both redshift and scale when neutrinos are massive, but we follow EC19 in neglecting this effect, given the small fiducial value considered. Therefore, we compute the growth factor accounting for massive neutrinos, but neglect any scale dependence. Note that the fiducial values used in this analysis are compatible with the fiducial cosmology of the Flagship simulation presented in Sec. 5.1.1 except for σ_8. This can be explained by the fact that the Flagship simulation does not account for massive neutrinos and therefore considers a slightly larger value for σ_8. However, since we are only extracting the galaxy bias and the galaxy distributions from Flagship, and we are computing Fisher forecasts, this difference in the fiducial σ_8 value does not have any impact on our results.

5.2.2 Fisher matrix formalism

We quantify the performance of photometric galaxy samples in constraining cosmological parameters through the metric figure of merit (FoM), as defined in Albrecht et al. (2006) but with the parameterisation defined in EC19. Our FoM is proportional to the inverse of the area of the error ellipse in the parameter plane of w_0 and w_a defined by the marginalized Fisher submatrix, $\tilde{\mathbf{F}}_{w_0 w_a}$,

$${\rm FoM}_{w_0 w_a} = \sqrt{\det\left(\tilde{\mathbf{F}}_{w_0 w_a}\right)} \,. \tag{5.12}$$

The Fisher matrix estimates the errors for cosmological parameter measurements and describes how fast the likelihood falls around the maximum. The Fisher matrix is defined as the expectation value of the second derivatives of the logarithmic likelihood function

$$F_{\alpha\beta} = \langle -\frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta} \rangle \tag{5.13}$$

where α and β are the subscripts that refer to the cosmological parameters θ_α and θ_β. If the likelihood is Gaussian, the Fisher matrix can be expressed as

$$F_{\alpha\beta} = \frac{1}{2}\mathrm{tr}\left[\frac{\partial \mathrm{C}}{\partial\theta_\alpha}\mathrm{C}^{-1}\frac{\partial \mathrm{C}}{\partial\theta_\beta}\mathrm{C}^{-1}\right] + \sum_{pq}\frac{\partial\mu_p}{\partial\theta_\alpha}(\mathrm{C}^{-1})_{pq}\frac{\partial\mu_q}{\partial\theta_\beta} \tag{5.14}$$

where $\mathrm{C} = \langle(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\rangle$ is the expected covariance of the data, $\boldsymbol{\mu}$ is the mean of the data vector $\mathbf{x}$, and q and p are the variables in the data vector. If the data $\mathbf{x}$ follows a Gaussian distribution with mean $\langle x\rangle = \mu$ and covariance C, either the mean is zero, or the covariance does not depend on the parameters. We assume that the covariance is does not depend on the parameters of the model and thus we will only use the second term of 5.14.

Once the Fisher matrix is determined, the covariance matrix can be constructed

$$C_{\alpha\beta}\left(F^{-1}\right)_{\alpha\beta}\,, \tag{5.15}$$

which gives the errors of the cosmological parameters. For example, marginalizing the uncertainties on each parameter, the diagonal elements of the covariance matrix give the 1σ error

$$\sigma_\alpha = \sqrt{C_{\alpha\alpha}}\,. \tag{5.16}$$

In this work, we will use the FoM using the Fisher matrix formalism as defined in equation 5.12. The higher the FoM value, the higher the cosmological constraining power.

5.3 Building forecasts for Euclid

So far, we have seen how the photometric depth and the spectroscopic training sample determine the overall redshift distributions of the resulting samples. We have selected six cases to cover a range of possible scenarios that we may encounter in the analysis of Euclid data complemented with ground-based surveys. Once the galaxy distributions

for the photometric cases under study have been obtained, we want to propagate the photo-z accuracy in determining tomographic subsamples to the final constraints on the cosmological parameters in order to understand how to optimize the photometric sample for galaxy clustering analyses.

In this work we follow the forecasting prescription presented in EC19. We consider the same Fisher matrix formalism and make use of the `CosmoSIS`[1] code validated for Euclid specifications therein. Our observable is the tomographically binned projected angular power spectrum, $C_{ij}(l)$, where l denotes the angular multipole, and i, j stand for pairs of redshift tomographic bins. This formalism is the same for WL, galaxy clustering (with the photometric sample), and GGL, with the only difference being the kernels used in the projection from the power spectrum of matter perturbations to the spherical harmonic-space observable. In this work we focus on the $\mathrm{GC_{ph}}$ cosmological probe, as well as its combination with GGL. The projection to $C_{ij}(l)$ is performed under the Limber, flat-sky and spatially flat approximations (Kitching et al., 2017; Kilbinger et al., 2017; Taylor et al., 2018). We also ignore redshift-space distortions, magnification, and other relativistic effects (Deshpande et al., 2020). To minimize the impact of neglecting relativistic effects, more relevant at large scales, in our analysis we consider multipole scales from $l \geq 10$ to $l \leq 750$.

In this section we will describe in more detail the different observables and forecast prescriptions for galaxy clustering and weak lensing analyses, as well as their combination.

5.3.1 Weak lensing observables

Although we do not use the weak lensing probe, we do use its combination with galaxy clustering so we take a look at the weak lensing observables. Here we describe the considered observables and forecast procedure for weak lensing as recommended in EC19.

As explained in Sec. 1.6.1, the light path of distant galaxies is deflected by the gravitational potential of foreground galaxies producing distortions in the background galaxies images, a process known as lensing. The distortion can produce changes in size and shape, which are related to the converge and shear, respectively. Since we are in the weak lensing regime, only small changes in ellipticity are considered, which are characterized by the shear. The shear field caused by the matter distribution

[1] `https://bitbucket.org/joezuntz/cosmosis/wiki/Home`

of large-scale structures contains cosmological information about the distribution of matter between the lensing galaxies and the lensed ones. The main observable in weak lensing analyses is the tomographic cosmic shear power spectrum, which is the angular spherical-harmonic measurement of the two-point correlation statistic computed in a series of redshift bins to extract the maximum cosmological information from the three-dimensional shear field.

To correctly model the tomographic cosmic shear power spectrum, the main effects that contribute to the shear signal should be taken into account. Following the model from EC19, the main observational effects are the theoretical cosmic shear power spectrum; the intrinsic alignment power spectrum that quantifies the main astrophysical contamination of cosmic shear, which is caused by the intrinsic alignment (IA) of galaxies; the small-scale part of the matter power spectrum where there are non-linearities effects; the photo-zs and number density that account for the uncertainty in the position of galaxies and the redshift distribution that affects the signal part of the cosmic shear power spectrum; and finally, the shot noise due to the discrete nature of galaxies as tracers of the shear field.

As already mentioned, the observable in weak lensing is the cosmic shear power spectrum, which is the spherical harmonic transform of the two-point correlation function. Its computation is complex (see e.g. Taylor et al. 2018) but thanks to the properties of the spherical Bessel functions under the flat-sky and Limber approximations (see e.g. Kitching et al. 2017), the cosmic shear power spectrum can be expressed as

$$C_{ij}^{\gamma\gamma}(l) \simeq \frac{c}{H_0} \int \frac{W_i^{\gamma}(z) W_j^{\gamma}(z)}{E(z)\chi^2(z)} P_{\delta\delta}\left(\frac{l+1/2}{\chi(z)}, z\right) \mathrm{d}z\,, \tag{5.17}$$

where i and j refers to pairs of redshift bins, $\chi(z)$ is the comoving distance, $P_{\delta\delta}(k,z)$ is the matter power spectrum evaluated at $k = k_l(z) \equiv (l+1/2)/\chi(r)$ due to the Limber approximation, and the weight function $W_i^{\gamma}(z)$ is

$$W_i^{\gamma}(z) = \frac{3H_0}{2c}\Omega_{\mathrm{m},0}(1+z)\tilde{\chi}(z) \int_z^{z_{\max}} n_i(z') \left[1 - \frac{\tilde{\chi}(z)}{\tilde{\chi}(z')}\right] \mathrm{d}z\,, \tag{5.18}$$

where $z_{\max}$ is the maximum redshift of the source redshift distribution, $\tilde{\chi} = \chi(z)/(c/H_0)$ is a dimensionless distance to highlight that $W_i^{\gamma}(z)$ depends on the cosmological h only through the factor (H_0/c), and $n_i(z)$ is the number density distribution of the observed galaxies in the redshift bin i that we obtain from our galaxy samples.

Remember that the observed ellipticity ϵ of a galaxy is given by the sum of the shear γ (defined in Sec. 1.6.1) and the intrinsic ellipticity ϵ^I of the galaxy

$$\epsilon = \gamma + \epsilon^I \,. \tag{5.19}$$

The intrinsically correlated orientation of galaxy shapes, known as IA, affects the two-point shear statistics becoming one of the principal contaminators. So the contributions of IA should be taken into account when modeling the two-point correlation function of equation 5.19:

$$C_{ij}^{\epsilon\epsilon}(l) = C_{ij}^{\gamma\gamma}(l) + C_{ij}^{I\gamma}(l) + C_{ij}^{\gamma I}(l) + C_{ij}^{II}(l) \,. \tag{5.20}$$

The power spectrum that correlates the foreground shear and background ellipticity $C_{ij}^{\gamma I}(l)$ is zero because a foreground shear should not be correlated with a background ellipticity. Then, following a model known as the linear-alignment model, the power spectrum that correlates the background shear and foreground intrinsic alignment $C_{ij}^{I\gamma}(l)$ and the intrinsic-intrinsic alignment autocorrelation power spectrum $C_{ij}^{\gamma I}(l)$ can be written as

$$C_{ij}^{I\gamma}(l) = \frac{c}{H_0} \int \frac{W_i^{\gamma}(z) W_j^{\mathrm{IA}}(z) + W_i^{\mathrm{IA}}(z) W_j^{\gamma}(z)}{E(z)\chi^2} P_{\delta I}\left(\frac{l+1/2}{\chi(z)}, z\right) \mathrm{d}z \,, \tag{5.21}$$

$$C_{ij}^{II}(l) = \frac{c}{H_0} \int \frac{W_i^{\mathrm{IA}}(z) W_j^{\mathrm{IA}}(z)}{E(z)\chi^2} P_{II}\left(\frac{l+1/2}{\chi(z)}, z\right) \mathrm{d}z \,, \tag{5.22}$$

where the weight function for IA is given by

$$W_i^{\mathrm{IA}}(z) = \frac{n_i(z)}{c/H(z)} = \frac{H_0}{c} n_i(z) E(z) \,. \tag{5.23}$$

Notice that a key ingredient for $W_i^{\mathrm{IA}}(z)$ and $W_i^{\gamma}(z)$ (5.18) is the number density distribution $n_i(z)$ of the observed galaxies in the bin i, which is obtained from the distribution of galaxies of our samples whose photo-z has been computed using a realistic method. Thus the photometric precision has a direct impact in the constraining power of the cosmological probe.

The power spectrum of $P_{\delta I}$ and P_{II} describe the effects of IA and can be related to the matter power spectrum as

$$P_{\delta I}(k,z) = -\mathcal{A}_{\rm IA} C_{\rm IA} \Omega_{\rm m,0} \frac{\mathcal{F}_{\rm IA}(z)}{D(z)} P_{\delta\delta}(k,z)\,, \tag{5.24}$$

$$P_{II}(k,z) = \left[-\mathcal{A}_{\rm IA} C_{\rm IA} \Omega_{\rm m,0} \frac{\mathcal{F}_{\rm IA}(z)}{D(z)}\right]^2 P_{\delta\delta}(k,z)\,, \tag{5.25}$$

where the function $\mathcal{F}_{\rm IA}(z)$ is the extended nonlinear alignment (eNLA) model

$$\mathcal{F}_{\rm IA}(z) = (1+z)^{\eta_{\rm IA}} \left[\frac{\langle L\rangle(z)}{L_\star(z)}\right]^{\beta_{\rm IA}}\,. \tag{5.26}$$

With respect to nonlinear alignment (NLA) model, the extended model includes the luminosity and thus redshift dependence of the IA contribution. The term $\langle L\rangle(z)/L_*(z)$ is the redshift-dependent ratio between the average source luminosity and the characteristic scale of the luminosity function (Hirata et al., 2007; Bridle, and King, 2007). For a detailed explanation on IA modeling see Samuroff et al. (2019). In this work we use the same ratio of luminosities for every galaxy sample. However, this ratio should in principle depend on the specific galaxy population. Since we select galaxies according to a $\rm m_{VIS}$ cut and not according to a particular galaxy type, we expect that the luminosity ratio does not change significantly between galaxy samples, and therefore use the same ratio for simplicity. We set the fiducial values for the intrinsic alignments nuisance parameters to

$$\{\mathcal{C}_{\rm IA}, \mathcal{A}_{\rm IA}, \eta_{\rm IA}, \beta_{\rm IA}\} = \{0.0134, 1.72, -0.41, 2.17\}\,, \tag{5.27}$$

in agreement with the recent fit to the IA contribution on the Horizon-AGN simulation (Chisari et al., 2015), although the amplitude $\mathcal{A}_{\rm IA}$ might be smaller in practice (Fortuna et al., 2020).

The last contribution to the observed cosmic shear power spectrum is the one from the shot noise. The contribution is non-zero for correlations of the power spectra with the same bin, but it is zero for correlations between different bins since ellipticities of galaxies at different redshifts should not be correlated. Then the shot noise contribution is expressed as

$$N^{\epsilon}_{ij}(l) = \frac{\sigma^2_\epsilon}{\bar{n}_i} \delta^{\rm Kr}_{ij} \tag{5.28}$$

where $\bar{n}_i$ is the galaxy surface density per bin expressed in inverse steradians, $\delta_{ij}^{\mathrm{Kr}}$ is the Kronecker delta, σ_ϵ^2 is the variance of the observed ellipticities.

Finally, the complete observed cosmic shear tomographic angular power spectrum for a flat-sky and Limber approximation is given by

$$C_{ij}^{\epsilon\epsilon}(l) = C_{ij}^{\gamma\gamma}(l) + C_{ij}^{II}(l) + C_{ij}^{I\gamma}(l) + N_{ij}^{\epsilon}(l)\,, \tag{5.29}$$

with an expanded form of

$$\begin{aligned} C_{ij}^{\epsilon\epsilon}(l) =& \frac{c}{H_0}\int \frac{W_i^{\gamma}(z)W_j^{\gamma}(z)}{E(z)\chi^2(z)} P_{\delta\delta}\left(\frac{l+1/2}{\chi(z)}, z\right) \mathrm{d}z + \\ & \frac{c}{H_0}\int \frac{W_i^{\mathrm{IA}}(z)W_j^{\mathrm{IA}}(z)}{E(z)\chi^2} P_{II}\left(\frac{l+1/2}{\chi(z)}, z\right) \mathrm{d}z + \\ & \frac{c}{H_0}\int \frac{W_i^{\gamma}(z)W_j^{\mathrm{IA}}(z) + W_i^{\mathrm{IA}}(z)W_j^{\gamma}(z)}{E(z)\chi^2} P_{\delta I}\left(\frac{l+1/2}{\chi(z)}, z\right) \mathrm{d}z + N_{ij}^{\epsilon}(l)\,. \end{aligned} \tag{5.30}$$

Lastly, we need to take into account the correction of the power spectrum at small scales due to nonlinear effects. A common approach is to use simulations to define a function of the linear power spectrum to convert it into the nonlinear power spectrum. There are two common recipes to model the nonlinear power spectrum: the halofit (introduced by Smith et al. 2003) and the halomodel (introduced by Cooray, and Sheth 2002, and extended by Mead et al. 2016) recipes. Both recipes compute both the 1-halo and 2-halo terms. The 1-halo term accounts for correlations between dark matter particles within the same dark matter halo and dominates on small scales. The 2-halo term describes correlations between dark matter haloes, dominates on larger scales and is proportional to the linear matter power spectrum multiplied by the bias. Here we follow the same halofit recipe as in EC19. There, the used halofit recipe is a revised version from Takahashi et al. (2012) that computes the 1- and 2-halo terms using fitting functions defined by cosmological parameters including the dark energy equation of state parameters. These functions were modeled using 16 different N-body simulations. The used halofit recipe also includes corrections in the nonlinear regime due to the presence of massive-neutrinos as described in Bird et al. (2012).

Fisher matrix for weak lensing

The Fisher matrix for the cosmic shear tomographic angular power spectrum is given by

$$F^{\rm L}_{\alpha\beta} = \sum_{l=l_{\rm min}}^{l_{\rm max}} \sum_{ij,mn} \frac{\partial C^{\epsilon\epsilon}_{ij}(l)}{\partial\theta_\alpha} \left[(\Delta C^{\epsilon\epsilon}(l))^{-1}\right]_{jm} \frac{\partial C^{\epsilon\epsilon}_{mn}(l)}{\partial\theta_\beta} \left[(\Delta C^{\epsilon\epsilon}(l))^{-1}\right]_{ni} \tag{5.31}$$

where the cross-correlation between modes with different l is not included since the covariance $C^{\epsilon\epsilon}(l)$ is assumed to be Gaussian. The covariance is defined as

$$\Delta C^{\epsilon\epsilon}_{ij}(l) = \sqrt{\frac{2}{(2l+1) f_{\rm sky} \Delta l}} C^{\epsilon\epsilon}_{ij}(l) \tag{5.32}$$

where $f_{\rm sky}$ is the fraction of the surveyed sky, and Δl is the multipole bandwidth.

5.3.2 Photometric galaxy survey observables

In this work we focus on the photometric galaxy clustering probe and its combination with weak lensing. Given that the photo-z measurements of the position of the galaxies carry uncertainties, galaxies are split in tomographic redshift bins to extract the maximum information, following the same procedure as with weak lensing. So the main $\rm GC_{ph}$ observable also is the tomographic angular galaxy clustering power spectrum. Since the photometric samples used for the $\rm GC_{ph}$ analyses are the same as for GGL, we decide that both probes will share the same tomographic bins, with the exact same number density in each redshift bin n_i and the same angular galaxy density $\bar{n}_i$ at each bin

$$\bar{n}_i = \int_{z_{\rm min}}^{z_{\rm max}} n_i(z) {\rm d}z \,. \tag{5.33}$$

Using the same galaxy sample for both probes is a bit of a simplification because in real observation the galaxy samples for both probes may be different due to the inherent intrinsic selection criteria based on photometry and galaxy shapes. Notice that the number density distribution $n_i(z)$ in each redshift bin comes from the photometric galaxy samples defined in the Flagship simulation, whose photo-z have been computed using a training set code and thus the galaxy positions in redshift carry uncertainties.

Following Euclid Collaboration: Blanchard et al. (2019), the radial weight function for $\mathrm{GC_{ph}}$ is given by

$$W_i^{\mathrm{G}}(k,z) = b_i(k,z)\frac{n_i(z)}{\bar{n}_i}H(z)\,, \tag{5.34}$$

where $b_i(k,z)$ is the galaxy bias in the tomographic bin i. We assume a constant bias in each redshift bin and that the bias does not depend on the scale k. We compute the bias using the Flagship simulation and following the relation 5.40. At the end of Sec. 5.3.2, we explain in more detail how we determine the bias.

Using the Limber approximation, the tomographic angular power spectrum for $\mathrm{GC_{ph}}$ is given by

$$C_{ij}^{\mathrm{GG}}(l) = \int \frac{W_i^{\mathrm{G}}(z)W_j^{\mathrm{G}}(z)}{H(z)\chi^2(z)}P_{\delta\delta}\left(\frac{l+1/2}{\chi(z)},z\right)\mathrm{d}z\,. \tag{5.35}$$

Regarding the modeling of the nonlinear scales, the same process as the one described for weak lensing (Sec. 5.3.1) is used for the $\mathrm{GC_{ph}}$ and GGL, which is the use of halofit with corrections for extended cosmological models of Takahashi et al. (2012) and correction for massive neutrinos of Bird et al. (2012).

Fisher matrix for photometric galaxy clustering

The Fisher matrix for the $\mathrm{GC_{ph}}$ tomographic angular power spectrum is given by

$$F_{\alpha\beta}^{\mathrm{G}} = \sum_{l=l_{\mathrm{min}}}^{l_{\mathrm{max}}}\sum_{ij,mn}\frac{\partial C_{ij}^{\mathrm{G}}(l)}{\partial\theta_\alpha}\left[\left(\Delta C^{\mathrm{G}}(l)\right)^{-1}\right]_{jm}\frac{\partial C_{mn}^{\mathrm{G}}(l)}{\partial\theta_\beta}\left[\left(\Delta C^{\mathrm{G}}(l)\right)^{-1}\right]_{ni} \tag{5.36}$$

where the $\mathrm{GC_{ph}}$ covariance $C^{\mathrm{G}}(l)$ is defined as

$$\Delta C_{ij}^{\mathrm{G}}(l) = \sqrt{\frac{2}{(2l+1)f_{\mathrm{sky}}\Delta l}}\left[C_{ij}^{\mathrm{G}}(l)+N_{ij}^{\mathrm{G}}(l)\right]\,, \tag{5.37}$$

where Δl is the width of the multipoles bins used when computing the angular power spectra, and $N_{ij}^{\mathrm{G}}(l)$ is the shot noise term for $\mathrm{GC_{ph}}$ defined as

$$N_{ij}^{\mathrm{G}}(l) = \frac{\delta_{ij}^{\mathrm{Kr}}}{\bar{n}_i}\,. \tag{5.38}$$

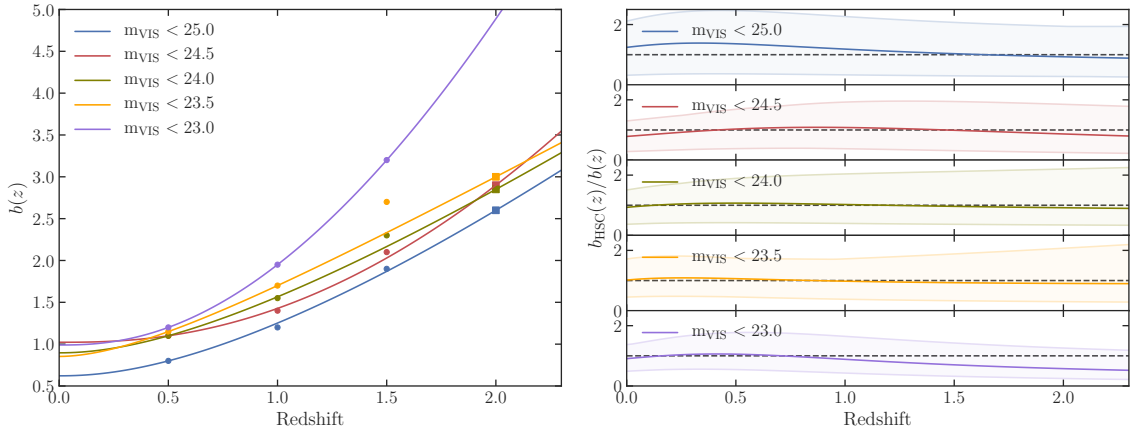

FIGURE 5.5: *Left panel:* Galaxy bias as a function of redshift. Dots correspond to the measured values in the Flagship simulation for different magnitude cuts and the solid lines are a fit following equation 5.40. We plot with squares the bias values obtained for $z = 2$ to indicate that at that redshift there are few objects and thus the values are slightly less reliable. At $m_{VIS} < 23$ there were not enough objects at $z = 2$ to compute the bias in Flagship. *Right panel:* Ratio between the HSC bias, b_{HSC}, from N20 and the Flagship bias for each magnitude-limited sample. To asses the 1σ uncertainty of b_{HSC} along the redshift range, we generate a set of Gaussian random numbers for the free parameter α, b_1, and b_0 of b_{HSC} with their values as mean and their errors as standard deviation. Then we evaluate b_{HSC} in the redshift range for all the set of free parameters previously generated. We pick the maximum and minimum b_{HSC} at each redshift. This corresponds to the shaded regions.

Galaxy bias for GC_{ph} and GGL

When considering GC_{ph} and GGL one of the primary sources of uncertainty is the relation between the galaxy distribution and the underlying total matter distribution, i.e. the galaxy bias (Kaiser, 1987). In this work we consider a linear galaxy bias relating the galaxy density fluctuation to the matter density fluctuation with a simple linear relation

$$\delta_g(\vec{x}, z) = b(z)\delta_m(\vec{x}, z)\,, \tag{5.39}$$

where we neglect any possible scale dependence. Note that a linear bias approximation is sufficiently accurate for large scales (Abbott et al., 2018a). However, when adding very small scales into the analysis, a more detailed modeling of the galaxy bias is required (see e.g., Sánchez et al., 2017). One of the approaches to this modeling is through perturbation theory, which introduces a nonlinear and nonlocal galaxy bias (Desjacques et al., 2018).

In this work we consider a constant galaxy bias in each tomographic bin. We get their fiducial values by fitting the directly measured bias in Flagship to the function

$$b(z) = \frac{Az^B}{1+z} + C\,, \tag{5.40}$$

where A, B and C are nuisance parameters. We select five subsamples with $\mathrm{m_{VIS}}$ limiting magnitudes: 25, 24.5, 24, 23.5, and 23 from the Flagship galaxy sample. We compute the bias values as a function of redshift for each of these magnitude-limited subsamples using directly the true redshift of Flagship at redshifts 0.5, 1, 1.5 and 2. The obtained bias and fitted functions are shown in the left panel of Fig. 5.5. To fit the bias-redshift relation we choose to use all the galaxy bias values computed with the Flagship simulation, although values at $z = 2$ were less reliable. The value of the bias at $z = 1.5$ falls outside the bias-redshift fit for the $\mathrm{m_{VIS}} < 23$ sample. However, we recomputed the bias fit neglecting the value at $z = 2$ and including the value at $z = 1.5$, with no significant difference. Therefore we keep the bias computed using the fits shown in Fig. 5.5..

To validate the bias obtained with Flagship, we compare our bias values to the ones obtained from the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP) data release 1 (DR1) by Nicola et al. (2020, N20 hereafter). The HSC survey has comparable survey depth and uses similar ground-based bands to the ones considered in this work. N20 fit galaxy bias on magnitude-limited galaxy samples down to $i < 24.5$. We compare their values to ours in the right panel of Fig. 5.5. We extrapolate their bias down to $i < 25$ for our last magnitude-limited bin. Strictly speaking, we are comparing i-band magnitude-selected samples from N20 to our $\mathrm{m_{VIS}}$-band magnitude-selected samples. We have checked in Flagship that the bias values for both i-band and $\mathrm{m_{VIS}}$-band selected samples cut at the same magnitude limit do not change by more than 10% and therefore our comparison is meaningful. N20 assume that bias can be split into two separated terms of redshift and limiting magnitude, and define it as

$$b_{\mathrm{HSC}}(z, m_{\mathrm{lim}}) = \bar{b}(m_{\mathrm{lim}}) D^{\alpha}(z)\,, \tag{5.41}$$

where α is a variable that takes into account the inverse relation between the growth factor and galaxy bias. By fitting α and $\bar{b}(m_{\mathrm{lim}})$ in a multi-step weighted process

they find

$$\begin{aligned} \alpha &= -1.30 \pm 0.19\,, \\ \bar{b}(m_{\rm lim}) &= b_1(m_{\rm lim} - 24) + b_0\,, \end{aligned} \tag{5.42}$$

where $b_1 = -0.0624 \pm 0.0070$ and $b_0 = 0.8346 \pm 0.161$. For a detailed explanation see Sect. 4.6 in N20. We compute $D(z)$ for our sample and use our $\mathrm{m_{VIS}}$ magnitude cuts as $m_{\rm lim}$ along with their fitted parameters to get a bias to compare. The ratio between the HSC bias, $b_{\rm HSC}$, and ours, $b(z)$, is shown in the right panel of Fig. 5.5. Note that N20 compute their bias up to redshift 1.25 and that we have extrapolated their behavior to higher redshifts for the comparison at $z > 1.25$. The values of the bias in Flagship stay within 1σ of the HSC values, $b_{\rm HSC}$ (shaded area in the right panel of Fig. 5.5), confirming that the bias values we use are consistent with the HSC observations.

5.3.3 Galaxy-galaxy lensing observables

Besides $\mathrm{GC_{ph}}$, in this work we focus on GGL, which is a cross-correlation between $\mathrm{GC_{ph}}$ and WL. Since the observables of both WL and $\mathrm{GC_{ph}}$ are quantified and modeled as a two-dimensional projection in different redshift bins, modeling the cross-correlation is simple and thus the same angular power spectrum formalism can be used for GGL. The angular power spectrum of GGL contains contributions from galaxy clustering and cosmic shear, but also from intrinsic galaxy alignments. We assume the latter is caused by a change in galaxy ellipticity that is linear in the density field. Note that such modeling is appropriate for large scales (Troxel et al., 2018), like the ones considered in this analysis, but more complex models should be used for the very small scales (see e.g. Blazek et al., 2019; Fortuna et al., 2020). Within this linear assumption we can define the density-intrinsic and intrinsic-intrinsic three-dimensional power spectra, $P_{\delta I}$ and P_{II}, respectively, that are related to the density power spectrum $P_{\delta\delta}$ following the relations presented in equations 5.24 and 5.25. Combining these equations with the observed tomographic shear angular power spectrum (5.30), the cosmic shear angular power spectrum can be defined as

$$C_{ij}^{\rm LL}(l) = \int_{z_{\rm min}}^{z_{\rm max}} \frac{W_i^{\rm L}(z) W_j^{\rm L}(z)}{H(z)\chi^2(z)} P_{\delta\delta}\left(\frac{l+1/2}{\chi(z)}, z\right) \mathrm{d}z\,, \tag{5.43}$$

where the weight function $W_i^{\rm L}(z)$ contains the radial weight function for cosmic shear $W_i^{\gamma}(z)$ defined in equation 5.18 and includes the effect of intrinsic alignment described in Sec. 5.3.1. It has the form

$$W_i^{\rm L}(z) = W_i^{\gamma}(z) - \mathcal{A}_{\rm IA} C_{\rm IA} \Omega_{\rm m} \frac{\mathcal{F}_{\rm IA}(z)}{D(z)} W_i^{\rm IA}(z)\,, \tag{5.44}$$

where the weight function of the intrinsic alignments $W_i^{\rm IA}(z)$ is defined in equation 5.23. The change of notation of the cosmic shear angular power spectrum is to make clearer its relation with the $\rm GC_{ph}$ power spectrum. Using the same Limber approximation and considering equations 5.43 and 5.35, the observable shear-galaxy angular power spectrum is given by

$$C_{ij}^{\rm GL}(l) = \int \frac{W_i^{\rm G}(z) W_j^{\rm L}(z)}{H(z)\chi^2(z)} P_{\delta\delta}\left(\frac{l+1/2}{\chi(z)}, z\right) {\rm d}z\,. \tag{5.45}$$

We consider the same redshift distributions $n_i(z)$ for both $\rm GC_{ph}$ and GGL and thus the same angular galaxy density $\bar{n}_i$ (5.33). In practice, this is an over-simplification, since these two probes will probably apply different selection criteria when determining their samples. GGL for instance will give some importance to the shape measurements of the galaxies. But for the present Fisher matrix analysis we limit ourselves to use the same sample for both probes. Since we use the same redshift distributions for both probes, we also consider the same linear galaxy bias obtained with Flagship for $\rm GC_{ph}$ as described in Sec. 5.3.2, which we consider constant in each tomographic redshift bin and evolves with redshift following the relation 5.40. In addition, the process to model the nonlinear scales is the same as the one described for weak lensing (Sec. 5.3.1) and used for the $\rm GC_{ph}$.

In this work we will consider all multipoles from $l \geq 10$ up to $l \leq 750$, which corresponds to the more conservative scenario in EC19.

Fisher matrix for the combination of probes

In this work we combine the $\rm GC_{ph}$ and GGL cosmological probes. The expression of the Fisher matrix for the combination of the angular power spectra with contributions from $\rm GC_{ph}$ and GGL is given by

$$F_{\alpha\beta}^{\rm XC} = \sum_{l=l_{\rm min}}^{l_{\rm max}} \sum_{ABCD} \sum_{ij,mn} \frac{\partial C_{ij}^{AB}(l)}{\partial\theta_\alpha} \left[(\Delta C(l))^{-1}\right]_{jm}^{AB} \frac{\partial C_{mn}^{CD}(l)}{\partial\theta_\beta} \left[(\Delta C(l))^{-1}\right]_{ni}^{CD}\,, \tag{5.46}$$

where the block descriptors AB and CD run over the combined probes GL and GG, thus including the observables of the cross-correlation between galaxy clustering and cosmic shear (5.45) and galaxy clustering auto-correlation (5.35). The subscripts ij and nm run over all unique pairs of tomographic bins for $C^{\mathrm{GG}}(l)$, and over all pairs for $C^{\mathrm{GL}}(l)$. The joint covariance matrix is defined as

$$\Delta C_{ij}^{AB}(l) = \sqrt{\frac{2}{(2l+1)f_{\mathrm{sky}}\Delta l}} \left[C_{ij}^{AB}(l) + N_{ij}^{AB}(l)\right] , \tag{5.47}$$

where Δl is the width of the multipoles bins used when computing the angular power spectra, and N_{ij}^{AB} is the shot noise term that for the galaxy clustering auto-correlation takes the form from 5.38 and for GGL is $N_{ij}^{\mathrm{GL}}(l) = 0$ since we assume that the Poisson errors of the cosmic shear and the photometric galaxy clustering are uncorrelated.

5.3.4 Brief overview of CosmoSIS

The CosmoSIS code is used to perform the analysis of the cosmological probes. Here we briefly overview its pipeline. First, the linear matter power spectrum is computed using the Boltzmann solver from the CAMB[2] module. Then the code rescales the linear matter power spectrum to match the normalization of σ_8. The nonlinear corrections of the matter power spectrum are computed using the halofit models with corrections to extended cosmological models of Takahashi and corrections for massive neutrinos of Bird. The corrections are added to the linear matter power spectrum. Then the effects of the intrinsic alignments are added using the model described in Sec. 5.3.1. The galaxy power spectrum is determined using the number density of each tomographic bin from the galaxy samples of Flagship and the galaxy bias computed from the galaxy samples, which follows the model described in Sec. 5.3.2. Then the power spectrum is projected to obtain the angular power spectrum. The intrinsic alignment angular power spectrum is computed and added to the cross-correlation angular power spectrum. Then using the angular power spectrum, the code computes the covariance matrix. Finally, the Fisher matrix is computed iteratively for variations of the parameters until it converges, giving the constrains for the cosmological parameters.

[2]See `camb.info`

5.4 Results

In this section we carry out a series of tests to optimize the sample selection for GC_{ph} analyses. We want to determine the best number and type of tomographic bins to constrain cosmological parameters. We explore the influence of the accuracy in the photo-z estimation and sample size in providing cosmological constraints. We split the data in tomographic redshift bins instead of subsamples in order to have more control in the variations of sample size and photo-z accuracy to better understand their impact in constraining cosmological parameters. We use the FoM defined in equation 5.12 to quantify the constraining power on the cosmological parameters. In addition, we also compute the FoM when combining GC_{ph} with GGL, assuming the same photo-z sample, which implies the same photo-z binning and number density. When computing the cosmological constraining power for GC_{ph} + GGL, we marginalize over the galaxy bias of each tomographic bin and intrinsic alignment parameters, whereas for GC_{ph} alone the galaxy bias parameters are fixed to their fiducial values. The main reason for this choice is that, under the linear galaxy bias approximation, there is a large degeneracy between the galaxy bias and σ_8. In this case, the Gaussianity assumption of the Fisher matrix approach breaks down and its constraints on the cosmological parameters are not reliable. Therefore, we fix the galaxy bias to break this degeneracy when considering GC_{ph} alone. Note that when we combine GC_{ph} with GGL, the additional information brought by the latter is enough to break such degeneracy and constrain σ_8 and the galaxy bias at the same time.

5.4.1 Optimizing the type and number of tomographic bins

We bin galaxies into different numbers of redshift bins to study the impact of the number of redshift bins on the cosmological parameter inference. When we define redshifts bins, we choose galaxies within the redshift range $[0, 2]$ since the maximum lightcone outputs generated in Flagship are at $z = 2.3$ and we prefer to avoid working at the limit of the simulation. We check the effect of using bins with the same redshift width (equidistant) and bins with the same number of objects (equipopulated). We also see the difference when using only GC_{ph} or both GC_{ph} and GGL probes. This analysis is performed using our fiducial sample (case 2) up to $m_{VIS} < 24.5$. We compute the FoM for all the cases mentioned and show the results in Fig. 5.6. The FoM are normalized to ten bins since it is the default number used to compute the forecasts in EC19.

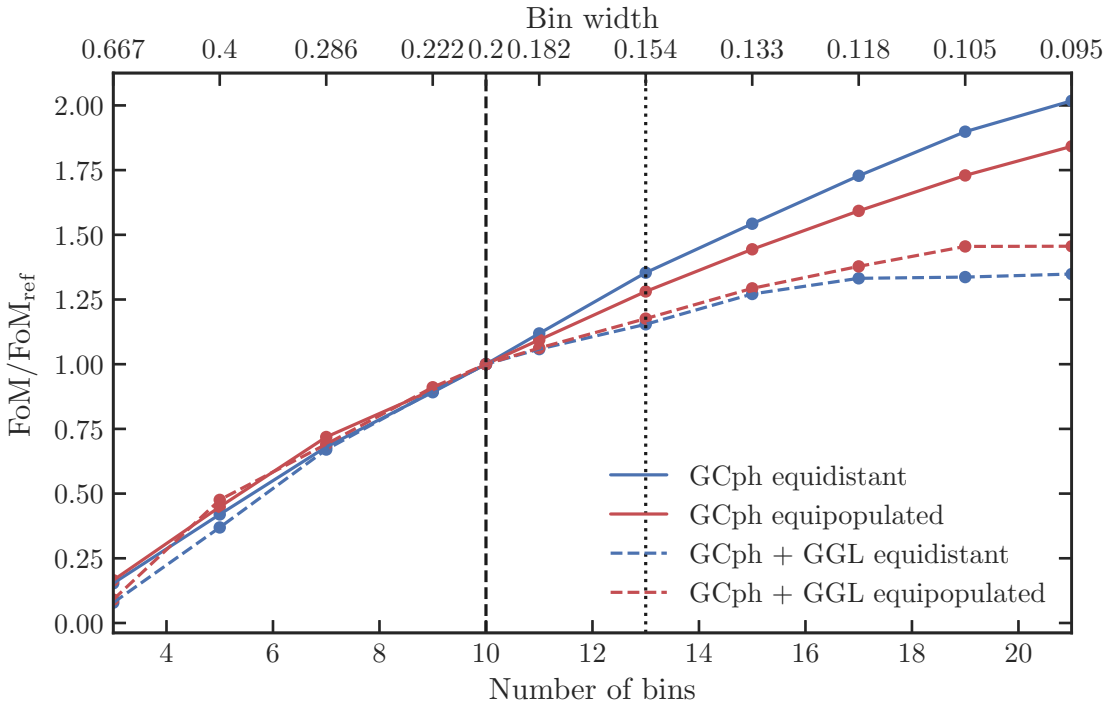

FIGURE 5.6: FoM as a function of the number of bins for GC_{ph} (solid) and GC_{ph} + GGL (dashed) and for bins with the same reshift width (blue) and with the same number of objects (red). The redshift width of the bins when they are equidistant is shown in the top x axis. The FoMs are normalized to the FoM at 10 bins, FoM_{ref}, which corresponds to the specifications for the number of bins used to compute the forecasts in EC19 and denoted by a vertical-dashed line. A vertical-dotted line shows the 13 bins used as our fiducial choice.

As seen in Fig. 5.6, the general tendency of the FoM is to consistently increase with the number of bins until it slowly saturates for a large number. EC19 used ten tomographic bins as their fiducial value. We observe that the FoM continues to increase for larger numbers of bins. For equidistant bins, the FoM increase when moving from ten to thirteen bins is 35.4% and 15.4%, for GC_{ph} only and for GC_{ph} + GGL, respectively. The FoM improvement we get from going to even more bins does not compensate the increase in computational time needed for the analysis. This is especially true when using both probes, where we notice that the curve flattens sooner. Moreover, our photo-z treatment may start to be too simplistic to realistically deal with too many photo-z bins. The FoM saturates with the increasing number of bins because it is not possible to extract more information on radial clustering when the width of the bins is smaller than the photo-z precision. At this limit, the uncertainty at which bin a particular galaxy belongs is greatly increased. For GC_{ph}

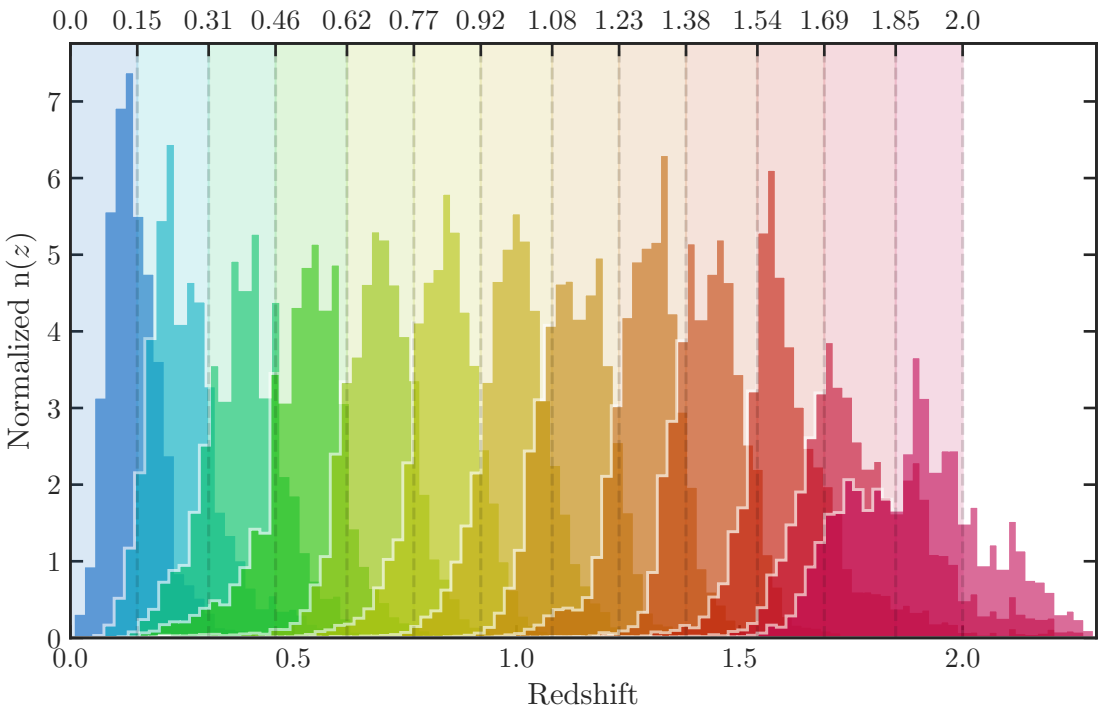

FIGURE 5.7: Redshift distributions ($z_{\rm mc}$) of the thirteen equidistant bins for the fiducial sample. The values in the top x-axis correspond to the limits of the redshift bins. The shaded regions indicate these limits.

+ GGL the curves flatten at lower number of bins since systematic effects in the marginalization of galaxy bias and intrinsic alignment free parameters also affect the cosmological information that can be extracted. Therefore, we choose thirteen to be our fiducial number of bins as a conservative choice.

In addition, we choose equidistant bins as the optimal way of partitioning the sample since we observe that, overall, for $\rm GC_{ph}$ the FoM is larger in this case than in the equipopulated one. For thirteen equidistant bins the FoM is 713 while it is 547 for equipopulated bins (recall that galaxy bias is fixed when considering $\rm GC_{ph}$ alone, which provides these large absolute values for the FoMs), which is an increase of 30%. For the $\rm GC_{ph}$ + GGL combined analysis, the FoM does not appreciably change between the equidistant and equipopulated cases. At thirteen bins, which is the fiducial choice, the FoM difference of using equidistant or equipopulated bins is negligible.

We will use these bin choices to analyze the dependency of cosmological constraints on the photo-z quality and size of the sample. In Fig. 5.7 we show the redshift

distributions of the thirteen equidistant bins for our fiducial case 2 sample.

5.4.2 FoM dependency on photometric redshift quality and number density

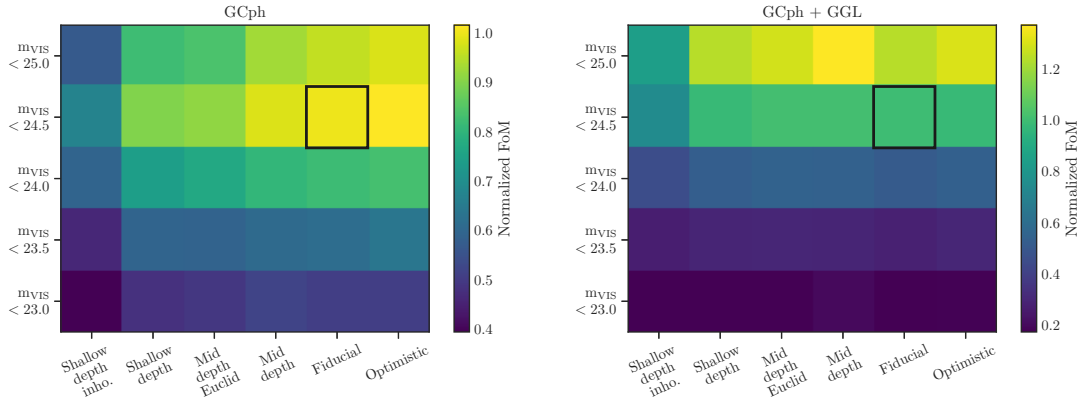

FIGURE 5.8: FoM for the samples defined in Sec. 5.1.3 with different photo-z accuracy and sample size. The size has been reduced by performing a series of cuts in m_{VIS}. The results are normalized to the FoM of the fiducial sample with $m_{VIS} < 24.5$ (highlighted cell). The figures correspond to the results for using only GC_{ph} (left) and for combining it with GGL (right).

Another aspect we want to study is the effect of the trade-off between photo-z accuracy and number density on the constraining power of cosmological parameters. For that purpose, we take the six photometric samples defined in Sec. 5.1.3 and apply five cuts (25, 24.5, 24, 23.5, 23) in m_{VIS} to modify the sample size (leading to a number density of about 41, 29, 18, 12 and 9 galaxies per arcmin^2 respectively). Besides reducing the number density of the photometric samples, the cut in m_{VIS} also affects the photo-z distribution and accuracy of the overall sample. A bright magnitude cut, that eliminates the fainter galaxies, mostly removes galaxies with higher and thus less reliable redshifts. We compute the FoM for all the cases mentioned before and normalize them to the FoM of our fiducial (case 2) sample at $m_{VIS} < 24.5$, for both GC_{ph} only and GC_{ph} + GGL. To help visualize the results, we present the resulting FoM in a grid format in Fig. 5.8 and the values themselves in Table 5.3. The configuration of tomographic bins used to perform the analysis is the optimum one found in the previous section, which is thirteen equidistant bins.

Let us first discuss the case of GC_{ph} alone. As seen in Fig. 5.8, in general, the FoM for GC_{ph} increases with deeper photometric data, which improves the photo-z

TABLE 5.3: Values of the FoM for samples defined in Sec. 5.1.3 with different photo-z accuracy and sample size (same cases as in Fig. 5.8). The results are normalized to the FoM of the fiducial sample with $m_{VIS} < 24.5$. For reference, the unnormalized value of our fiducial sample is 713 for GC_{ph} and 411 for GC_{ph} + GGL. Note that galaxy bias and intrinsic alignments nuisance parameters are free in the latter, which provides a lower FoM than in GC_{ph} alone.

m_{VIS}	Shallow depth inho.	Shallow depth	Mid depth Euclid	Mid depth	Fiducial	Optimistic
GC_{ph}						
25	0.57	0.82	0.84	0.93	0.96	0.98
24.5	0.67	0.90	0.91	0.98	1.00	1.02
24	0.59	0.74	0.77	0.81	0.82	0.83
23.5	0.46	0.59	0.59	0.61	0.62	0.64
23	0.39	0.48	0.50	0.52	0.51	0.51
GC_{ph} and GGL						
25	0.85	1.24	1.29	1.37	1.24	1.30
24.5	0.75	0.98	1.01	1.01	1.00	0.98
24	0.46	0.53	0.55	0.54	0.52	0.54
23.5	0.27	0.30	0.30	0.29	0.28	0.30
23	0.17	0.17	0.18	0.20	0.17	0.18

performance (increasing along the x-axis in the figure). The FoM also increases with number density, determined by the magnitude limit imposed (increasing along the y-axis). We notice a larger increase in the FoM with sample size in those samples where the photo-z quality is better (e.g., the optimistic, fiducial and mid depth ground-based photometry cases). In these cases, increasing the sample size from a m_{VIS} cut from 23.5 to 24 and from 24 to 24.5 leads to an increase of the FoM of about 20%. Clearly, having a fainter magnitude cut results in larger samples that yield higher FoM values. This trend is in agreement with the results presented in Tanoglidis et al. (2020).

The trend of increasing FoM as we take fainter magnitude limit cuts and increase the number density continues as long as the photo-z performance is not degraded. Once we push to faint magnitudes where there are no objects to train the photo-z algorithms, their performance degrades and the photo-z bins start to be wider. There are many object that do not belong to the bins and spurious cross-correlations between different bins appear. As a result, the strength of the cosmological signal is diminished and the FoM decreases. This effect can be seen in Fig. 5.8 for the GC_{ph}

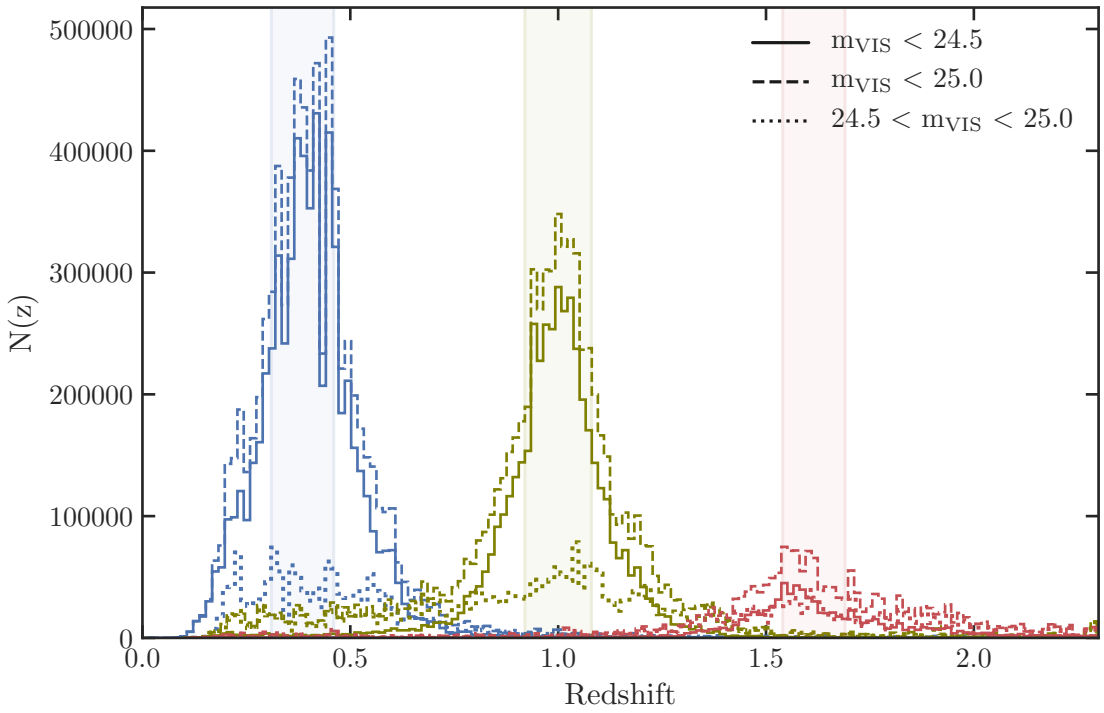

FIGURE 5.9: Redshift distributions (z_{mc}) of three of the thirteen tomographic bins selected with z_{mean} corresponding to the shaded regions $[0.31, 0.46)$ (blue), $[0.92, 1.08)$ (green) and $[1.54, 1.69)$ (red). For each bin we plot objects with $m_{VIS} < 24.5$ (solid), $m_{VIS} < 25.0$ (dashed) and $24.5 < m_{VIS} < 25.0$ (dotted). The photometric sample used is the mid depth Euclid (case 4).

case (left panel), where we can appreciate a reduction in the FoM when we move from a magnitude-limited sample cut at $m_{VIS} < 24.5$ (second row from the top) to a magnitude-limited sample cut at $m_{VIS} < 25.0$ (top row). With this change, we are increasing the sample, but with galaxies that cannot be located in redshift as their photo-z cannot be calibrated. As a result, the clustering strength is diluted and some spurious cross-correlation signal appears resulting in a decreased FoM compared to a shallower sample with better photo-zs.

To illustrate this effect, in Fig. 5.9 we show the redshift distribution of three tomographic bins for three samples with galaxies down to $m_{VIS} < 24.5$, < 25, and with galaxies only between 24.5 and 25. Galaxies with m_{VIS} between 24.5 and 25 are mostly outside their tomographic bin increasing the width of the distribution and diluting the signal. We conclude that the GC_{ph} probe is sensitive to the actual location of their tracer galaxies inside their tomographic bins. Both the photo-z performance and

the number density are important contributing factors when performing cosmological inference with GC_{ph}. When pushing to faint magnitudes, there is no improvement including galaxies that cannot be located in redshift.

Let us now discuss the case where we add GGL to GC_{ph} (right panel in Fig. 5.8). We observe that increasing the sample size (moving along the y-axis) has a more significant impact on the improvements of the FoM than the photo-z quality (changes along the x-axis). The greatest improvement, of about 50% for the best photo-z quality samples, takes place going from $m_{VIS} < 24$ to 24.5. The second largest improvement is of about 25 – 30% when adding objects from $m_{VIS} < 24.5$ to 25. In the GGL case, source galaxies outside the tomographic bin of the lens galaxy contribute to the signal. The lensing kernel is quite extended in redshift and galaxies beyond the lens contribute to the signal with only a mild dependence of their precise redshift, making the photo-z performance less important compared to the GC_{ph} only case. On the other hand, the statistical nature of detecting the lensing signal makes the number density (and therefore the magnitude limit cut) a more important factor in determining the GGL cosmological inference power.

In the FoM grid, we find a non-intuitive behavior for some samples when combining GC_{ph} and GGL (Fig. 5.8 right panel). If we compare the mid depth and mid depth Euclid samples to the fiducial and optimistic samples at the same number density (along the x-axis), we find that the former pair gives better FoM constraints despite having larger photo-z scatter. This is counter-intuitive as fewer galaxies are properly located in redshift and still the FoM cosmological constraints are slightly better. As we mentioned before, whenever the photo-z performance degrades, more galaxies supposedly being in our tomographic bin belong to other bins. This effect can increase the effective number of sources for our lenses and thus boost the GGL signal. However, this is at the expense of reducing the cosmological constraining power of the GC_{ph} probe. The interplay between these two effects is difficult to gauge. The GGL increase appears slightly more prominent when pushing to fainter magnitude limits that produce a sizeable increase in number density.

The representativeness of the training sample also determines the photo-z performance and thus the cosmological constraining power. For GC_{ph}, if we check the difference in FoM between our fiducial sample, trained with a spectroscopic sample that has a completeness drop at faint m_{VIS}, and the same photometric sample trained with a fully representative training sample (optimistic sample) we see a gain of about 1–2% in the FoM. Note that the spectroscopic incompleteness in this case is small and

only affecting faint magnitudes, so the effect on the FoM is also small. This difference greatly increases when we compare the FoM performance of shallower samples and higher incompleteness in the spectroscopic training sample. If we compare the shallow depth sample that was trained with a sample that has a completeness drop in faint m_{VIS} magnitude to the shallow depth inhomogeneous sample that was trained with a sample that is incomplete in the spectroscopic $n(z)$, the difference between of FoM can be up to 25% for GC_{ph} and 39% for both probes combined.

Finally, we look at the difference due to the ground-based photometric depth. The difference between our fiducial and shallow depth cases may represent the change in depth to be achieved in the Southern and Northern hemispheres. For these cases the difference in cosmological constraint power is about 10% at $m_{VIS} < 24.5$ for GC_{ph}. This percentage reduces to 2% if we also consider GGL.

5.4.3 Impact on the cosmological parameters constrains

We further investigate the forecasts of the constraints on the cosmological parameters by looking at the parameters uncertainties, $\sigma_i = ((\mathbf{F}^{-1})_{ii})^{\frac{1}{2}}$, given by the square root of the diagonal elements of the inverse of the Fisher matrix. The uncertainties are computed for all the photometric samples defined in Sec. 5.1.3 and for the different sample sizes. For visual clarity, we present the results in grid form in Figs. 5.10–5.11.

In Fig. 5.10 we show the uncertainties for the GC_{ph} probe. We can appreciate that, in general, the uncertainties have a similar behavior to the FoM, where the sample down to $m_{VIS} < 24.5$ gives the higher FoM. However, there are parameters, such as Ω_b, w_a, and h, that do not degrade as much their performance when going to the deeper $m_{VIS} < 25$ sample.

In Fig. 5.11 we show the uncertainties of the cosmological parameters when we combine the GC_{ph} and GGL cosmological probes. Again, we see similar trends compared to the FoM case, but with minor changes in the behavior of how the uncertainties in some of the parameters vary. The addition of galaxies, increasing the survey depth, and the improvement of the photo-z performance produce lower uncertainties in the Ω_b and h parameters. The reduction of the uncertainty obtained when considering the deepest $m_{VIS} < 25$ case compared to the $m_{VIS} < 24.5$ is minimal, though.

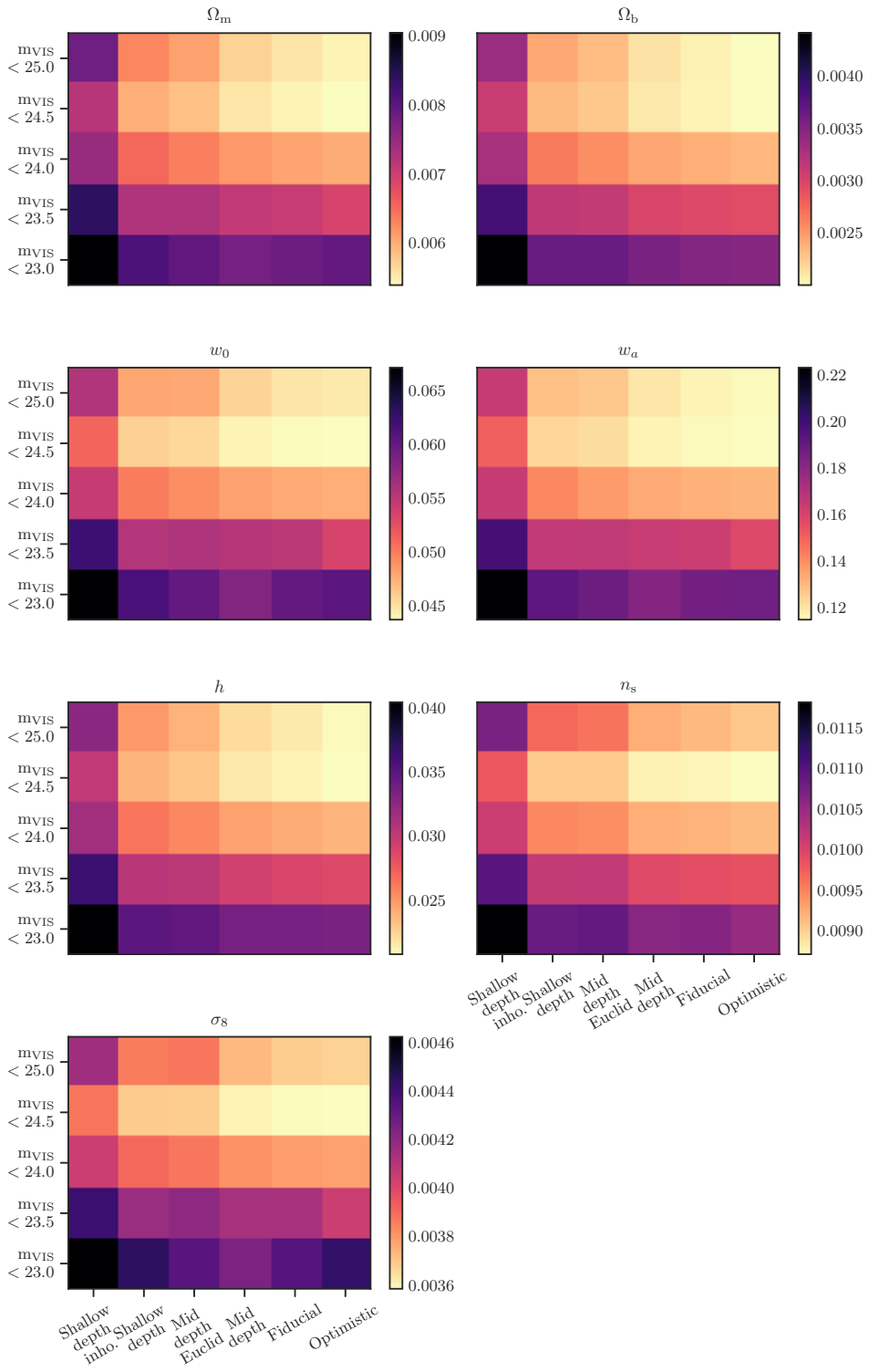

FIGURE 5.10: Uncertainties of the cosmological parameters for all the cases considered in Sec. 5.4.2 for GC_{ph}.

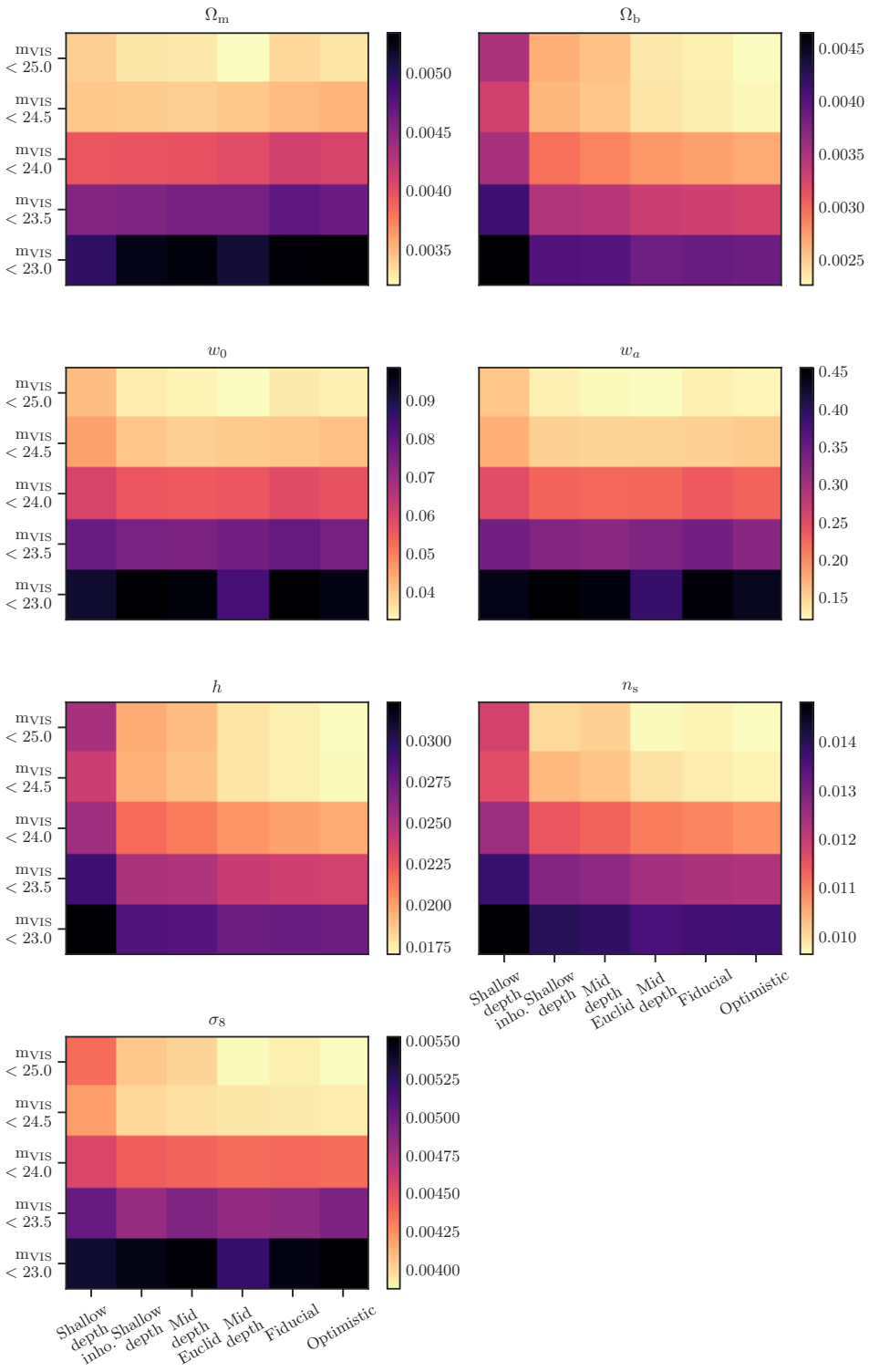

FIGURE 5.11: Uncertainties of the cosmological parameters for all the cases considered in Sec. 5.4.2 for combined GC_{ph} and GGL.

Confidence regions

In addition to the values of the FoM and the uncertainties in the parameters, it is also informative to study the distribution of those uncertainties and the error contours in the determination of pairs of parameters.

The Fisher matrix can be used to project the confidence region of the determination of two parameters. Assuming that the Fisher matrix is Gaussian, the confidence region is an ellipse with semi-minor axis a, semi-major axis b and orientation angle ϕ that are given by:

$$a = A\left[\frac{1}{2}(\mathrm{C}_{\alpha\alpha}+\mathrm{C}_{\beta\beta}) + \left[\frac{1}{4}(\mathrm{C}_{\alpha\alpha}-\mathrm{C}_{\beta\beta})^2+\mathrm{C}^2_{\alpha\beta}\right]^{\frac{1}{2}}\right]^{\frac{1}{2}},$$
$$b = A\left[\frac{1}{2}(\mathrm{C}_{\alpha\alpha}+\mathrm{C}_{\beta\beta}) - \left[\frac{1}{4}(\mathrm{C}_{\alpha\alpha}-\mathrm{C}_{\beta\beta})^2+\mathrm{C}^2_{\alpha\beta}\right]^{\frac{1}{2}}\right]^{\frac{1}{2}},$$
$$\phi = \frac{1}{2}\mathrm{atan}\left(\frac{2\mathrm{C}_{\alpha\beta}}{\mathrm{C}_{\alpha\alpha}-\mathrm{C}_{\beta\beta}}\right),$$

where the constant factor $A^2 = 2.3,\ 6.18,\ 11.8$ depending on the 1, 2, or 3 σ confidence level that one wants to plot, and the covariance matrix elements are defined as $\mathrm{C}_{\alpha\beta} = (\mathrm{F}^{-1})_{\alpha\beta}$, where F is the Fisher matrix as defined in equation 5.14.

In Fig. 5.12 we present the confidence contour plots for our fiducial sample at $\mathrm{m_{VIS}} < 24.5$ and 23.5, to check how the number density affects the constraining power, and compare them to our shallow sample at $\mathrm{m_{VIS}} < 24.5$, to see the impact of having a sample with shallower ground-based photometry. The contours for the $\mathrm{GC_{ph}}$ case are shown in the upper panel and the $\mathrm{GC_{ph}}$ and GGL case in the lower panel. For both probes we see that the fiducial sample gives the best constraints and the largest improvement is gained when the sample size increases. The increase in constraining power with sample size is more prominent in the $\mathrm{GC_{ph}}$ and GGL combined case in general, and for the parameters that characterize dark energy, w_0 and w_a in particular.

5.4.4 Redshift distribution of the photometric redshift bins

To better understand the behavior of the FoM in Sec. 5.4.2 and the constraints in Sec. 5.4.3, we take a closer look at the $n(z)$ of some of the samples used to perform the study. In the top panel of Fig. 5.13 we compare our fiducial photometric sample for $\mathrm{m_{VIS}}$ cuts at < 25, < 24.5, and < 23.5 to see the effects in the $n(z)$ when changing

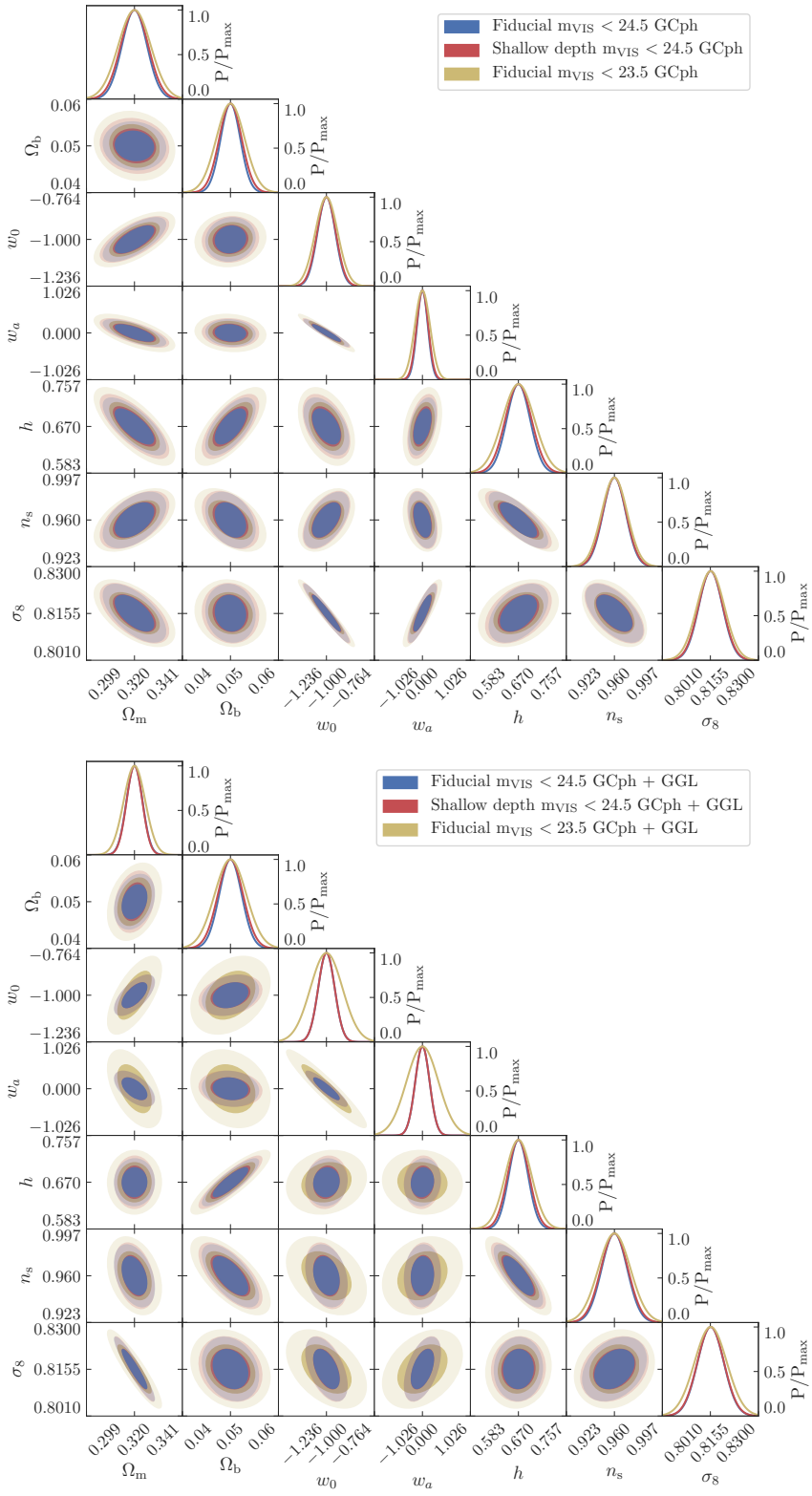

FIGURE 5.12: Fisher matrix contours for our fiducial sample down to $m_{VIS} < 24.5$ (blue) and 23.5 (yellow), and the sample with ground-based photometry degraded by 1.75 magnitudes (red). *Top panel*: For GC_{ph}. *Bottom panel*: For GC_{ph} and GGL.

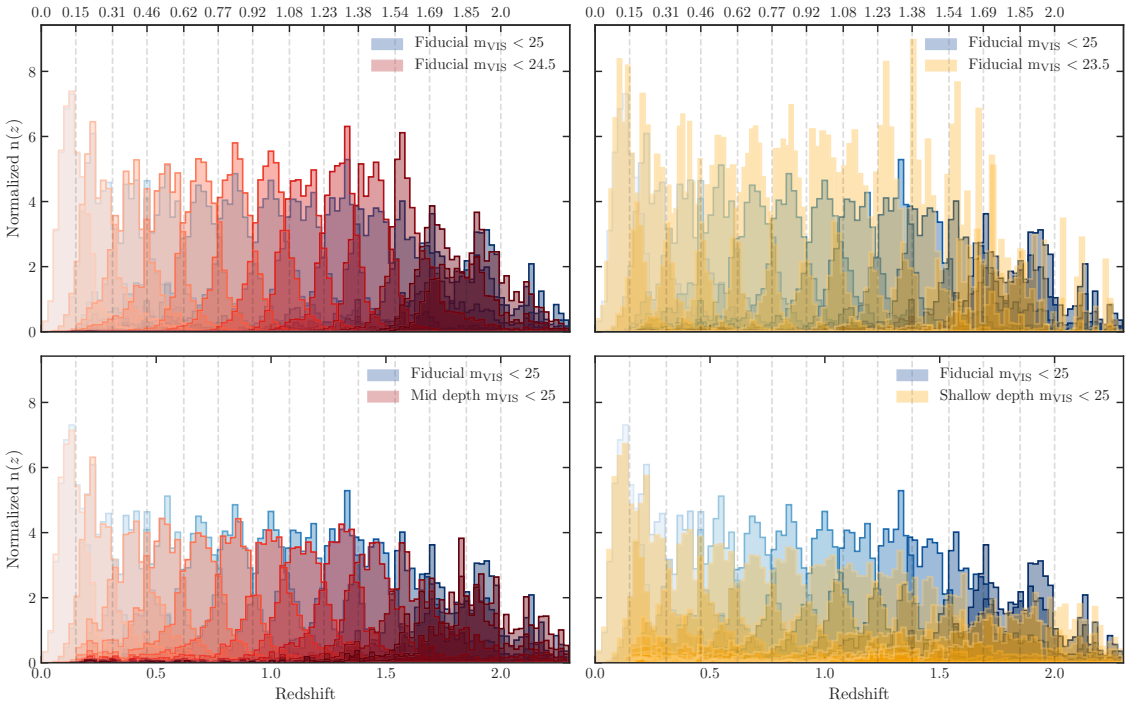

FIGURE 5.13: *Top panels:* Redshift distribution ($z_{\rm mc}$) of each tomographic bin for the fiducial sample at $m_{VIS} < 25$ compared to the fiducial at $m_{VIS} < 24.5$ (left) and 23.5 (right). *Bottom panels:* Redshift distribution ($z_{\rm mc}$) of each redshift bin for the fiducial sample compared to the mid depth (left) and shallow depth samples at $m_{VIS} < 25$ (right).

the magnitude limit and therefore the sample size. A shallower cut in magnitude removes objects at higher redshift. In the bottom panel of the figure we compare the $n(z)$ for the fiducial, mid depth, and shallow depth samples at $m_{VIS} < 25$ to see how the behavior of the $n(z)$ changes with the depth of the ground-based photometry and therefore with the photo-z performance. Overall, the shallower the photometry, the larger the width of the $n(z)$ distributions, especially at higher redshift. This effect spuriously dilutes the correlation signal inside bins and increases the cross-correlation signal between bins, bringing down the GC_{ph} constraining power. On the contrary, for the GGL case the widening of the redshift distributions is less important given the width of the lensing kernel. In addition, the effect of an increase in the number density dominates the performance of the FoM that in general increases with depth.

5.5 Summary and conclusions

Our primary goal was to study the cosmological constraints that can be derived from galaxy clustering studies of photometrically-selected samples using the combination of Euclid and ground-based surveys. For that purpose we used the figure of merit, FoM, defined in equation 5.12 as our performance metric. We wanted to explore the impact of the ground-based photometry depth as well as the photo-z performance on the FoM constraints. To explore the photo-z performance, we varied both the survey depth and the spectroscopic sample available to train the photo-z algorithms. We used the Flagship simulation to create realizations of the expected observed magnitudes and their errors for the survey depths under study. To add a layer of realism to the study, we computed the photo-z using the machine learning code DNF in order to obtain a realistic photo-z estimation for each of the photometric samples under study. We also tried to mimic the training of the photo-z method using spectroscopic samples with different completeness levels. Given the scaled degradation of the photometric quality among the samples, we obtained a gradient of photo-z quality. We chose as our fiducial sample the one corresponding to the photometric depth expected to be available in the Southern hemisphere with a survey like Rubin-LSST. We performed our FoM analysis using the same Fisher forecast formalism as in EC19.

First, we studied the optimization of the FoM with respect to the number and type of tomographic bins. We normalized our results to the case of ten equidistant bins since it is the specifications used in EC19. For this analysis we used the fiducial photometric sample defined in Sec. 5.1.3. Fig. 5.6 shows the variations in the normalized FoM as a function of the number and type of bins. We found the best compromise for an optimal configuration to be:

- Number of bins: A number slightly larger than ten is preferred. We adopted a default value of thirteen bins for our study. For equidistant bins, the FoM increased when moving from ten to thirteen bins by 35.4% and 15.4% for GC_{ph} only and for GC_{ph} + GGL, respectively. We found that a larger number of bins still provided an increase in the FoM for the GC_{ph} only case. However, the photo-z scatter started to be comparable to the bin width for such a large number of bins and our assumptions on how we trained and compute the photo-z started to be too simplistic.

- Type of bins: equidistant. For the GC_{ph} case the FoM increased by 30% for thirteen equidistant bins compared to equipopulated bins. When combining

with GGL the difference in the FoM as a function of bin type was almost negligible.

These results were in nice agreement with Kitching et al. (2019) where they find similar conclusions of the optimum type of bins when optimizing the binning of photometric galaxy samples for cosmic shear analysis. The need of a larger number of bins, especially with good photo-z accuracy and the inclusion of intrinsic alignment parameters, to extract all the necessary information for cosmic shear was also found in Bridle, and King (2007). In this latter study, they also concluded that the model and freedom of the intrinsic alignment parameters greatly impacted the FoM of dark energy.

We further studied the dependence of the FoM on the quality of the photo-z and the size of the sample. We studied possible scenarios of complementary ground-based data for Euclid that could be available in the Southern and Northern hemispheres and in the region in between. We take several magnitude limit cuts and generate realizations of the survey using the Flagship simulation. We also explored different possibilities of spectroscopy data available to train the photo-z techniques. We ended up with a variety of samples with different number densities and photo-z performance properties that tried to encompass the possible samples that will be available for Euclid analyses. We computed the dark energy FoM for all these samples to study its variation. Our results were summarized in Fig. 5.8 and Table 5.3. For the GC_{ph} case, we found a FoM of 713 for our fiducial sample with $m_{VIS} < 24.5$ (remember that galaxy bias is fixed for the GC_{ph} case, providing larger absolute values for the FoM than in the combination of GC_{ph} + GGL). The FoM improved with photo-z quality and sample size. The trend with sample size or magnitude depth reversed when adding galaxies in a magnitude range (between 24.5 and 25 in our case) where photo-zs could not be calibrated and were therefore of poor quality. There was a faster increase of the FoM with sample size in those samples where the photo-z performance was better. For example, in the optimistic, fiducial and mid depth cases increasing the sample size from m_{VIS} 23.5 to 24 and from 24 to 24.5 led to an increase in the FoM of about 20%. When combining GC_{ph} and GGL the FoM for the fiducial sample at $m_{VIS} < 24.5$ was 411. The FoM depended more strongly on the sample size (or survey depth) than on the photo-z performance. The greatest FoM increase, of about 50%, took place when adding galaxies from $m_{VIS} < 24$ to 24.5. The FoM had a weak dependence on the photo-z performance. Generally, it improved with better photo-z accuracy.

The photo-z performance depended on the signal-to-noise of the photometry available and on the spectroscopic sample used in the photo-z algorithm. In our study, we used a machine learning technique, DNF. The representativeness of the training sample had a significant influence on the photo-z quality. The impact on the FoM was larger when the photometry was shallower. For the optimistic photometry, the improvement in the FoM was minimal, 1–2%, when we trained the photo-zs with a representative subsample or with a subsample with a completeness drop at faint m_{VIS}. This minimum variation was because the spectroscopic sample incompleteness in the second case only affected the very faintest galaxies. In the cases where the spectroscopy incompleteness was representative of a larger fraction of the galaxy sample the FoM variation was larger. For example, for our shallowest photometric sample, the relative variation in FoM when trained with an incomplete $n(z)$ and with just a completeness drop only at the faintest m_{VIS}, could be of around 30%.

We also investigated the uncertainties in the constrains of our cosmological parameter set across the photo-z quality and sample density space. Cosmological parameters presented similar trends to those of the FoM. But there were small differences between the different parameters. For GC_{ph}, in general the smallest uncertainty was achieved when we got the highest FoM, which was the optimistic sample at $m_{VIS} < 24.5$. However, Ω_b, w_a and h got the smallest uncertainties for the same optimistic sample but for $m_{VIS} < 25$. The balance between the degradation of the photo-z and the increase in number density affected these parameters slightly differently. For GC_{ph} combined with GGL, the uncertainty in the cosmological parameters presented a similar behavior to the FoM trends. The lowest uncertainty in the parameters was achieved when the number density was largest, at $m_{VIS} < 25$. The trend with photo-z performance did not influence the level of uncertainty. In general the parameters were better constrained when the accuracy on the photo-z determination was higher. However, for some parameters this trend was different in the deepest sample.

To conclude, there was significant gain in the FoM when using a larger number of redshift bins than the nominal ten bins choice of Euclid, especially for GC_{ph}. We studied the effect that the accuracy of the photo-zs and the survey depth had on the FoM. When using the GC_{ph} probe, the FoM increased with survey depth and with the reduction in photo-z uncertainties. We studied the influence of the training sample in the photo-z performance and its implications on the FoM. We found than adding faint galaxies whose redshifts could not be properly determined because there were no galaxies of those magnitudes in the training sample decreased the FoM.

For the combination of the GC_{ph} and GGL probes, there was even more gain on the cosmological constraining power when using larger samples than for GC_{ph} alone. The photo-z quality had slightly less impact on the FoM than for GC_{ph} alone. In general for the combination of probes, the number density had a stronger influence on the FoM than the photo-z accuracy.

Conclusions

As we have highlighted along the thesis, photometric redshifts (photo-zs) are one of the main ingredients necessary to perform cosmological inference when analyzing photometric surveys. Their accuracy directly affects the amount of reliable cosmological information that can be extracted from observations. Their determination is an important process that needs to be thoroughly and carefully handled. In this thesis, we focused on three projects related to the performance of photo-zs and their impact on the cosmological analysis.

Photometry remapping in simulations to mimic observational data

Simulations should accurately represent the photometry and redshifts of real observations in order to perform reliable photo-z related analyses. In the first project of this thesis, we aimed to reduce the existing discrepancies in photometry between real data and simulations. To reduce these differences we transformed the photometry in existing simulations by transferring the photometric statistical properties of real data to the simulations. This process was not trivial as one needs to keep the correlations between different bands not to change the spectral energy distribution of individual galaxies. We explored a few methods to transfer the properties. We focused on obtaining simulated photometry that resembled DES observations in order to have a simulation that allowed to crosscheck ongoing and future analyses in DES that involved photo-zs.

One of the methods we used to transform simulations was the N-dimensional pdf transfer function. This was the first time this method was used in cosmology to transfer statistical properties of observables. Originally the N-dimensional pdf transfer function was designed to transfer the color palettes between pictures. This method allowed us to recover the same photometric distributions of DES in simulations. When determining the photo-z using this transformed photometry, we recovered the same photo-z distribution as in DES. We applied this transformation to the whole catalogue. In addition, we wanted to check that we could recover appropriate subsamples

when introducing selection criteria to the input photometry. So, we applied in the simulations the sample selections for lens samples used in the DES collaboration – the magnitude-limited and the RedMaGiC samples. We recovered similar samples to the ones obtained in DES. Although the recovered distributions were good, the results of cosmological analyses still need to be carefully studied when the modified simulation is used. This is left for future work.

The N-dimensional pdf transfer function transferred the statistical properties in an efficient way and its application did not need much tuning. We think the algorithm has a lot of potential to be used to transfer any other observable property over any pair of real and simulated data. As a future project, this method will be used to transform the photometry of the Dark Energy Spectroscopic Instrument Legacy Imaging Surveys (DESI-Legacy Survey; Dey et al. 2019), which is a combination of public surveys. Among other cosmological motivations, the Legacy Survey was designed to provide the input photometry to select spectroscopic targets for DESI (Dark Energy Spectroscopic Instrument Collaboration: Aghamousa et al., 2016), a Stage IV ground-based dark energy survey whose data will be very useful for synergies with DES, Rubin-LSST and Euclid. As future work, the photometry of the publicly available Data Release 8 [3] (DR8), and soon the Data Release 9 (DR9), of the Legacy Surveys will be transferred to simulations in order to be able to carry out a study of the future performance of DESI. The DR9 will be the photometry used for target selection in DESI, so the transfer of its photometry to simulations can be used to explore target selection including also the emission line strengths.

Target selection and analysis of the color-redshift map with SOMs

In the first part of this project, we used the SOM defined by the C3R2 project to identify and target galaxies that will be useful to fill the regions of the color-redshift map without spectroscopic representation in order to obtain a fully representative spectroscopic sample to calibrate the photo-zs of Euclid. The time awarded at the VLT for the project included the ECDFS field. In this field, the original C3R2 photometry was very limited and we decided to use the DES data instead. For that purpose we computed the color transformations from the DES to the C3R2 system using the VVDS filed where they had photometry in common. We selected galaxies in the ECDFS field to be observed with the VLT telescope with this remapped photometry.

[3] `http://legacysurvey.org/dr8/`

In the second part of this project, we used the SOM technique to explore the color space of the photometry of galaxies from PAUS in order to study the spectroscopic redshift coverage in color space. We considered the same data that are used in photo-z analyses of PAUS to construct the SOM. We explored several sets of parameters as input features to the SOM to obtain the best description of the color space. We found the best input parameters to be the six broad bands magnitude used to compute the photo-z instead of the 40 narrow bands colors. We believe this is due to the high degree of correlation between the narrow bands and the relative large errors at the PAUS magnitude limit considered. Once the SOM was created, we populated it with the spectroscopic sample used to validate the photo-z performance of PAUS. We found that 27.14% of the color-space had no high confidence spectroscopic redshift and that the fraction of galaxies without spectroscopic redshift increased for redshifts larger that 0.7. If galaxies with less reliable spectroscopic redshifts were also considered, the empty color-space decreased to 12.88%. We think that this percentage of missing spectroscopic redshift coverage of the color-redshift mapping can be a source of bias when evaluating the accuracy of photo-zs and when the spectroscopic redshifts are used to determine the photo-zs with training-based algorithms. Similar to the C3R2 project, the SOM would be useful to identify the galaxies that should be observed to have a better spectroscopic coverage.

In addition, we also tried to use the SOM as a tool to detect galaxies with strange colors since the SOM finds patterns in the color space and removing this kind of galaxies can reduce the outliers in photo-z analyses. In order to detect strange galaxies, we used different parameters to try to determine a pattern that would easily detect these galaxies. We did not find a suitable set of parameters to achieve that. So we left further investigation for future work.

In Buchs et al. (2019), the authors present a methodology that uses two SOMs to group galaxies by phenotype in Deep (high confidence redshift) and Wide Field (noisier photometry, less secure redshift with observations in few broad-bands) observations. This phenotype classification of galaxies helps to combine both type of observations by calibrating the likelihood probability that relates the phenotypes between them, thus breaking degeneracies in photo-z between these types of observations. This approach is being applied in surveys such as DES. As a future work, we would apply this approach in the PAUS data by considering it as Deep Field data and using it to describe a phenotype classification with the SOM. We would also like to apply this methodology using the Euclid Consortium Flagship simulation to see

what likelihood probability can be obtain in Euclid, and thus what capability will have Euclid to break photo-z degeneracies.

Optimization of the photometric sample of Euclid for GC analyses

In the last project of this thesis, we studied how the variation in the depth of ground-based observations combined with Euclid observations affected the accuracy of photo-z and hence the cosmological constraining power of Euclid focusing on $\mathrm{GC_{ph}}$ and GGL analyses. In addition, we explored which would be the optimum tomographic redshift binning configuration that would give the tightest cosmological constrains. In order to achieve that, we defined with simulations several Euclid and ground-based galaxy samples with different photometric depth and determined their photo-zs using the same machine learning method and mimicking the training samples that we would use with real observations instead of modeling the photo-zs uncertainties. We also checked how the number density of the galaxy samples impacted the cosmological constrains by taking several magnitude limit cuts. Then, the Fisher matrix formalism was used to perform a realistic forecast of the constraining power of the different galaxy samples and configurations.

We found that increasing the number of tomographic bins compared to the fiducial number of ten in the Euclid cosmological probes analysis improved the resulting Figure of Merit (FoM). We also found that equidistant bins in redshift performed slightly better than equipopulated bins. Increasing the number of redshift bins from ten to thirteen improved the FoM by 35% and 15% for $\mathrm{GC_{ph}}$ and its combination with GGL, respectively. So for future Euclid analyses it might be worth using more than ten tomographic bins to increase the cosmological constraining power. When varying the photometric depth and therefore the number of galaxies used in the samples under study, we saw that, for $\mathrm{GC_{ph}}$, an increase of the survey depth provided a higher FoM. However, when the number size increased, the inclusion of faint galaxies beyond the limit of the spectroscopic training data decreased the FoM due to the addition of unreliable photo-zs with larger errors. For the combination of $\mathrm{GC_{ph}}$ and GGL, the number density of the sample was the factor that most influenced the variations in the FoM. Adding galaxies at faint magnitudes and thus mostly at high redshift increased the FoM even when they were beyond the spectroscopic limit, since the increase of the number density compensated the photo-z degradation in this case. The broad shape of the lensing kernel makes the constraining power not be affected as much by the increasing photo-z uncertainty, while the higher density reduces the shot noise.

In this work, we varied the photometric depth of ground-based observation assuming that the magnitude limit in each band was isotropic. However, Euclid will observe a large fraction of the sky using several ground-based surveys of the Northern and Southern hemisphere to complement its data. This photometry will have different magnitude limit depths and uncertainties at different positions on the sky. The next layer of realism in our forecasts would be to generate several sets of ground-based photometry according to the specific characteristics of each ground-based survey in the region of the sky covered, in order to reproduce the expected anisotropy of the photometry. Then the different sets of ground-based photometry would be combined and added to the Euclid photometry in order to determine the photo-zs and redo the optimization analysis performed in this work. We could also include the photometric redshift pipeline being developed by the Euclid Photometric Redshifts Organization Unit to compute the photo-z used in this forecast.

To summarize, we have developed several tools to help us understand the performance and analysis of cosmological surveys. We have paid special attention to reproduce the photometric properties of observational imaging surveys. In particular, we have included the novel tool of N-dimensional pdf transfer to remap photometry. We have used the Self Organizing Map method to partition the color space to be able to understand the mapping between colors and redshift, help in the spectroscopic target selection in the C3R2 program and understand the photometric properties of the PAUS survey. We have worked in the optimization of the clustering analysis of photometric samples and galaxy-galaxy lensing in the Euclid survey. For that purpose, we have run photometric redshift codes, Fisher Matrix analysis and the CosmoSIS package to provide cosmological forecasts.

Bibliography

(1) Abbott, T. M. C. et al. (2018). Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D 98* 043526, 043526.

(2) Abbott, T. M. C. et al. (2019). Dark Energy Survey year 1 results: Joint analysis of galaxy clustering, galaxy lensing, and CMB lensing two-point functions. *Physical Review D 100* 023541, 023541.

(3) Abbott, T. M. C. et al. (2018). The Dark Energy Survey: Data Release 1. *The Astrophysical Journal Supplement Series 239* 18, 18.

(4) Aihara, H. et al. (2018). The Hyper Suprime-Cam SSP Survey: Overview and survey design. *Publications of the Astronomical Society of Japan 70* S4, S4.

(5) Albrecht, A., Bernstein, G., Cahn, R., Freedman, W. L., Hewitt, J., Hu, W., Huth, J., Kamionkowski, M., Kolb, E. W., Knox, L., Mather, J. C., Staggs, S., and Suntzeff, N. B. (2006). Report of the Dark Energy Task Force. *arXiv e-prints*.

(6) Alpher, R. A., and Herman, R. (1948). Evolution of the Universe. *Nature 162*, 774–775.

(7) Appenzeller, I. et al. (1998). Successful commissioning of FORS1 - the first optical instrument on the VLT. *The Messenger 94*, 1–6.

(8) Atsushi, T., *Structural Formation and Dark Matter in Space*, 2006.

(9) Bartelmann, M., and Schneider, P. (2001). Weak gravitational lensing. *Physics Reports 340*, 291–472.

(10) Benítez, N. et al. (2009). Measuring Baryon Acoustic Oscillations Along the Line of Sight with Photometric Redshifts: The PAU Survey. *The Astrophysical Journal 691*, 241–260.

(11) Benítez, N. (2000). Bayesian Photometric Redshift Estimation. *The Astrophysical Journal 536*, 571–583.

(12) Bielby, R., Hudelot, P., McCracken, H. J., Ilbert, O., Daddi, E., Le Fèvre, O., Gonzalez-Perez, V., Kneib, J. P., Marmo, C., Mellier, Y., Salvato, M., Sanders, D. B., and Willott, C. J. (2012). The WIRCam Deep Survey. I. Counts, colours, and mass-functions derived from near-infrared imaging in the CFHTLS deep fields. *Astronomy & Astrophysics 545* A23, A23.

(13) Bird, S., Viel, M., and Haehnelt, M. G. (2012). Massive neutrinos and the non-linear matter power spectrum. *Monthly Notices of the Royal Astronomical Society 420*, 2551–2561.

(14) Blanton, M. R., Lupton, R. H., Schlegel, D. J., Strauss, M. A., Brinkmann, J., Fukugita, M., and Loveday, J. (2005). The Properties and Luminosity Function of Extremely Low Luminosity Galaxies. *The Astrophysical Journal 631*, 208–230.

(15) Blanton, M. R., Schlegel, D. J., Strauss, M. A., Brinkmann, J., Finkbeiner, D., Fukugita, M., Gunn, J. E., Hogg, D. W., Ivezić, Ž., Knapp, G. R., Lupton, R. H., Munn, J. A., Schneider, D. P., Tegmark, M., and Zehavi, I. (2005). New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys. *The Astronomical Journal 129*, 2562–2578.

(16) Blanton, M. R. et al. (2003). The Broadband Optical Properties of Galaxies with Redshifts $0.02<z<0.22$. *The Astrophysical Journal 594*, 186–207.

(17) Blazek, J. A., MacCrann, N., Troxel, M. A., and Fang, X. (2019). Beyond linear galaxy alignments. *Physical Review D 100* 103506, 103506.

(18) Bridle, S., and King, L. (2007). Dark energy constraints from cosmic shear power spectra: impact of intrinsic alignments on photometric redshift requirements. *New Journal of Physics 9*, 444.

(19) Buchs, R. et al. (2019). Phenotypic redshifts with self-organizing maps: A novel method to characterize redshift distributions of source galaxies for weak lensing. *Monthly Notices of the Royal Astronomical Society 489*, 820–841.

(20) Capak, P. et al. (2019). Enhancing LSST Science with Euclid Synergy. *arXiv e-prints*.

(21) Capak, P. et al. (2007). The First Release COSMOS Optical and Near-IR Data and Catalog. *Astrophysical Journal Supplement Series 172*, 99–116.

(22) Carrasco Kind, M., and Brunner, R. J. (2014). SOMz: photometric redshift PDFs with self-organizing maps and random atlas. *Monthly Notices of the Royal Astronomical Society 438*, 3409–3421.

(23) Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., and Fosalba, P. (2015). An algorithm to build mock galaxy catalogues using MICE simulations. *Monthly Notices of the Royal Astronomical Society 447*, 646–670.

(24) Casas, R., Cardiel-Sas, L., Castander, F. J., Díaz, C., Gaweda, J., Jiménez Rojas, J., Jiménez, S., Lamensans, M., Padilla, C., Rodriguez, F. J., Sanchez, E., and Sevilla Noarbe, I. (2016). Characterization and performance of PAUCam filters. *9908* 99084K, 99084K.

(25) Chambers, K. C. et al. (2016). The Pan-STARRS1 Surveys. *arXiv e-prints* arXiv:1612.05560, arXiv:1612.05560.

(26) Chevallier, M., and Polarski, D. (2001). Accelerating Universes with Scaling Dark Matter. *International Journal of Modern Physics D 10*, 213–223.

(27) Chisari, N., Codis, S., Laigle, C., Dubois, Y., Pichon, C., Devriendt, J., Slyz, A., Miller, L., Gavazzi, R., and Benabed, K. (2015). Intrinsic alignments of galaxies in the Horizon-AGN cosmological hydrodynamical simulation. *Monthly Notices of the Royal Astronomical Society 454*, 2736–2753.

(28) Cooray, A., and Sheth, R. (2002). Halo models of large scale structure. *Physics Reports 372*, 1–129.

(29) Costille, A., Caillat, A., Rossin, C., Pascal, S., Sanchez, P., Barette, R., Laurent, P., Foulon, B., and Pariès, C. In *Space Telescopes and Instrumentation 2018: Optical, Infrared, and Millimeter Wave*, ed. by Lystrup, M., MacEwen, H. A., Fazio, G. G., Batalha, N., Siegler, N., and Tong, E. C., 2018; Vol. 10698, 106982B.

(30) Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., and Carretero, J. (2015). The MICE Grand Challenge lightcone simulation - II. Halo and galaxy catalogues. *Monthly Notices of the Royal Astronomical Society 453*, 1513–1530.

(31) Crocce, M. et al. (2019). Dark Energy Survey year 1 results: galaxy sample for BAO measurement. *Monthly Notices of the Royal Astronomical Society 482*, 2807–2822.

(32) Crocce, M. et al. (2019). Dark Energy Survey year 1 results: galaxy sample for BAO measurement. *Monthly Notices of the Royal Astronomical Society 482*, 2807–2822.

(33) Crocce, M. et al. (2016). Galaxy clustering, photometric redshifts and diagnosis of systematics in the DES Science Verification data. *Monthly Notices of the Royal Astronomical Society 455*, 4301–4324.

(34) Cropper, M. et al. In *Space Telescopes and Instrumentation 2018: Optical, Infrared, and Millimeter Wave*, ed. by Lystrup, M., MacEwen, H. A., Fazio, G. G., Batalha, N., Siegler, N., and Tong, E. C., 2018; Vol. 10698, p 1069828.

(35) Dark Energy Spectroscopic Instrument Collaboration: Aghamousa, A. et al. (2016). The DESI Experiment Part I: Science, Targeting, and Survey Design. *arXiv e-prints*.

(36) Dark Energy Survey Collaboration (2005). The Dark Energy Survey. *arXiv: astro-ph/0510346*.

(37) Dark Energy Survey Collaboration: Abbott, T. et al. (2018). The Dark Energy Survey: Data Release 1. *The Astrophysical Journal Supplement Series 239* 18, 18.

(38) Dark Energy Survey Collaboration: Abbott, T. et al. (2016). The Dark Energy Survey: more than dark energy - an overview. *Monthly Notices of the Royal Astronomical Society 460*, 1270–1299.

(39) Davis, M. et al. (2007). The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets. *Astrophysical Journal Letters 660*, L1–L6.

(40) de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., and Valentijn, E. A. (2013). The Kilo-Degree Survey. *Experimental Astronomy 35*, 25–44.

(41) De Vicente, J., Sánchez, E., and Sevilla-Noarbe, I. (2016). DNF - Galaxy photometric redshift by Directional Neighbourhood Fitting. *Monthly Notices of the Royal Astronomical Society 459*, 3078–3088.

(42) Debono, I., and Smoot, G. F. (2016). General Relativity and Cosmology: Unsolved Questions and Future Directions. *Universe 2* 23, 23.

(43) DeRose, J. et al. (2019). The Buzzard Flock: Dark Energy Survey Synthetic Sky Catalogs.

(44) Deshpande, A. C. et al. (2020). Euclid: The reduced shear approximation and magnification bias for Stage IV cosmic shear experiments. *Astronomy & Astrophysics 636* A95, A95.

(45) Desjacques, V., Jeong, D., and Schmidt, F. (2018). Large-scale galaxy bias. *Physics Reports 733*, 1–193.

(46) Dey, A. et al. (2019). Overview of the DESI Legacy Imaging Surveys. *The Astronomical Journal 157* 168, 168.

(47) Diehl, H. T. et al. (2017). The DES Bright Arcs Survey: Hundreds of Candidate Strongly Lensed Galaxy Systems from the Dark Energy Survey Science Verification and Year 1 Observations. *The Astrophysical Journal Supplement Series 232* 15, 15.

(48) Drlica-Wagner, A. et al. (2018). Dark Energy Survey Year 1 Results: The Photometric Data Set for Cosmology. *The Astrophysical Journal Supplement Series 235*, 33.

(49) Einstein, A. (1915). Die Feldgleichungen der Gravitation. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin*, 844–847.

(50) Einstein, A. (1917). Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin*, 142–152.

(51) Einstein, A. (1905). Zur Elektrodynamik bewegter Körper. *Annalen der Physik 322*, 891–921.

(52) Eisenstein, D. J. et al. (2001). Spectroscopic Target Selection for the Sloan Digital Sky Survey: The Luminous Red Galaxy Sample. *The Astronomical Journal 122*, 2267–2280.

(53) Elvin-Poole, J. et al. (2018). Dark Energy Survey year 1 results: Galaxy clustering for combined probes. *Physical Review D 98* 042006, 042006.

(54) Eriksen, M. et al. (2019). The PAU Survey: early demonstration of photometric redshift performance in the COSMOS field. *Monthly Notices of the Royal Astronomical Society 484*, 4200–4215.

(55) Eriksen, M. et al. (2020). The PAU Survey: Photometric redshifts using transfer learning from simulations. *Monthly Notices of the Royal Astronomical Society 497*, 4565–4579.

(56) Eriksen, M., and Gaztañaga, E. (2015). Combining spectroscopic and photometric surveys using angular cross-correlations - II. Parameter constraints from different physical effects. *Monthly Notices of the Royal Astronomical Society 452*, 2168–2184.

(57) Euclid Collaboration: Blanchard, A. et al. (2019). Euclid preparation: VII. Forecast validation for Euclid cosmological probes. *arXiv e-prints* arXiv:1910.09273, arXiv:1910.09273.

(58) Faber, S. M. et al. (2003). The DEIMOS spectrograph for the Keck II Telescope: integration and testing. *4841*, ed. by Iye, M., and Moorwood, A. F. M., 1657–1669.

(59) Flaugher, B. et al. (2015). The Dark Energy Camera. *The Astronomical Journal 150* 150, 150.

(60) Fortuna, M. C., Hoekstra, H., Joachimi, B., Johnston, H., Chisari, N. E., Georgiou, C., and Mahony, C. (2020). The halo model as a versatile tool to predict intrinsic alignments. *arXiv e-prints* arXiv:2003.02700, arXiv:2003.02700.

(61) Fosalba, P., Crocce, M., Gaztañaga, E., and Castander, F. J. (2015). The MICE grand challenge lightcone simulation - I. Dark matter clustering. *Monthly Notices of the Royal Astronomical Society 448*, 2987–3000.

(62) Fosalba, P., Gaztañaga, E., Castander, F. J., and Crocce, M. (2015). The MICE Grand Challenge light-cone simulation - III. Galaxy lensing mocks from all-sky lensing maps. *Monthly Notices of the Royal Astronomical Society 447*, 1319–1332.

(63) Freeman, K. C. (1970). On the Disks of Spiral and S0 Galaxies. *Astrophysical Journal 160*, 811.

(64) Friedmann, A. (1922). Über die Krümmung des Raumes. *Zeitschrift fur Physik 10*, 377–386.

(65) Friedmann, A. (1924). Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes. *Zeitschrift fur Physik 21*, 326–332.

(66) Furusawa, H., Kosugi, G., Akiyama, M., Takata, T., Sekiguchi, K., and Furusawa, J. (2008). Subaru/XMM-Newton Deep Survey (SXDS) - Optical Imaging Survey and Photometric Catalogs. *399*, ed. by Kodama, T., Yamada, T., and Aoki, K., 131.

(67) Gamow, G. (1946). Expanding Universe and the Origin of Elements. *Physical Review 70*, 572–573.

(68) Gatti, M. et al. (2018). Dark Energy Survey Year 1 results: cross-correlation redshifts - methods and systematics characterization. *Monthly Notices of the Royal Astronomical Society 477*, 1664–1682.

(69) Gawiser, E. et al. (2006). The Multi-wavelength Survey by Yale-Chile (MUSYC): Survey Design and Deep Public UBVRIzÍmages and Catalogs of the Extended Hubble Deep Field-South. *Astrophysical Journal 162*, 1–19.

(70) Gaztañaga, E., Eriksen, M., Crocce, M., Castander, F. J., Fosalba, P., Marti, P., Miquel, R., and Cabré, A. (2012). Cross-correlation of spectroscopic and photometric galaxy surveys: cosmology from lensing and redshift distortions. *Monthly Notices of the Royal Astronomical Society 422*, 2904–2930.

(71) Gschwend, J. et al. (2018). DES science portal: Computing photometric redshifts. *Astronomy and Computing 25*, 58–80.

(72) Guglielmo, V. et al. (2020). Euclid preparation: VIII. The Complete Calibration of the Colour-Redshift Relation survey: VLT/KMOS observations and data release.

(73) Guzzo, L. et al. (2014). The VIMOS Public Extragalactic Redshift Survey (VIPERS). An unprecedented view of galaxies and large-scale structure at $0.5 < z < 1.2$. *Astronomy & Astrophysics 566* A108, A108.

(74) Harrison, E. R. (1970). Fluctuations at the Threshold of Classical Cosmology. *Physical Review D 1*, 2726–2730.

(75) Heymans, C. et al. (2012). CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey. *Monthly Notices of the Royal Astronomical Society 427*, 146–166.

(76) Hildebrandt, H. et al. (2012). CFHTLenS: improving the quality of photometric redshifts with precision photometry. *Monthly Notices of the Royal Astronomical Society 421*, 2355–2367.

(77) Hirata, C. M., Mandelbaum, R., Ishak, M., Seljak, U., Nichol, R., Pimbblet, K. A., Ross, N. P., and Wake, D. (2007). Intrinsic galaxy alignments from the 2SLAQ and SDSS surveys: luminosity and redshift scalings and implications for weak lensing surveys. *Monthly Notices of the Royal Astronomical Society 381*, 1197–1218.

(78) Hoyle, B. et al. (2018). Dark Energy Survey Year 1 Results: redshift distributions of the weak-lensing source galaxies. *Monthly Notices of the Royal Astronomical Society 478*, 592–610.

(79) Hu, W., and Sawicki, I. (2007). Parametrized post-Friedmann framework for modified gravity. *Physical Review D 76*, DOI: 10.1103/PhysRevD.76.104043.

(80) Hubble, E. (1929). A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science 15*, 168–173.

(81) Ibata, R. A. et al. (2017). The Canada-France Imaging Survey: First Results from the u-Band Component. *The Astrophysical Journal 848* 128, 128.

(82) Ivezić, Ž. et al. (2019). LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal 873* 111, 111.

(83) Kaiser, N. (1987). Clustering in real space and in redshift space. *Monthly Notices of the Royal Astronomical Society 227*, 1–21.

(84) Kilbinger, M., Heymans, C., Asgari, M., Joudaki, S., Schneider, P., Simon, P., Van Waerbeke, L., Harnois-Déraps, J., Hildebrand t, H., Köhlinger, F., Kuijken, K., and Viola, M. (2017). Precision calculations of the cosmic shear power spectrum projection. *Monthly Notices of the Royal Astronomical Society 472*, 2126–2141.

(85) Kitching, T. D., Alsing, J., Heavens, A. F., Jimenez, R., McEwen, J. D., and Verde, L. (2017). The limits of cosmic shear. *Monthly Notices of the Royal Astronomical Society 469*, 2737–2749.

(86) Kitching, T. D., Taylor, P. L., Capak, P., Masters, D., and Hoekstra, H. (2019). Rainbow Cosmic Shear: Optimisation of Tomographic Bins. *arXiv e-prints*.

(87) Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics 43*, 59–69.

(88) Laigle, C. et al. (2016). The COSMOS2015 Catalog: Exploring the 1 < z < 6 Universe with Half a Million Galaxies. *Astrophysical Journal Supplement Series 224* 24, 24.

(89) Laureijs, R. et al. (2011). Euclid Definition Study Report. *arXiv e-prints* arXiv:1110.3193, arXiv:1110.3193.

(90) Le Fèvre, O., Mellier, Y., McCracken, H. J., Foucaud, S., Gwyn, S., Radovich, M., Dantel-Fort, M., Bertin, E., Moreau, C., Cuillandre, J. C., Pierre, M., Le Brun, V., Mazure, A., and Tresse, L. (2004). The VIRMOS deep imaging survey. I. Overview, survey strategy, and CFH12K observations. *Astronomy & Astrophysics 417*, 839–846.

(91) Le Fèvre, O. et al. (2013). The VIMOS VLT Deep Survey final data release: a spectroscopic sample of 35 016 galaxies and AGN out to z ~6.7 selected with $17.5 \leq \mathrm{i}_{AB} \leq 24.75$. *Astronomy & Astrophysics 559*, A14.

(92) Lehmer, B. D. et al. (2005). The Extended Chandra Deep Field-South Survey: Chandra Point-Source Catalogs. *Astrophysical Journal Supplement Series 161*, 21–40.

(93) Lilly, S. J. et al. (2007). zCOSMOS: A Large VLT/VIMOS Redshift Survey Covering $0 < \mathrm{z} < 3$ in the COSMOS Field. *The Astrophysical Journal Supplement Series 172*, 70–85.

(94) Lilly, S. J. et al. (2009). The zCOSMOS 10k-Bright Spectroscopic Sample. *The Astrophysical Journal Supplement Series 184*, 218–229.

(95) Lima, J. A. S., and Santos, R. C. (2017). 100 Years of Relativistic Cosmology (1917-2017). Part I: From Origins to the Discovery of Universal Expansion (1929). *arXiv e-prints*.

(96) Linder, E. V. (2005). Cosmic growth history and expansion history. *Physical Review D 72*, DOI: `10.1103/PhysRevD.72.043529`.

(97) LSST Science Collaboration: Abell et al. (2009). LSST Science Book, Version 2.0. *arXiv e-prints*.

(98) Lupton, R. H., Gunn, J. E., Szalay, A. S., and et, a. (1999). A Modified Magnitude System that Produces Well-Behaved Magnitudes, Colors, and Errors Even for Low Signal-to-Noise Ratio Measurements. *The Astronomical Journal 118*, 1406–1410.

(99) Martí, P., Miquel, R., Castander, F. J., Gaztañaga, E., Eriksen, M., and Sánchez, C. (2014). Precise photometric redshifts with a narrow-band filter set: the PAU survey at the William Herschel Telescope. *Monthly Notices of the Royal Astronomical Society 442*, 92–109.

(100) Masters, D. et al. (2015). Mapping the Galaxy Color-Redshift Relation: Optimal Photometric Redshift Calibration Strategies for Cosmology Surveys. *The Astrophysical Journal 813* 53, 53.

(101) Masters, D. C., Stern, D. K., Cohen, J. G., Capak, P. L., Rhodes, J. D., Castander, F. J., and Paltani, S. (2017). The Complete Calibration of the Color-Redshift Relation (C3R2) Survey: Survey Overview and Data Release 1. *Astrophysical Journal 841* 111, 111.

(102) Masters, D. C., Stern, D. K., Cohen, J. G., Capak, P. L., Stanford, S. A., Hernitschek, N., Galametz, A., Davidzon, I., Rhodes, J. D., Sand ers, D., Mobasher, B., Castander, F., Pruett, K., and Fotopoulou, S. (2019). The Complete Calibration of the Color-Redshift Relation (C3R2) Survey: Analysis and Data Release 2. *Astrophysical Journal 877* 81, 81.

(103) McCracken, H. J., Radovich, M., Bertin, E., Mellier, Y., Dantel-Fort, M., Le Fèvre, O., Cuilland re, J. C., Gwyn, S., Foucaud, S., and Zamorani, G. (2003). The VIRMOS deep imaging survey. II: CFH12K BVRI optical data for the 0226-04 deep field. *Astronomy & Astrophysics 410*, 17–32.

(104) McCracken, H. J. et al. (2012). UltraVISTA: a new ultra-deep near-infrared survey in COSMOS. *Astronomy & Astrophysics 544* A156, A156.

(105) McLean, I. S. et al. (2012). MOSFIRE, the multi-object spectrometer for infra-red exploration at the Keck Observatory. *8446* 84460J, 84460J.

(106) Mead, A. J., Heymans, C., Lombriser, L., Peacock, J. A., Steele, O. I., and Winther, H. A. (2016). Accurate halo-model matter power spectra with dark energy, massive neutrinos and modified gravitational forces. *Monthly Notices of the Royal Astronomical Society 459*, 1468–1488.

(107) Morganson, E. et al. (2018). The Dark Energy Survey Image Processing Pipeline. *Publications of the Astronomical Society of the Pacific 130*, 074501.

(108) Neal, R. (2003). Slice sampling. *Annals of Statistics 31*, 705–767.

(109) Newman, J. A. et al. (2015). Spectroscopic needs for imaging dark energy experiments. *Astroparticle Physics 63*, 81–100.

(110) Newman, J. A. et al. (2013). The DEEP2 Galaxy Redshift Survey: Design, Observations, Data Reduction, and Redshifts. *The Astrophysical Journal Supplement Series 208* 5, 5.

(111) Nicola, A., Alonso, D., Sánchez, J., Slosar, A., Awan, H., Broussard, A., Dunkley, J., Gawiser, E., Gomes, Z., Mand elbaum, R., Miyatake, H., Newman, J. A., Sevilla-Noarbe, I., Skinner, S., and Wagoner, E. L. (2020). Tomographic galaxy clustering with the Subaru Hyper Suprime-Cam first year public data release. *Journal of Cosmology and Astroparticle Physics 2020* 044, 044.

(112) Oke, J. B., Cohen, J. G., Carr, M., Cromer, J., Dingizian, A., Harris, F. H., Labrecque, S., Lucinio, R., Schaal, W., Epps, H., and Miller, J. (1995). The Keck Low-Resolution Imaging Spectrometer. *Publications of the Astronomical Society of the Pacific 107*, 375.

(113) Padilla, C. et al. (2019). The Physics of the Accelerating Universe Camera. *The Astronomical Journal 157* 246, 246.

(114) Peebles, P. J. E., and Yu, J. T. (1970). Primeval Adiabatic Perturbation in an Expanding Universe. *Astrophysical Journal 162*, 815.

(115) Penzias, A. A., and Wilson, R. W. (1965). A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal 142*, 419–421.

(116) Perlmutter, S. et al. (1999). Measurements of Ω and Λ from 42 High-Redshift Supernovae. *The Astronomical Journal 517*, 565–586.

(117) Pitié, F., Kokaram, A. C., and Dahyot, R. (2005). N-Dimensional Probability Density Function Transfer and its Application to Colour Transfer. *Proceedings of the Tenth IEEE International Conference on Computer Vision*, 1550–5499.

(118) Planck Collaboration et al. (2020). Planck 2018 results. I. Overview and the cosmological legacy of Planck. *Astronomy & Astrophysics 641* A1, A1.

(119) Planck Collaboration et al. (2020). Planck 2018 results. X. Constraints on inflation. *Astronomy & Astrophysics 641* A10, A10.

(120) Pocino, A., Tutusaus, I., Castander, F., Fosalba, P., Crocce, M., and Porredon, A. (in prep.). Euclid preparation. XII. Optimizing the photometric sample of the Euclid survey for galaxy clustering analyses.

(121) Porredon, A. et al. (2020). Dark Energy Survey Year 3 Results: Optimizing the Lens Sample in Combined Galaxy Clustering and Galaxy-Galaxy Lensing Analysis. *arXiv e-prints*.

(122) Potter, D., Stadel, J., and Teyssier, R. (2017). PKDGRAV3: beyond trillion particle cosmological simulations for the next era of galaxy surveys. *Computational Astrophysics and Cosmology 4* 2, 2.

(123) Prakash, A., Licquia, T. C., Newman, J. A., and Rao, S. M. (2015). Luminous Red Galaxies: Selection and Classification By Combining Optical and Infrared Photometry. *The Astrophysical Journal 803* 105, 105.

(124) Racca, G., Laureijs, R., and Mellier, Y. (2018). The Euclid Mission at the Critical Design Review. *42*, E1.16–3–18.

(125) Racca, G. D. et al. (2016). The Euclid mission design. *9904* 99040O, ed. by MacEwen, H. A., Fazio, G. G., Lystrup, M., Batalha, N., Siegler, N., and Tong, E. C., 99040O.

(126) Rau, M. M., Hoyle, B., Paech, K., and Seitz, S. (2017). Correcting cosmological parameter biases for all redshift surveys induced by estimating and reweighting redshift distributions. *Monthly Notices of the Royal Astronomical Society 466*, 2927–2938.

(127) Rhodes, J. et al. (2017). Scientific Synergy between LSST and Euclid. *The Astrophysical Journal Supplement Series 233* 21, 21.

(128) Riess, A. G. et al. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *The Astronomical Journal 116*, 1009–1038.

(129) Rozo, E. et al. (2016). redMaGiC: selecting luminous red galaxies from the DES Science Verification data. *Monthly Notices of the Royal Astronomical Society 461*, 1431–1450.

(130) Rubin, V. C., and Ford, J., W. Kent (1970). Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. *Astrophysical Journal 159*, 379.

(131) Rykoff, E. S. et al. (2016). The RedMaPPer Galaxy Cluster Catalog From DES Science Verification Data. *The Astrophysical Journal Supplement Series 224* 1, 1.

(132) Samuroff, S. et al. (2019). Dark Energy Survey Year 1 results: constraints on intrinsic alignments and their colour dependence from galaxy clustering and weak lensing. *Monthly Notices of the Royal Astronomical Society 489*, 5453–5482.

(133) Sánchez, A. G. et al. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the configuration-space clustering wedges. *Monthly Notices of the Royal Astronomical Society 464*, 1640–1658.

(134) Sánchez, C., Raveri, M., Alarcon, A., and Bernstein, G. M. (2020). Propagating sample variance uncertainties in redshift calibration: simulations, theory, and application to the COSMOS2015 data. *Monthly Notices of the Royal Astronomical Society 498*, 2984–2999.

(135) Scoville, N. et al. (2007). The Cosmic Evolution Survey (COSMOS): Overview. *Astrophysical Journal Supplement Series 172*, 1–8.

(136) Sevilla-Noarbe, I. et al. (in prep.). The Dark Energy Survey Year 3 Results: Photometric Data Set for Cosmology. *The Astrophysical Journal Supplement Series*.

(137) Sharples, R. et al. (2013). First Light for the KMOS Multi-Object Integral-Field Spectrometer. *The Messenger 151*, 21–23.

(138) Smith, J. A. et al. (2002). The u'g'r'i'z' Standard-Star System. *The Astronomical Journal 123*, 2121–2144.

(139) Smith, R. E., Peacock, J. A., Jenkins, A., White, S. D. M., Frenk, C. S., Pearce, F. R., Thomas, P. A., Efstathiou, G., and Couchman, H. M. P. (2003). Stable clustering, the halo model and non-linear cosmological power spectra. *Monthly Notices of the Royal Astronomical Society 341*, 1311–1332.

(140) Takahashi, R., Sato, M., Nishimichi, T., Taruya, A., and Oguri, M. (2012). Revising the Halofit Model for the Nonlinear Matter Power Spectrum. *The Astrophysical Journal 761* 152, 152.

(141) Taniguchi, Y. et al. (2015). The Subaru COSMOS 20: Subaru optical imaging of the HST COSMOS field with 20 filters*. *Publications of the Astronomical Society of Japan 67* 104, 104.

(142) Tanoglidis, D., Chang, C., and Frieman, J. (2020). Optimizing galaxy samples for clustering measurements in photometric surveys. *Monthly Notices of the Royal Astronomical Society 491*, 3535–3552.

(143) Taylor, E. N. et al. (2009). A Public, K-Selected, Optical-to-Near-Infrared Catalog of the Extended Chandra Deep Field South (ECDFS) from the Multiwavelength Survey by Yale-Chile (MUSYC). *The Astrophysical Journal Supplement Series 183*, 295–319.

(144) Taylor, P. L., Kitching, T. D., McEwen, J. D., and Tram, T. (2018). Testing the cosmic shear spatially-flat universe approximation with generalized lensing and shear spectra. *Physical Review D 98*, 023522.

(145) Tonello, N. et al. (2019). The PAU Survey: Operation and orchestration of multi-band survey data. *Astronomy and Computing 27* 171, 171.

(146) Troxel, M. A. et al. (2018). Dark Energy Survey Year 1 results: Cosmological constraints from cosmic shear. *Physical Review D 98* 043528, 043528.

(147) Tutusaus, I. et al. (2020). Euclid: The importance of galaxy clustering and weak lensing cross-correlations within the photometric Euclid survey. *arXiv e-prints* arXiv:2005.00055, arXiv:2005.00055.

(148) van Uitert, E. et al. (2018). KiDS+GAMA: cosmology constraints from a joint analysis of cosmic shear, galaxy-galaxy lensing, and angular clustering. *Monthly Notices of the Royal Astronomical Society 476*, 4662–4689.

(149) Walcher, J., Groves, B., Budavári, T., and Dale, D. (2011). Fitting the integrated spectral energy distributions of galaxies. *Astrophysics and Space Science 331*, 1–52.

(150) Williams, S. C., Hook, I. M., Hayden, B., Nordin, J., Aldering, G., Boone, K., Goobar, A., Lidman, C. E., Perlmutter, S., Rubin, D., Ruiz-Lapuente, P., Saunders, C., and Supernova Cosmology Project (2020). See Change: VLT spectroscopy of a sample of high-redshift Type Ia supernova host galaxies. *Monthly Notices of the Royal Astronomical Society 495*, 3859–3880.

(151) Wright, A. H., Hildebrandt, H., van den Busch, J. L., and Heymans, C. Photometric Redshift Calibration with Self Organising Maps, 2019.

(152) York, D. G. et al. (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal 120*, 1579–1587.

(153) Zehavi, I. et al. (2011). Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. *The Astrophysical Journal 736* 59, 59.

(154) Zeldovich, Y. B. (1972). A hypothesis, unifying the structure and the entropy of the Universe. *Monthly Notices of the Royal Astronomical Society 160*, 1P.

(155) Zeldovich, Y. B. (1970). Gravitational instability: an approximate theory for large density perturbations. *Astronomy & Astrophysics 5*, 84–9.

(156) Zhou, R., Newman, J. A., Mao, Y.-Y., Meisner, A., Moustakas, J., Myers, A. D., Prakash, A., Zentner, A. R., Brooks, D., Duan, Y., Landriau, M., Levi, M. E., Prada, F., and Tarle, G. (2020). The Clustering of DESI-like Luminous Red Galaxies Using Photometric Redshifts. *arXiv e-prints*.

(157) Zwicky, F. (1933). Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta 6*, 110–127.